

Generative AI–Driven Multimedia Semantic Communication
for Wireless Networks: Semantic Representation, Robust
Transmission, and Predictive Modeling

PhD Thesis

Maheshi Udara Lokumarambage

Human Centric AI Research Group

Department of Computer and Information Sciences

University of Strathclyde, Glasgow

April 14, 2026

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by University of Strathclyde Regulation 3.50. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

Abstract

The rapid integration of image and video-based intelligent applications into wireless networks has demanded communication paradigms that go beyond bit-level reconstruction. Semantic communication (SemCom) has emerged as a novel paradigm that shapes the next generation of wireless networks by transmitting “meaning”. However, existing designs suffer from critical gaps due to the lack of alignment with the evolving SemCom theories. These limitations are characterized by the absence of minimally sufficient semantic representations, limited robustness of learned semantics under channel noise, poor scalability across tasks and data distributions, and a lack of evaluation methodologies for semantic distortion, noise, and predictive sufficiency in image and video transmission. This research addresses these gaps through a theoretically grounded approach that leverages representation learning, discrete semantic abstraction, semantic coding, and generative artificial intelligence (AI) coupled with classical channel coding techniques.

First, the thesis lays the foundation for minimally sufficient semantic symbol transmission for generative modeling by conveying semantic meaning as segmentation maps, achieving compression gains of up to $20\times$ and stable semantic reconstruction in low signal-to-noise ratio (SNR) regions where conventional codecs fail. This study laid the groundwork for identifying theoretically grounded practical design considerations for subsequent research.

Second, the study advances the concept of minimal semantic symbol transmission and explores vector-quantized and contrastively aligned semantic representations that function as shared semantic language between transmitter and receiver. This model is an advancement of edge-computing-based SemCom system initially developed,

Abstract

which has used a vector-quantized variational autoencoder (VQ-VAE) to ground representation learning into the semantic extraction process. By adopting index-based transmission protected by classical channel coding, mapped to a shared semantic codebook, the proposed systems achieve improved robustness, scalability, and multi-task accuracy without requiring channel-specific retraining. A Wasserstein-distance-based discriminator is introduced as the main novelty in this framework to stabilize adversarial training and guide learned representations toward a semantic manifold defined by the training image distribution under noisy transmission conditions. The model uses novel multilevel contrastive optimization at pixel, latent, perpetual, and semantic task levels to combat semantic and channel noise. In wireless image transmission, the model achieved a 56.2% reduction in symbol transmission load with lower model complexity and latency than joint source-channel coding (JSCC). In machine-perceived task performance, it achieved 94% classification accuracy, and human-perceived task performance, it achieved 91% image captioning accuracy for SNRs above 4 dB. The design demonstrated learned perceptual image patch similarity (LPIPS) of approximately 0.3 compared to 5.5 with better portable graphics (BPG) in fading channels.

Third, the thesis presents a codebook optimization strategy and a formal quantification of semantic noise arising from symbol and representational ambiguities. A novel evaluation methodology based on the fast gradient sign method (FGSM) is introduced to evaluate the system's robustness to semantic noise. The novel quantization methodology is presented using soft assignments, adversarial discrimination, and entropy-based regularization to minimize semantic ambiguity. The model achieved up to 96.88% codebook usage for codebook sizes of 64 or higher, with higher perplexity than vector quantized variational auto encoder (VQ-VAE). FGSM-based latent semantic noise evaluation showed classification accuracy of 60%-90% across the highest and lowest perturbation magnitudes.

Contents

Abstract	ii
List of Figures	viii
List of Tables	xv
List of Acronyms	xvii
Acknowledgements	xxi
1 Introduction	2
1.1 Motivation	4
1.1.1 Conceptual Framework	6
1.2 Research hypothesis	10
1.3 Research objectives	11
1.4 Publications	12
1.5 Key contributions	13
1.6 Thesis organization and achievements	14
2 Literature review	19
2.1 Theoretical foundations	20
2.2 Semantic communication	28
2.2.1 Knowledge enhanced semantic communication	29
2.2.2 Goal oriented semantic communication	31
2.2.3 Task oriented semantic communication	32

Contents

2.2.4	Joint source channel coding (JSCC)	34
2.2.5	Generative models in semantic communication	35
2.3	Semantic representation in next-generation communication systems	37
2.3.1	Representation learning in semantic communication	38
2.3.2	World models	40
2.3.3	World models and next generation communication	42
2.4	Semantic noise mitigation	42
2.5	Deep learning techniques	45
2.5.1	Generative adversarial network	45
2.5.2	Contrastive learning	46
2.5.3	Vector quantized variational autoencoder	48
2.5.4	Transformers	50
2.6	Semantic communication and edge intelligence	51
2.7	Video semantic communication	52
2.8	Semantic communication applications	54
2.9	Communication channel models and channel coding	57
2.10	Summary	60
3	Segmentation guided semantic symbol encoding for generative semantic communication	61
3.1	Problem formulation	63
3.2	System model	64
3.2.1	Semantic encoder (transmitter)	67
3.2.2	Common knowledge base	68
3.2.3	Semantic decoder (receiver)	70
3.2.4	Communication framework	71
3.2.5	Error concealment block	72
3.3	Simulation results	73
3.3.1	Compression efficiency	74

3.3.2	Impact of channel noise on semantic symbol accuracy and edge preservation	76
3.3.3	Semantic symbol transmission under varying channel noise levels.	78
3.3.4	Quantization effect with zero channel noise	79
3.3.5	Joint quantization and channel noise effects	80
3.3.6	Human perceived semantic fidelity: A qualitative evaluation . . .	82
3.4	Limitations	83
3.5	Conclusion	83
4	Edge computing based semantic communication with vector quantization	86
4.1	Problem formulation	88
4.2	System model	91
4.2.1	Semantic encoder	93
4.2.2	Discrete representation of latent space	93
4.2.3	Semantic decoder (generator)	94
4.2.4	Loss function	95
4.2.5	Communication framework	96
4.3	Simulation results	97
4.3.1	Experimental results on CIFAR-10 dataset	97
4.3.2	Generalization	100
4.3.3	Scalability and qualitative illustration	103
4.3.4	Computational complexity analysis	104
4.4	Limitations and future works	105
4.5	Conclusion	106
5	Contrastive vector-quantized generative semantic representation for robust semantic communication	107
5.1	Problem formulation	111
5.2	System model	112
5.2.1	Semantic encoder	114

Contents

5.2.2	Semantic decoder	117
5.2.3	Wasserstein distance and critic design	118
5.2.4	Model loss function	120
5.2.5	Multilevel contrastive optimization	122
5.2.6	Communication architecture	124
5.2.7	Performance metrics	126
5.3	Theoretical analysis	128
5.4	Simulation results	131
5.4.1	Rate distortion analysis	135
5.4.2	Task performance	139
5.4.3	Bandwidth efficiency	142
5.4.4	Model complexity	143
5.4.5	Ablation study	145
5.4.6	Generalization	147
5.4.7	Semantic distortion	148
5.4.8	Visual analysis	148
5.5	Limitations	150
5.6	Conclusion	151
6	Semantic noise modeling and evaluation through shared knowledge	
	base optimization	153
6.1	Problem formulation	155
6.2	System model	156
6.2.1	The quantization process	157
6.2.2	Discriminator	160
6.2.3	Loss function	160
6.2.4	Semantic inference pipeline	163
6.2.5	Interpretation of codebook clustering	163
6.3	Simulation results	166
6.3.1	Codebook analysis	166
6.3.2	Clustering	169

Contents

6.3.3	Semantic noise robustness	170
6.3.4	Quality of transmission	172
6.3.5	Generalization and scalability to diverse datasets	174
6.3.6	Computational cost and scalability	175
6.4	Limitations and future work	176
6.5	Conclusion	176
7	Conclusion	178
7.1	Future work	181
	Bibliography	183
A	Appendix A	215
A.1	Mathematical proofs	215
A.1.1	Proof of proposition 5.3.1	215
A.1.2	Proof of Proposition 5.3.2	216
A.1.3	Proof of proposition 5.3.3	217
A.1.4	Proof of proposition 5.3.4	218
A.1.5	Proof of proposition 5.3.5	219
A.2	Training stability	220

List of Figures

1.1	Conceptual framework on semantic communication that guides the thesis objectives.	8
2.1	Traditional communication system.	21
2.2	Semantic symbols and language definition.	25
2.3	Literature review flow	29
2.4	World-model pipeline illustrating the construction of implicit latent representations from visual observations and the prediction of future states through learned dynamics. In a practical sense, the semantic encoder compresses raw image frames into compact latent variables that capture semantics, while the dynamics model simulates prospective environment evolution conditioned on past states or actions.	41
2.5	Contrastive learning: make the positive pairs closer and push negative pairs apart.	47
2.6	Semantic communication applications	54
3.1	Conceptual three-layer model of semantic communication, comprising the technical layer for signal transmission, the semantic layer for meaning extraction, and the effectiveness layer for task-driven interpretation.(Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)	66

List of Figures

3.2	Semantic communication system for image transmission, illustrating semantic extraction, channel transmission, and semantic-driven generation. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)	67
3.3	Visualization of segmentation masks and the associated transmitted images, showing how semantic regions are encoded and preserved in semantic communication. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)	69
3.4	Adversarial training loop of a GAN, illustrating generator–discriminator updates. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)	71
3.5	Error-correction performance of polar codes under different block lengths.(Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)	72
3.6	Visual illustration of the transmitted image size comparison for a sample image. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)	74
3.7	Comparison of compression ratios for sample images with the proposed model and traditional compression scheme. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)	75
3.8	Illustration of JPEG image compression with added noise. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)	76
3.9	Effect of channel noise on the semantic segmentation map, illustrating how increased noise levels degrade semantic consistency and object boundary clarity. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)	77
3.10	Effect of median filtering on segmentation maps corrupted with channel noise, demonstrating effective noise removal while preserving sharp edges and structural details. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)	77

List of Figures

3.11 PSNR versus SNR for a sample of 10 semantic segmentation map images after noise concealment. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)	78
3.12 PSNR versus SNR transmitted image sample of 10. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)	79
3.13 PSNR versus SNR without noise concealment for transmitted semantic segmentation map image sample of 10. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)	80
3.14 Impact of combined quantization effects and channel noise on transmitted compressed semantic segmentation map images, with noise concealment applied to mitigate distortions and preserve semantic regions. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)	81
3.15 Image quality at different noise levels for generated images with joint quantization and channel noise. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)	82
4.1 Theoretical semantic communication (SemCom) system model illustrating semantic noise at the extraction and coding stages and channel noise during physical transmission. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	90
4.2 Illustration of the implementation of the proposed system	92
4.3 The VQ-based index transmission system includes the physical channel with channel coding.	93
4.4 Classification accuracy vs. signal-to-noise ratio (SNR) on the CIFAR-10 test set under additive white Gaussian noise (AWGN) noise for $K = 32, 64, 128$, and $L = 16$	98
4.5 Error correction performance for simulated AWGN and fading channels with Sionna for information bit size 512 with different coding rates for low-density parity-check (LDPC) using quadrature phase shift keying (QPSK)	99

List of Figures

4.6	(a) Image quality (CIFAR-10) at different SNR levels over AWGN channel with LDPC coding for QPSK, (b) Image quality (CIFAR-10) at different SNR levels for flat fading channel with LDPC coding for QPSK (the proposed only requires 75% of the transmitted symbols compared to joint source-channel coding (JSCC) with code rate = 0.5).	100
4.7	Tiny ImageNet dataset : (a) PSNR at different SNR levels over AWGN channel with LDPC coding for QPSK, (b) PSNR at different SNR levels for flat fading channel with LDPC coding for QPSK.	102
4.8	Tiny ImageNet dataset : (a) multi-scale structural similarity index (MS-SSIM) at different SNR levels over AWGN channel with LDPC coding for QPSK, (b) MS-SSIM at different SNR levels for flat fading channel with LDPC coding for QPSK.	102
4.9	Tiny ImageNet dataset : (a) learned perceptual image patch similarity (LPIPS) at different SNR levels over AWGN channel with LDPC coding for QPSK, (b) LPIPS at different SNR levels for flat fading channel with LDPC coding for QPSK.	103
4.10	Illustration of image reconstruction quality for Kodak and CelebA-HQ dataset for Rayleigh flat fading channel with LDPC coding with code rate of 0.5. The top of each image shows the values for peak signal-to-noise ratio (PSNR)/MS-SSIM/LPIPS.	104
5.1	Training phase of semantic vector quantized Wasserstein generative adversarial network (SVQ-WGAN): quantization of continuous features using the learned codebook, noise injection in the quantized latent space, and random element masking with probability $\mathbb{P}_{mask}$. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	113
5.2	Image transmission over a noisy channel: codebook indices from the semantic-encoded and quantized image are mapped to a bitstream and transmitted over AWGN and fading channels with LDPC channel coding. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	114

List of Figures

5.3	Residual block. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	115
5.4	Learning rate schedule (a) critic learning rate schedule (b) generator learning rate schedule. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	132
5.5	Spatial attention heat maps for sample images from the Kodak dataset illustrating edge-focused attention distribution. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	133
5.6	Image transmission performance on the Kodak dataset over an AWGN channel across different SNR levels using QPSK modulation and LDPC coding with multiple coding rates. Results are presented for a codebook size of 256: (a) average PSNR (higher is better), (b) average MS-SSIM (higher is better), and (c) average LPIPS (lower is better). (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	137
5.7	Image transmission performance over fading channel for varying SNRs for Kodak dataset with different coding rates for LDPC for QPSK: (a) average PSNR (higher the better), (b) average MS-SSIM (higher the better), and (c) average LPIPS (lower the better). (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	137
5.8	robustness index (RI) evaluation with different channel conditions (a) AWGN channel, (b) fading channel, (c) bit error rate (BER). (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	138
5.9	Quality versus BER for Kodak dataset over Rayleigh fading channel with LDPC coding for QPSK : (a) average PSNR (higher the better), (b) average MS-SSIM (higher the better), and (c) average LPIPS (lower the better). (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	139
5.10	Classification accuracy on the CIFAR-10 dataset across different SNR levels. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	141

List of Figures

5.11	Ablation study evaluating different codebook sizes and training configurations, where n/n denotes no noise applied during training: (a) PSNR vs. SNR, (b) MS-SSIM vs. SNR, (c) LPIPS vs. SNR, and (d) RI vs. SNR. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	146
5.12	Model performance under different configurations across varying SNR levels: (a) PSNR (b) LPIPS. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	147
5.13	Image quality across different SNR levels for multiple datasets using QPSK modulation with LDPC coding at coding rate (CR)=0.5 over a fading channel, evaluated (PSNR/MS-SSIM/LPIPS). (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	149
5.14	Zoomed image patches under zero channel noise conditions for images from the Kodak and DIV2K datasets. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	150
5.15	Zoomed image patches from the CelebA-HQ dataset and a 4K image under zero channel noise conditions. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	150
6.1	Training phase: continuous feature representations are quantized using the learned codebook and soft assignments. The complete training process is defined in algorithm 4.	156
6.2	(a) Codebook usage for different codebook sizes for the proposed model and VQ-VAE. (b) Perplexity and codebook embedding entropy for different codebook sizes.	168
6.3	density-based spatial clustering of applications with noise (DBSCAN) clustering of codebook embeddings for CIFAR-10 class labels and clustering circles.	169
6.4	Classification accuracy under semantic-level (latent) and pixel-level perturbations for varying codebook sizes 64,128, and 256 of the proposed system.	171

List of Figures

6.5	LPIPS performance on Tiny-ImageNet for image transmission over a fading channel at $R_c = 0.5$. The comparison includes better portable graphics (BPG), VQ-VAE, and the proposed method, demonstrating generalization under a more diverse data distribution.	174
6.6	Visual quality of transmitted representative Imagenet samples with fading channel at code rate of 0.5. The quality metrics illustrated are (PSNR/MS-SSIM/LPIPS) at SNR = 3–6 dB, GT – ground truth image.	175
A.1	losses (a) Reconstruction Loss (b) vector quantization (VQ) Loss. (c) Discriminator Loss (d) Total Loss.	221

List of Tables

2.1	Summary of recent advancements in generative AI-based SemCom systems. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	37
4.1	Computational cost and inference latency for image reconstruction . . .	105
5.1	Comparison of SimCLR, MoCo v3, and SVQ-WGAN	124
5.2	BERTScore F1 performance across varying SNR levels for BPG and the proposed model. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	140
5.3	Effective bitrate comparison between the proposed compact index-based transmission paradigm and BPG with LDPC coding. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	143
5.4	Encoder/Decoder FLOPs ($\times 10^9$) and asymptotic complexity at 256×256 resolution. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	144
5.5	Comparison of SVQ-WGAN and deep JSCC in terms of parameters. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	144
5.6	Model evaluation. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	148
5.7	Task-oriented semantic distortion. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)	148

List of Tables

6.1	Perplexity and codebook embedding entropy values for different embedding sizes.	168
6.2	PSNR (dB) (higher is better) versus SNR on CIFAR-10 over fading channel for different code rates (R_c).	172
6.3	MS-SSIM (higher is better) versus SNR on CIFAR-10 over fading channel for different code rates (R_c).	173
6.4	LPIPS (lower is better) versus SNR on CIFAR-10 over fading channel for different code rates (R_c).	173

List of Acronyms

5G	fifth generation
6G	sixth generation
AI	artificial intelligence
AoII	age of incorrect information
AR	augmented reality
AWGN	additive white Gaussian noise
BER	bit error rate
BERT	bidirectional encoder representations from transformers
BLEU	bilingual evaluation understudy
BLIP	bootstrapping language-image pre-training
BPG	better portable graphics
BPSK	binary phase-shift keying
bpp	bits per pixel
CABAC	context-adaptive binary arithmetic coding
CL	curriculum learning
CNN	convolutional neural networks
CPP	channel-per-pixel
CR	coding rate
CSI	channel state information

List of Acronyms

DBSCAN	density-based spatial clustering of applications with noise
DCT	discrete cosine transform
DMC	discrete memoryless channel
ELBO	evidence lower bound
EMA	exponential moving average
EP	error probability
FEC	forward error correction
FGSM	fast gradient signed method
FID	Fréchet inception distance
FLOP	floating-point operations per second
FPGA	field-programmable gate array
GAN	generative adversarial network
GNN	graph neural network
GO-SemCom	goal-oriented semantic communication
GPU	graphics processing unit
HARQ	hybrid automatic repeat request
IoT	internet of things
IoU	intersection over union
IoV	internet of vehicles
ISAC	integration of sensing and communication
ITS	intelligent transportation systems
JEPA	joint embedding predictive architecture
JPEG	joint photographic experts group
JSCC	joint source-channel coding
KB	knowledge base

List of Acronyms

KG	knowledge graph
KID	kernel inception distance
LDPC	low-density parity-check
LLM	large language model
LLR	log-likelihood ratio
LMMSE	linear minimum mean square error
LPIPS	learned perceptual image patch similarity
LSTM	long short-term memory
M2H	machine-to-human
M2M	machine-to-machine
MMSE	minimum mean square error
MS-SSIM	multi-scale structural similarity index
mmWave	millimeter-wave
NOMA	non-orthogonal multiple access
POMDP	partially observable Markov decision process
PSNR	peak signal-to-noise ratio
QAM	quadrature amplitude modulation
QoE	quality of experience
QPSK	quadrature phase shift keying
ReLU	rectified linear unit
RI	robustness index
RL	reinforcement learning
RNN	recurrent neural network
ROI	regions of interest
RONI	region of non-interest

List of Acronyms

SAM	segment anything model
SemCom	semantic communication
SF	semantic fidelity
SMI	semantic mutual information
SNR	signal-to-noise ratio
STE	straight-through estimator
SVQ-WGAN	semantic vector quantized Wasserstein generative adversarial network
t-SNE	t-distributed stochastic neighbor embedding
THz	terahertz
TOSC	Task-oriented semantic communication
UAV	unmanned aerial vehicle
VAE	variational auto encoder
ViT	vision transformer
VQA	visual question answering
VQ	vector quantization
VQ-VAE	vector quantized variational auto encoder
VR	virtual reality
WGAN	Wasserstein generative adversarial network
XR	extended reality

Acknowledgements

This doctoral thesis is the result of the continued support and assistance of many individuals and organizations over the years, without whom its completion would not have been possible. First, I want to express my gratitude to my supervisor, Prof. Anil Fernando, for this opportunity and advice. I am incredibly grateful to my supervisor, Prof. Feng Dong, whose support was invaluable during challenging times. His thoughtful direction and advice were instrumental in helping me move forward. My sincere appreciation also goes to the University of Strathclyde for awarding me the scholarship that made this journey possible.

I extend my heartfelt thanks to my collaborators at the University of Oulu, Finland, Prof. Nandana Rajathewa, Dr. Hoessein Razei, and Dr. Thushan Sivalingam, for their continued support and insightful contributions. A special thank you goes to Thushan, not only for being a dedicated research collaborator but also for being a true friend throughout this journey.

I am deeply grateful to Sri Lanka for providing me with free education from primary school through bachelor's, enabling me to grow. I am deeply grateful to all my school teachers at Vihara Maha Devi Girls' School, Badulla, Sri Lanka, and Badulla Central College, Sri Lanka, for providing me with a strong foundation in primary and secondary education. I would also like to thank my lecturers at the University of Peradeniya, where I completed my bachelor's degree, and my master's supervisor, Prof. Chandana Gamage, and all the lecturers at the University of Moratuwa, Sri Lanka, for their guidance and support during my postgraduate studies. I am also thankful to my past managers and colleagues at Sri Lanka Telecom PLC, especially Mr. Rohitha Somaratne, for training me to navigate the technological landscape with confidence and independence.

My family has been the foundation for everything. I am forever grateful to my mother for her unconditional love and constant encouragement throughout my life. My heartfelt thanks also go to my brother, Manoj, who always listened to me and supported me with patience. I am grateful to my husband, Kumudu, for his support throughout this journey and for encouraging me to pursue this opportunity. Finally, to my son, Ayen, thank you for your patience and love throughout this journey. Your resilience in adapting to new environments and constant changes, by moving to a new country at such a young age, has inspired and determined me to overcome every obstacle along the way. This thesis is lovingly dedicated to you, Ayen!

Chapter 1

Introduction

The evolution of human societies has been profoundly shaped and accelerated by successive waves of technological innovation, especially by the rapid pace driven by the rise of modern communication systems, which improved connectivity. The cornerstone is the transformative leap that connects the individuals, societies, and intelligent systems that made the vast knowledge that contributed to this development. While the basic principle governing both human and animal communication is the exchange of signals with a receiver, the nature of these exchanges differs significantly. Human communication is a system of open generative system capable of expressing abstract concepts, creating new meanings and planning far beyond the present. In contrast, animal communication remains largely a context-specific signal serving immediate needs such as “food”. The nature of human communication has led to the vast development of societies. Despite this rich natural system of human communication, the design of human-engineered communication systems has historically ignored the meaning of the information being exchanged. Rooted in Shannon’s mathematical theory [3], modern wireless systems operate purely at the syntactic level, optimizing bit-level reconstruction without regard to semantic content, intent, or task relevance.

This disconnection has become increasingly problematic as wireless networks have integrated unprecedented levels of intelligence and applications, especially in image and video data. The sixth generation (6G) of communication is believed to deliver higher data rates, sub-millisecond latencies, global-scale connectivity, and seamless

fusion of the physical and digital worlds [4–6]. This connected ecosystem is expected to be comprised of artificial intelligence (AI), sustainable green technologies, and terahertz (THz) communication, supporting applications such as holographic telepresence, digital twins, and the Internet of Everything (IoE) [7]. As we enter this new era of intelligence, 6G represents not just an incremental upgrade in bits, but a paradigm shift that demands modern communication systems understand and transmit meaning, similar to the human communication system. This emerging necessity is not just about reducing strain on wireless communication systems, but also about motivating a fundamental rethinking of how information should be represented, transmitted, and interpreted in next-generation networks, to facilitate the next generation of intelligent applications and devices, which will consist of intelligent agents.

SemCom redefines the design grounds of information exchange by shifting the optimization into semantically relevant transmission from bit-accurate reconstruction. In essence, SemCom aims to extract and convey the intended meaning of the underlying data so the receiver can understand the meaning and infer new states or act upon or reason based on the meaning to go to new states or predict the future. This contrasts with traditional communication systems, whose primary objective is to reproduce the transmitted bitstream with minimal distortion. This paradigm is grounded in the fundamental principle that not all the data transmitted is relevant to the receiver, specifically for a given task or context. This feature of SemCom helps to reduce the bandwidth strain on next-generation wireless networks. This can be realized by adopting deep learning-based technologies, shared data, or knowledge and context-aware reasoning. This development inherently reduces redundancy, resulting in lower bit rates and aligning communication performance with intelligent purposes [8,9]. These qualities are essential for next-generation networks relying heavily on intelligent decision-making.

In lieu of this transformative shift in the communication system design, SemCom remains in its formative stage, with several design directions and concepts introduced. In another line of research, the basic theoretical frameworks are being developed that govern the SemCom. Despite these advancements, several fundamental challenges remain unresolved. Existing approaches in designing SemCom systems often rely

on deterministic data-driven approaches, rather than knowledge-driven, and these pipelines lack generality, interpretability, and measurable semantic fidelity. Furthermore, these designs often do not align with fundamental theories that govern SemCom or a standardized framework ensuring robust representations, thereby restricting their implementations to a level of complexity required for current 6G applications already in pipelines, that may require machine-to-machine (M2M) interaction and low data rates. This thesis is motivated by the identified gaps and is based on a carefully designed framework for modern SemCom.

1.1 Motivation

This section outlines the motivation for this research in two directions. First is the motivation for the necessity of SemCom, and the second is the motivation towards the generative AI-based design paradigm that optimizes the semantic extraction and generation.

Why semantic communication The rapid growth in mobile data demand, driven largely by image and video applications, has increasingly strained current communication systems. According to estimates from the ITU [10], global mobile traffic is expected to rise dramatically across all service categories in the coming years. This report estimated three categories of traffic: non-video traffic (text, audio, web browsing, messaging), video traffic (streaming, video calls, images), and M2M communication, such as internet of things (IoT) sensors and smart devices. In 2030, video traffic is expected to increase by a factor of 10 over the next 5 years, starting from 2025. Meanwhile, data generated by M2M communication is projected to account for 7% of total mobile traffic in 2020, rising to 12% by 2030. This trend underscores the urgent need for more efficient communication paradigms beyond traditional approaches, which are limited by Shannon’s theoretical channel capacity for transmitting image and video data.

To address the escalating demand for wireless data transmission, it is essential to reconsider the limitations imposed by traditional communication theories on physical channel limits. The fundamental constraints are described by Shannon’s theorem [3],

which defines the maximum achievable data rate or channel capacity of a communication channel subject to noise. This limit is given as

$$C_{Phy} = W \log_2 \left(1 + \frac{P}{N_0 W} \right) \quad , \quad (1.1)$$

where C_{Phy} is the channel capacity in bits per second, W is the bandwidth of the channel in hertz, P is the average signal power, and $N_0/2$ is the power spectral density of the AWGN. Accordingly, to increase the data rate C_{Phy} , we need to increase the bandwidth W or the signal power P . Increasing signal power leads to an inefficient system design, which contradicts sustainable goals. In the current literature, expanding the bandwidth W is leveraged with higher frequency bands, such as millimeter-wave (mmWave) and terahertz (THz) spectrums [11]. Operating in these frequencies presents challenges, such as high propagation losses due to atmospheric absorption and limited diffraction, which lower communication ranges and reliability. This remains a critical area of research to develop advanced solutions such as highly directional antennas and sophisticated beamforming techniques [12, 13]. These traditional communication systems cannot theoretically operate beyond Shannon’s theoretical capacity limit. There is a growing need for a paradigm shift in designing SemCom systems. SemCom is a promising technology that can reduce the strain on communication channels by reducing data rates.

Generative AI-based semantic communication The emergence of deep learning techniques, spanning representation learning, vector-quantized latent modeling, and generative AI-based models, provides the research community with the opportunity to rethink how communication systems are designed through the concepts of semantics, abstraction, and prediction. In principle, visual generative models learn minimally sufficient representations from data distributions that capture high-level structures, which are inherently compact and rich in meaningful features. These capabilities are directly aligned with the objectives of SemCom, which in essence are not to reproduce the signal at the receiver, but to preserve its meaning, its functional relevance, and its predictive value with minimal information.

The second motivation factor for generative models in SemCom stemmed from the uncertainty in wireless environments and the semantic coding process. While traditional communication systems mitigate channel noise solely at the physical layer, SemCom must address it across multiple layers, including symbol corruption, insufficient representation, and semantic ambiguity [14]. Generative models offer a structured methodology to handle these by allowing corrupted information or perturbed representations to be projected back into a coherent semantic manifold. Modern intelligent applications, such as tele-surgery, tele-robotics, drone applications, extended reality (XR), and video understanding-based applications, require not only semantics but also temporal and causal structures to make intelligent decisions.

One of the key motivations for the research objectives of this thesis is the practical deployment considerations in SemCom, where they shall be lightweight, scalable, and compatible with existing communication standards, supporting real-world communication scenarios. This moves semantic communication through a theoretically oriented approach to a deployable technology, where developing SemCom systems are aligned with modern theories and design paradigms as well as practical deployability.

Collectively, the aforementioned motivating factors drive the direction towards a visual generative AI-based SemCom that integrates effective representation learning, robust semantic coding, and predictive modeling. Such systems can fundamentally shift the intent of communication systems from bit transmission to reconstruction-driven SemCom systems that convey abstract meaning to recover intent and reactive responses to proactive predictions. The design decisions for semantic extraction for robust SemCom systems are motivated by the conceptual framework that goes beyond traditional SemCom.

1.1.1 Conceptual Framework

The main objectives of this thesis are to move beyond isolated encoder-decoder architectures that solely compress data by extracting semantic features, but to adopt a principled view of how semantics are represented and utilized across the communication pipeline. While existing literature shows unprecedented advances in task-oriented

coding, semantic extraction, and multimodal generative reconstruction, these efforts remain fragmented and lack a unifying conceptual structure. This thesis is principally motivated by the following conceptual framework that guides the design of the future direction of SemCom. The thesis has identified five key areas [9] for the design and operation of SemCom systems.

- **Minimal representations:** Minimal representation in SemCom captures the essential information required to understand, reconstruct meaning, or execute a task. This differs from deterministic latent compression, which relies on feature mappings and was originally inspired by representation learning [15]. The minimal semantic representations aim to model the underlying probability distributions governing the data rather than its pixel-level appearance. By encoding the generative factors, structure, and semantic regularities of the data distribution, these representations enable the system to preserve meaning and infer missing information, which in the SemCom context, remains robust under physical channel degradation.
- **Semantic language:** The semantic language defines the shared vocabulary through which meaning is communicated between the transmitter and receiver. This can be a consistent mapping between semantic symbols and their semantic interpretations. The practical implementation can be a codebook, a set of class labels that represent pixel-based spatial mappings (semantic segmentations) or task-specific symbols. The objective of efficient and robust SemCom is to implement a semantic language that minimizes semantic ambiguity and supports reliable semantic inference even under noisy or resource-constrained channels.
- **Computing and wireless resources:** This pillar addresses the practical constraints under which SemCom systems expect to operate, including limited bandwidth, energy, latency, and computational power. Efficient semantic encoders, lightweight models for edge devices, and symbol-efficient transmission strategies are required to meet these constraints.
- **Semantic-driven evaluation:** Semantic-driven evaluation involves assessing communication performance not through pixel-level distortion but through metrics

that reflect meaning, task relevance, and contextual relevance or applicability. These metrics can quantify how well the transmitted minimal representations preserve meaning, support downstream inference, and how well they align with the intended semantic tasks of the receiver.

- **Causal structure:** Causal structure refers to the ability of semantic encoding to capture the underlying causal relationships within the data. By learning latent dynamics and dependencies that govern how the states evolve, the system can generalize beyond observed samples and predict future outcomes, which is particularly important in video SemCom. Incorporating causality strengthens robustness, supports reasoning, and enables predictive semantics, reducing the need for continuous, high-volume data transmission.

These pillars form the conceptual backbone that serves as a guiding framework for the methodological developments presented in later chapters, as illustrated in figure 1.1.

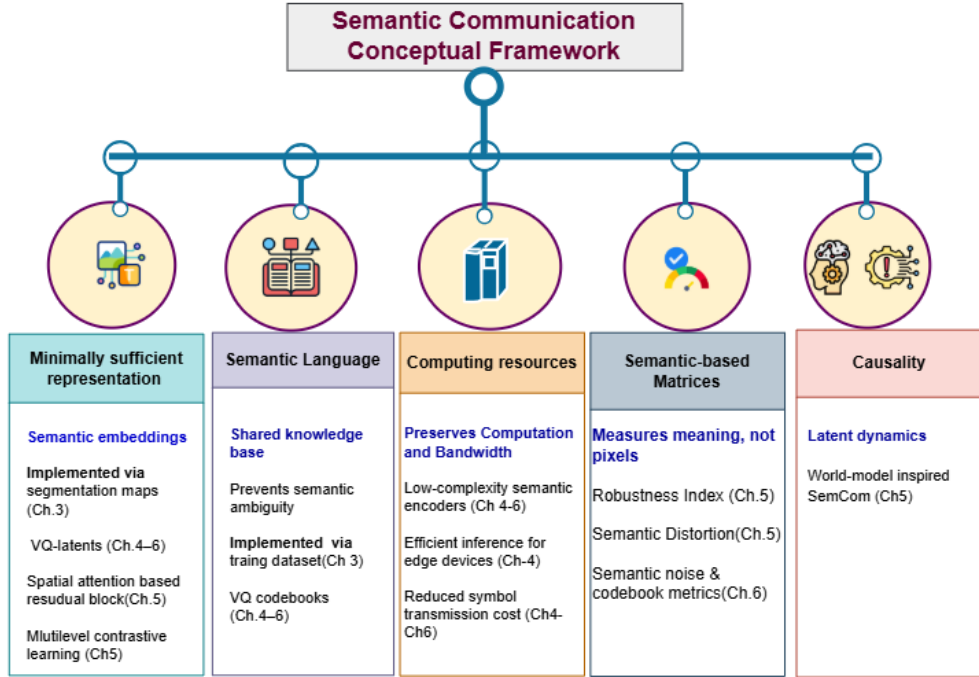


Figure 1.1: Conceptual framework on semantic communication that guides the thesis objectives.

This conceptual framework defines what SemCom should be like and what it should

achieve in design. The following subsection presents how those conceptual design pillars motivate the operational architectural choices and deep learning techniques employed in this thesis, grounded in existing research gaps.

Vector quantization–based semantic representation A primary limitation of existing SemCom systems is the transmission of continuous latents over physical channels, trained jointly with channel conditions [16]. These systems are often effective with the matching training conditions, but limit their adaptability in heterogeneous channel conditions. When using deterministic deep learning techniques such as standard autoencoders, deepJSCC, convolutional neural networks (CNN), or transformer encoders, the representations are highly sensitive to channel perturbations, where small channel perturbations can lead to large semantic deviations. Data-driven approaches lack semantic interpretability but provide only semantic-based compression, which does not align with the shared semantic language as interpreted by modern SemCom theories.

VQ representations discretize semantics into stable codebook symbols, enabling graceful degradation and robust meaning preservation under noisy wireless channels. The codebook functions as a shared semantic vocabulary between transmitter and receiver, complying with the conceptual design of SemCom. This discretization introduces structured redundancy at the semantic level, improves robustness to channel errors, and can enable index-based transmission, which can achieve compression. In contrast to task-driven tokenization or ad hoc semantic symbol definitions, VQ learns semantic manifolds that interpret abstract semantic concepts, while remaining compatible and interpretable.

Wasserstein Distance and Adversarial Training Generative semantic decoders can reconstruct meaningful content from noisy or corrupted semantic representations. The conventional generative adversarial network (GAN)s are known for their instability during training [17]. The Wasserstein distance [18] is based on the minimum cost required to convert one probability distribution to another and provides an effective measure of distributional discrepancy between the real and the generated. By employing a Wasserstein-distance-based critic [19], a stable training process can be achieved. Further,

the decoder can learn to map corrupted or missing symbol sequences into a meaningful semantic manifold rather than performing pixel-based reconstruction. The design is motivated by the need to retain semantic integrity under channel and semantic perturbations, rather than pixel-based reconstructions from latents. Wasserstein distance-based adversarial learning is adopted to stabilize semantic generation under imperfect semantic inputs explicitly.

Index-based transmission with classical channel coding Transmitting full latents over a communication channel has two main drawbacks. First, while they achieve semantic-based image compression, they are susceptible to channel noise because neural networks implicitly learn channel protection. The second is that the compression achieved is limited by poor pixel-based reconstruction. By mapping the latent vectors into a codebook and transmitting the representative set of indices, it is expected to achieve high compression while maintaining semantic integrity. Further, index-based transmission decouples the semantic representation from the physical-layer reliability, enabling the use of standardized and mature channel coding techniques [20]. Owing to these clues, this thesis is motivated by the position that semantic communication should not replace physical-layer reliability mechanisms but should coexist with them. Leveraging classical channel coding benefits by providing predictable error behaviour to explicitly evaluate the semantic representation effectiveness, hardware compatibility, and adaptability to diverse channel conditions. The design choice shifts from channel-aware semantic encoders to channel-agnostic semantic representations, encouraging practical deployability. Based on these motivations, the remainder of this chapter presents the research objectives and design paradigms, aligning with the conceptual framework that systematically structures the key architectural and deep learning based model decisions adopted in this thesis.

1.2 Research hypothesis

- **H1:** SemCom systems that transmit minimally sufficient semantic symbols when coupled with generative AI decoders, can reliably convey semantic under wireless

channel impairments, even when pixel-level information is corrupted or lost.

- **H2:** SemCom systems that leverage deep representation learning via vector quantization provide robustness, interpretability, and task-agnostic generalization under channel noise.
- **H3:** Index-based semantic transmission with classical channel coding enables comparable semantic fidelity while maintaining high bandwidth efficiency and practical implementation capabilities.
- **H4:** Wasserstein-distance-based adversarial training enables a more stable training and guides the training process towards the semantic manifold, even when semantic symbols are corrupted or lost during physical channel transmission.

1.3 Research objectives

The primary objective of this research is to design and validate semantic representation methodology optimized for SemCom that goes beyond classical data-driven approaches [9]. This work aims to operationalize the theoretical constructs and design paradigms presented in next-generation semantic communication frameworks and theories [21–23] by introducing practical, robust, and scalable system designs capable of transmitting meaning, not merely data. The extracted semantics shall be compatible with the existing channel coding techniques, where channel-level optimization is not within the scope of this thesis; instead, we expect to use the existing channel coding methodologies. The objectives of the research are listed below.

- **To develop a lightweight semantic encoder that supports semantic representation learning, and extracts compact and noise-robust semantic representations for images.** This objective serves as the basis for minimal sufficient semantic representations aligned with the thesis’s conceptual framework. The extracted semantic representations will be the basis for an end-to-end training and inference pipeline for the designed image transmission systems.

- **To build a practical, resource-constrained semantic communication system suitable for edge computing scenarios.** This aligns with the design paradigm of discrete semantic representations via index transmission while reducing model complexity, inference cost, and bandwidth consumption, enabling real-world deployability in edge–cloud systems.
- **To formally model semantic noise and introduce an evaluation methodology for shared knowledgebase optimization and semantic noise.**

This objective assesses how the shared knowledge base affects the meaning preservation in SemCom. By introducing metrics and evaluation methods that capture semantic ambiguity and representational drift, the goal is to ensure that learned semantic codebooks remain stable and meaning-consistent. This strengthens the reliability of semantic communication where traditional distortion measures are insufficient.

- **To evaluate SemCom systems comprehensively.** Develop novel metrics considering semantic fidelity, noise robustness, and downstream task achievement ability, and comprehensively evaluate the developed SemCom, complementing traditional and existing semantic metrics.

1.4 Publications

The research conducted during this PhD has been either published in or submitted to peer-reviewed venues. Each publication is directly linked to the corresponding chapter that discusses its content

1. **Wireless end-to-end image transmission system using semantic communications**, M. U. Lokumarambage, V. S. S. Gowrisetty, H. Rezaei, T. Sivalingam, N. Rajatheva, and A. Fernando, in *IEEE Access*, vol. 11, pp. 37149-37163, 2023, doi: 10.1109/ACCESS.2023.3266656.
2. **Edge computing based vector quantized semantic communication system for wireless image transmission**. M. Lokumarambage, T. Sivalingam, H. Rezaei, N. Rajatheva, and A. Fernando. [Submitted to EUCNC 2024 (November).

(Accepted as a Poster)] [Improved manuscript accepted in IEEE Transactions on Consumer Electronics.]

3. **Generative AI-based vector quantized end-to-end semantic communication system for wireless image transmission.** M. Lokumarambage, T. Sivalingam, F. Dong, N. Rajatheva, A. Fernando. IEEE Transactions on Machine Learning in Communications and Networking, vol. 3, pp. 1050-1074, 2025, doi: 10.1109/TMLCN.2025.3607891
4. **Codebook-enabled semantic communication with contrastive and entropy-based optimization to reduce semantic noise.** M. Lokumarambage, T. Sivalingam, F. Dong, N. Rajatheva, A. Fernando. (2025). [IEEE Transactions on Vehicular Technology (Under Review)]
5. **Robust semantic communication with soft-assignment based vector quantization and joint regularization.** M. Lokumarambage, F. Dong, A. Fernando. (2025). [University of Strathclyde CIS student conference - 2025 (Poster Presenter)]

1.5 Key contributions

This thesis makes the following scientific contributions to the field of semantic communication, advancing the theoretical understanding, algorithm design, and practical realization of generative AI-driven SemCom systems for image transmission.

Establishing generative modeling-based representation learning as a foundational paradigm for semantic extraction in images This thesis advances SemCom theory-based practical model design by demonstrating that generative, vector-quantized, and contrastively aligned representations provide a principled foundation for learning minimally sufficient semantics that remain stable under channel and semantic distortions. Unlike prior work that ties semantics to task-specific or JSCC encoders, this thesis introduces a unified representation framework that progressively improves from symbolic segmentation maps to discrete latent spaces and

contrastively regularized embeddings, which consistently preserves meaning across physical impairments, tasks, and data distributions. The contributions demonstrate that quantized latent spaces that function as semantic manifolds enable robust decoding, scalable high-resolution modeling (up to 4K images), and task-agnostic generalization in image transmission. This establishes a scientific approach to semantic abstraction that bridges machine perception, rate-distortion, and communication-system design.

Introducing a formal treatment of semantic noise and shared knowledge structures for reliable semantic transmission and measuring in image transmission. A scientifically novel contribution of this thesis is the formalization of semantic noise, that is, the distortions arising from quantization ambiguity, representation mismatches, and concept-level perturbations that are not captured by traditional metrics. The thesis establishes a methodology for evaluating how knowledge bases (codebooks) structure meaning, how semantic distortions affect semantics, and how symbolic representations can be optimized in lossy environments. The resulting models offer deeper scientific insights into how semantics should be encoded, optimized, and protected in next-generation networks. The semantics-based novel metrics introduced in each chapter provide a comprehensive evaluation of SemCom systems, providing scientific knowledge contribution to SemCom metrics.

1.6 Thesis organization and achievements

This thesis presents its technical chapters validating the research hypothesis that generative AI can fundamentally enhance semantic communication by enabling minimally sufficient representations, structured semantic languages, temporal modeling, and the use of semantically meaningful performance metrics, resulting in scalable communication architectures. Each chapter contributes to the aforementioned objectives mentioned in the section.

Chapter 2 Chapter 2 provides the foundational background for the subsequent technical developments by surveying existing research, challenges, and open problems in

Chapter 1. Introduction

semantic communication. It begins by outlining the shift from bit-level communication to meaning-centric transmission, examining key concepts such as semantic extraction, knowledge representation, and task-oriented encoding. Existing semantic communication architectures, including JSCC-based systems, attention-driven encoders, contrastive representation learning methods, and transformer-based semantic modules, are reviewed with emphasis on their design principles and limitations. Techniques ranging from invariant and discrete neural representations to knowledge-graph embeddings and large-model-based semantic priors are summarized, highlighting their impact on robustness, compression, and generalization. The chapter further discusses emerging directions, including causal reasoning, world models, and prediction-aware semantics, and illustrates their relevance for next-generation AI-native wireless networks. Finally, applications across intelligent transportation systems (ITS), XR systems, edge intelligence, and distributed sensing are examined to contextualize practical challenges, thereby motivating the need for unified, generalizable SemCom frameworks developed in the following chapters.

Chapter 3 Chapter 3 lays the methodological groundwork by introducing a generative AI-based SemCom architecture built on a three-layer model comprising semantic extraction, shared knowledge, and semantic reconstruction. This chapter establishes the feasibility of transmitting minimal semantic representations, specifically segmentation maps, over noisy channels. The framework achieves substantial compression gains, yielding a $17.2\times$ reduction using raw semantic maps and up to $20.6\times$ with semantic-map compression, corresponding to approximately 5% of the payload required by conventional joint photographic experts group (JPEG) across all coding rates. Under noisy channel conditions, the proposed system maintains stable PSNR and MS-SSIM performance within the 2.1–3.0 dB range, where traditional codecs such as JPEG collapse, demonstrating the advantage of meaning-centric regeneration over pixel-centric reconstruction. Human qualitative evaluation further confirms semantic fidelity, with an average score of 4.19/5, indicating consistent preservation of object-level meaning. The system remains operational up to 2.2 dB noise with polar coding, beyond which

Chapter 1. Introduction

GAN-based reconstruction fails. The results reveal that edge perturbations in semantic maps constitute the primary source of distortion, while core semantic content remains recoverable within the defined noise tolerance. Collectively, these results validate the construction of minimally sufficient semantic representations and motivate the generative and robustness-driven extensions developed in subsequent chapters.

Chapter 4 Chapter 4 proposes a VQ based SemCom system for resource-constrained environments. This design validates the discrete representation principles that are the main design principles used throughout this thesis and optimizes them for lightweight inference, reduced computational footprint, and edge-cloud adaptive transmission. This chapter extends the minimal symbolic semantic representation towards a more generalized representation, based on the limitations identified in the previous chapter. A two-step codebook stabilization and task-adaptive optimization procedure is introduced to ensure consistent semantic representations under strict compute budgets. By performing semantic extraction and quantization at the edge and transmitting only compact codebook indices protected by LDPC coding, the proposed architecture achieves a 25% reduction in transmitted symbols at code rate 0.5 compared to JSCC-based systems, while reducing inference latency to 1.2–1.6 ms (compared to 10–12 ms for JSCC). The design remains channel-agnostic, operating robustly across AWGN and Rayleigh fading channels without SNR-specific retraining, while higher classification accuracy than JPEG. Experimental results demonstrate consistently higher PSNR, improved MS-SSIM, and lower LPIPS than BPG across all SNR levels, validating perceptual robustness in bandwidth-limited conditions. Generalization experiments on Tiny ImageNet, Kodak, and CelebA-HQ datasets validate scalability across diverse data distributions, demonstrating the practical applicability of discrete semantic communication for edge and IoT deployment scenarios.

Chapter 5 Chapter 5 proposes a novel contrastive-driven semantic representation learning framework that extracts compact, task-agnostic latent features for robust semantic communication for image transmission as the main contribution of this thesis. This SemCom pipeline is based on generative AI, and the chapter proposes

multi-level contrastive objectives at the pixel, latent, perceptual, and task-output levels under lossy channel conditions to learn consistent semantic embeddings. The learned latent representations are systematically evaluated to assess both machine-perceived and human-perceived semantic fidelity. Experimental evaluation demonstrates that the proposed framework consistently outperforms conventional JSCC-based SemCom systems and traditional image coding methods. In terms of perceptual fidelity, SVQ-WGAN achieves substantially lower LPIPS than BPG across all SNRs, maintaining values around 0.30 at 2 dB in AWGN (compared to 0.45 for BPG) and stabilizing near 0.30 at 4 dB under Rayleigh fading while BPG saturates around 0.55, with approximately 56.2% bits per pixel (bpp) savings. Robustness analysis using the proposed robustness index (RI) demonstrates that SVQ-WGAN maintains a higher RI around 0.8 for SNRs higher than 5 dB across coding rates, whereas BPG drops below 0.4. Downstream task evaluation confirms higher CIFAR-10 classification accuracy and improved COCO captioning scores compared to evaluated baselines, demonstrating preservation of machine- and human-perceived semantic fidelity. The framework further exhibits strong generative generalization, achieving an Fréchet inception distance (FID) of 5.496 and maintaining performance when scaled to 4K images without additional training. The theoretical justifications further validate that the quantized latent representation, index transmission, adversarial training, and contrastive regularization jointly contribute to semantic compression, preservation, and robustness under channel and semantic noise.

Chapter 6 This chapter advances the theoretical dimension of the thesis by formulating a model of semantic noise, the distortions arising not from the physical channel but from representation mismatches, quantization ambiguities, and concept-level perturbations. A novel codebook (shared knowledgebase) optimization mechanism is developed using soft assignments, adversarial discrimination, and entropy-based regularization to minimize semantic ambiguity and ensure stable symbol grounding. This chapter contributes to the objective of enabling consistent semantic representations. It proposes evaluation methodologies for analysing codebook structure, clustering

Chapter 1. Introduction

behaviour, and semantic robustness under both latent-level and pixel-level perturbations, using fast gradient signed method (FGSM)- based evaluation, thereby addressing a significant gap in semantic reliability assessment identified in the literature review. We validated higher codebook usage of 96.88% for codebook size of 64 and higher perplexity than VQ-VAE, indicating rich and balanced semantic representation across varying codebook sizes and reflecting higher perpetual fidelity with the higher codebook usage, supporting the claim that semantically optimized and representative knowledge bases support higher semantic fidelity at the receiver. Clustering the shared codebook showed that the semantic knowledge base effectively represents the underlying semantics of the data distributions and, hence, improves the semantic fidelity of the received content at the receiver. We demonstrated this concept using DBSCAN clustering. The learned codebook vectors are closely aligned with CIFAR-10 class semantics, forming compact, interpretable clusters, in contrast to VQ-VAE’s dispersed embeddings. Further, we demonstrated strong generalization on Tiny-ImageNet, maintaining reconstruction quality and semantic stability across highly diverse images.

Chapter 7 concludes the thesis by consolidating the core contributions and reflecting on the contributions of the research on SemCom frameworks. The chapter highlights how representation learning and generative SemCom collectively advance the knowledge area toward communication systems that prioritize meaning and context. The conclusion further outlines several promising directions for future research, including the development of unified multimodal semantic encoders, lightweight causal world models suitable for edge deployment, and the standardisation of evaluation metrics for semantic similarity and predictive sufficiency. Additionally, it discusses extending the presented frameworks to real-world settings such as autonomous systems, distributed sensing, and intelligent healthcare applications. Through these contributions, the chapter positions the thesis as a step toward SemCom systems for next-generation wireless networks with efficient representation learning.

Chapter 2

Literature review

This chapter reviews the existing literature, highlighting the shift from bit-level information transfer toward the transfer of meaning and intent. The discussion begins by revisiting the foundational theoretical work of semantic information and SemCom. The review distinguishes traditional communication systems optimized for bit accuracy and channel efficiency, with semantic-driven, task-oriented SemCom approaches that emphasize the task-driven nature and goal relevance. It further explores recent advances of deep learning based methodologies that have shaped this field, including the application of Generative AI models such as GANs, variational auto encoder (VAE)s, and VQ-VAEs for semantic encoding and reconstruction, methods for modeling and mitigating semantic noise, the emergence of edge intelligence for real-time, resource-efficient semantic processing. The literature review concludes by identifying key research gaps, particularly the need for generalizable semantic representations, reasoning-aware receiver architectures, robust semantic evaluation metrics, and the integration of causal temporal models based on world models.

This literature review aims to position the present research in the scholarly framework of communication theory and its more recent semantic development. The accuracy of data symbol transmission over noisy channels has long been the primary focus of classical communication systems. However, the shortcomings of such bit-centric models have been identified by recent advances [14] in AI models and the growing requirements of autonomous vehicles, intelligent vision applications such as anomaly detection, and

robotic and drone-based teleoperation [8]. Semantic communication, which emphasizes the transmission of meaning rather than raw data, has attracted attention as a way to improve communication effectiveness and reduce the burden on communication channels. The review is basically a methodical literature review that used the SCOPUS database.

There has been a significant paradigm shift in theory and practice due to switching from classical to semantic communication. Classical communication is based on Shannon’s theory [3] and emphasizes bit-level accuracy and bandwidth efficiency as determined by mutual information, channel capacity, and entropy. However, by their very nature, these models ignore the significance or task-relevance of the data being communicated. To minimize semantic entropy and maximize semantic mutual information, semantic communication systems transmit only the information required for the recipient to understand and act upon a message. Since communication frequently involves machines that shall understand, make judgments, and respond based on transmitted information, this new paradigm is fundamental in the context of 6G networks and AI-native systems.

This literature review focuses on several key components that provide the basis for creating and evaluating semantic communication systems. It starts by looking at the image and video transmission field, which presents unique challenges due to the high dimensionality and perceptual complexity of visual data. It then explores generative and VQ models, such as GANs and VQ-VAE, as methods for producing distinct, semantically rich representations that facilitate effective decoding and transmission. The integration of predictive ability into communication systems, specifically how machines may deduce, categorize, or respond to the semantic content they receive, is also included in the review.

2.1 Theoretical foundations

Traditional communication systems are based on the Shannon’s [3] mathematical foundation of classical information theory. The fundamental theoretical frameworks that support the semantic structures created in this thesis are described in this section, along

with how the fundamental ideas of Shannon’s theory are expanded to facilitate semantic communication. Shannon’s original communication model conceptualizes information transmission as a process involving an information source, an encoder, a noisy channel, a decoder, and a destination as depicted in figure 2.1. The objective is to transmit data with minimal bit errors and maximal efficiency.

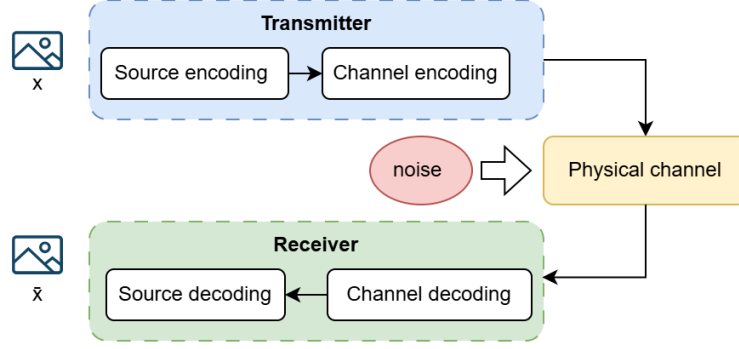


Figure 2.1: Traditional communication system.

Central to this model are three concepts: entropy, which quantifies the uncertainty or information content in a source; mutual information, which measures the amount of information successfully transmitted through the channel; and channel capacity, the theoretical limit of reliable data transmission.

Entropy Entropy is a fundamental concept in information theory that quantifies the uncertainty inherent in a discrete random variable. It is formally defined as,

$$H(X) = - \sum_x p(x) \log p(x) \quad (2.1)$$

where X is a discrete random variable representing the information source, and $p(x)$ is the probability mass function corresponding to each possible outcome x of X . The term $\log p(x)$ captures the information content associated with the occurrence of symbol x ; Entropy provides a measure of how unpredictable or informative a source is. A high entropy implies a high degree of uncertainty, indicating that the source generates symbols that are difficult to reconstruct. Conversely, low entropy suggests a more deterministic or repetitive source. From a practical standpoint, entropy sets the theoretical limit on

how efficiently the source can be compressed without loss of information, forming the basis of source coding theorems. In the context of images, each pixel value is treated as an outcome of a set of possible values, each occurring with a probability, and entropy quantifies the uncertainty and compressibility.

Mutual information Mutual information [24] extends the concept of entropy by quantifying the amount of information that one random variable contains about another. In the context of communication, mutual information $I(X; \hat{X})$ measures how much information the received signal $\hat{X}$ conveys about the transmitted signal X . Mutual information provides a measure of the output $\hat{X}$ that reduces uncertainty about the input X , and thus quantifies the fundamental efficiency of information transfer over a communication channel as:

$$I(X; \hat{X}) = \sum_{x, \hat{x}} p(x, \hat{x}) \log \frac{p(x, \hat{x})}{p(x)p(\hat{x})} \quad (2.2)$$

Here, X and $\hat{X}$ are discrete random variables representing the input and output signals of a communication system, respectively. The joint probability distribution $p(x, \hat{x})$ represents the likelihood of observing specific input-output symbol pairs, while $p(x)$ and $p(\hat{x})$ denote the marginal distributions of the input and output. Mutual information is maximized when $\hat{X}$ perfectly predicts X . Thus, mutual information provides an interpretable metric for the efficiency of information transfer in communication systems. In images, it measures the statistical difference between the transmitted and received images, reflecting how well the output retains the structure and information content of the input.

Channel capacity Channel capacity (C_I) is the maximum achievable mutual information between input X and output $\hat{X}$, optimized over all possible input distributions $p(x)$. It represents the highest data rate at which information can be transmitted over a channel with arbitrarily low error probability [24].

$$C_I = \max_{p(x)} I(X; \hat{X}) \quad (2.3)$$

This quantity determines the upper bound of reliable communication for any given channel and is the cornerstone of the Shannon capacity theorem. This result serves as a foundational element in Shannon’s information theory and underpins the development of modern coding theory and digital communication systems. In practical applications, achieving rates close to channel capacity remains a major design goal in the development of efficient and reliable transmission protocols.

Semantic entropy A key conceptual extension is the transition from Shannon entropy to semantic entropy. Shannon entropy, defined above, measures the average uncertainty of a source’s output symbols, without any regard for what those symbols represent or mean. Semantic entropy concerns the uncertainty in the meaning or relevance of transmitted messages. Let the semantic representation of a message be $S_{sem}(X)$, then semantic entropy captures the ambiguity of the meaning represented by these semantics:

$$H_{sem}(X) = - \sum_{s_I \in S_{sem}} p(s_I) \log p(s_I) \quad (2.4)$$

, where s_I is a conceptual unit of meaning and $p(s_I)$ is the probability of the semantic unit s_I appearing in the message. The s_I can be object categories in a scene, latent variables, scene concepts (direction, temporal dynamics), or any learned discrete semantic tokens as a representation.

Semantic mutual information In semantic communication systems, minimizing semantic entropy [23] becomes a primary objective, as it leads to more semantically relevant minimal information transmissions. This gives rise to the concept of semantic mutual information (SMI), which quantifies how much semantic content is shared between the transmitted message and its reconstruction at the receiver. It is defined as the mutual information between the semantic interpretations of the sent and received signals:

$$\text{SMI}(X; \hat{X}) = I(S_{sem}(X); S_{sem}(\hat{X})) \quad (2.5)$$

By focusing on $S_{sem}(X)$ and $S_{sem}(\hat{X})$, the semantic mappings of the original and

received messages SMI directly evaluate the alignment of meaning rather than symbol accuracy. Systems designed with high SMI can ensure that the receiver correctly interprets the sender’s intended meaning, even if the exact data values are distorted.

Semantic capacity The concept of semantic capacity [25] extends Shannon’s channel capacity by quantifying the maximum rate at which meaningful semantic information can be transmitted reliably between a transmitter and a receiver, and can be defined as,

$$C_{sem} = \max_{p(x)} I(S_{sem}(X); S_{sem}(\hat{X})) \quad (2.6)$$

Unlike the traditional capacity that focuses on raw bits, semantic capacity considers the shared semantic knowledge base between sender and receiver. Semantic capacity can be higher than Shannon’s capacity when meaning can be compressed by transmitting only symbolic abstractions, but it can also be problematic when misalignment in the semantic knowledge base introduces additional semantic noise. In practice, this means that semantic decoders must be equipped with prior knowledge, often in the form of pretrained models or codebooks, enabling them to reconstruct meaning from sparse or noisy inputs. This shared knowledge transforms the communication channel from a purely statistical construct to a knowledge-aware semantic link. The common knowledge is defined in terms of a common language, and it is defined by language grammar as shown in figure 2.2.

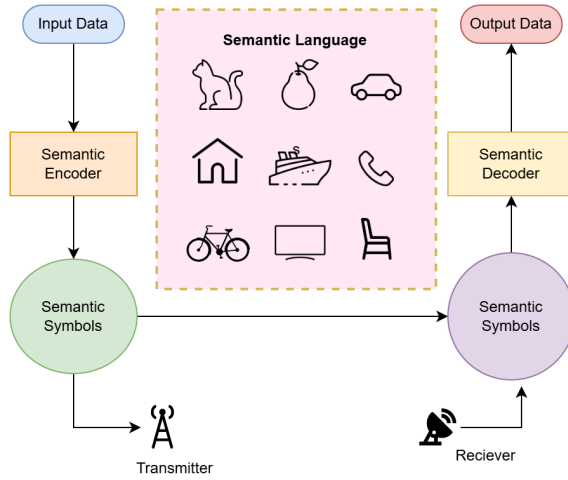


Figure 2.2: Semantic symbols and language definition.

In early research on semantic language, the study in [26] introduces the concept of a common language comprising a hierarchical set of beliefs, enabling a “speaker” to describe their observations of the environment and a “listener” to interpret those descriptions and take action accordingly. To optimise the system, the authors formulate an optimisation problem whose objective is to select the smallest set of beliefs that fully describes the observed event, while minimising both task execution time and transmission cost. To solve this problem, they develop a novel top-down framework that combines curriculum learning (CL) and reinforcement learning (RL), allowing the speaker-listener pair to gradually refine the belief set and improve task performance over time. Thus, while belief-driven goal-oriented systems are a promising first step, moving toward symbolic or codebook-driven frameworks offers a more scalable, efficient, and modular path forward, particularly in the context of next-generation wireless networks where semantic compression, shared understanding, and task relevance are paramount.

In the study in [27], the authors articulate how the design of semantic communication systems benefits from the explicit definition and sharing of a semantic language [28], a structured codebook of meaning units or semantic symbols that both transmitter and receiver internalise. The paper proposes a well-established semantic communication theory that expresses communication beyond raw bit-transmission. A semantic encoder maps input data into semantic symbols drawn from the codebook, which are then

transmitted and interpreted by a decoder operating over the same semantic language. This approach enables more efficient communication by aligning transmissions with task-relevant semantic units rather than generic bits. Importantly for codebook-based semantic communication [29], the authors demonstrate that the size, granularity, and structure of the semantic symbol set critically determine both achievable throughput and end-user reasoning capability. By modelling multi-user scenarios with reasoning-capable devices, they show how semantic codebooks must be carefully designed [30, 31] to support user inference when communication is intermittent or degraded. Although their simulations are conceptual, the focus on semantic language points toward formalising codebook-based semantic frameworks, where shared symbol dictionaries replace raw bitstreams and enable the extraction and transmission of meaning, not just data.

SemCom shall be backed by established theories on semantics. “Information Value Theory” [22] introduces a decision-theoretic framework for quantifying the value of information, arguing that information is meaningful only as it improves expected decision outcomes. Instead of treating information as a purely syntactic or statistical quantity, they treat it as something whose usefulness depends on how it influences knowledge and thereby guides action. This perspective aligns closely with semantic communication theories, which support our design decisions and emphasize that the goal of communication is not merely transmitting bits but conveying useful, decision-relevant meaning. Further, the work supports key thesis themes, including task-oriented semantic metrics, meaningful distortion measures, and knowledge-informed encoding/decoding. In SemCom, the receiver must reconstruct meaning by integrating context, knowledge, and intent rather than relying on raw signal content, which aligns with the theories in [22], where the signal acquires interpretation through contextual cues and inferred relationships. The authors treat meaning-making not as a fixed property of a symbol but as an active procedure for extracting informational value from ambiguous messages. Further, in [21], the authors argue that standard semantic theories fail to provide enough information for a system to fully determine meaning, identifying what they call the information problem. Their proposed solution is to treat meanings as constitutive semantic essences, which allow a semantic system to uniquely and transparently represent

what an expression means. This idea aligns directly with semantic communication theories in the thesis, which emphasize that communication requires transmitting “meaningful representations” rather than merely signals or syntactic content. By highlighting that meaning must be specified in a uniquely recoverable way, their approach supports core thesis ideas such as semantic fidelity, interpretability, and knowledge-centered encoding/decoding (shared knowledge base). Overall, the work justifies why semantic communication systems must model deeper meaning structures, not just surface-level linguistic or statistical correlations, to achieve reliable understanding.

Semantic rate-distortion These developments lead to a reformulation of rate-distortion theory in the semantic domain. Classical rate-distortion theory determines the minimum transmission rate required to achieve a certain level of distortion between sent and received data. Traditionally, this distortion is typically measured at the bit level, such as mean square error or bit error rate:

$$R(D) = \min_{p(\hat{x}|x): \mathbb{E}[d(X, \hat{X})] \leq D} I(X; \hat{X}) \quad (2.7)$$

Semantic rate-distortion [14] defines the distortion function in the semantic interpretation of messages:

$$R_{sem}(D) = \min_{p(\hat{x}|x): \mathbb{E}[d(S_{sem}(X), S_{sem}(\hat{X}))] \leq D} I(X; \hat{X}) \quad (2.8)$$

This framework enables the system to tolerate substantial data distortion provided that the semantic outcome remains intact. For example, a highly compressed image may exhibit significant visual degradation yet still suffice for an object detection task, thereby resulting in low semantic distortion.

Semantic communication theory represents a paradigm shift from classical symbol-centric information theory to a meaning-based approach to systems. Entropy, mutual information, and rate distortion are interpreted by the semantics, reasoning, and shared knowledge. These are the basic theoretical building blocks that form the basis for future intelligent communication systems that are channel efficient and transmit only “what is

needed”.

2.2 Semantic communication

The advancement of 6G wireless networks necessitates a paradigm shift in the design of wireless networks from data-centric to goal-oriented and semantics-driven communication frameworks. The authors in [32] highlight this transition by positioning artificial intelligence as the core enabler of intelligent, context-aware wireless networks. They argue that, beyond throughput and latency, future systems must convey meaning to support applications such as autonomous agents and immersive experiences, laying a conceptual foundation for semantic communication. Further, the survey in [16] formalizes semantic communication by introducing a comprehensive framework encompassing semantic information extraction, transmission, and effectiveness evaluation. It highlights the limitations of conventional channel coding in task-oriented scenarios and the use of matrices of semantic effectiveness that align transmission with communicative intent, which is an essential evolution for AI-native networks. Complementing these theoretical advancements, the article [33] situates semantic communication within intelligent wireless systems, emphasizing architectures that support distributed learning and low-latency inference, particularly within the broader context of edge computing, machine learning, and distributed decision-making. Their survey underscores the inadequacy of conventional bit-level transmission models for scenarios involving resource constraints, dynamic environments, and real-time inference, particularly within the broader context of edge computing, machine learning, and distributed decision-making. A notable contribution of the paper is its detailed taxonomy of semantic communication layers and its proposal for modular system design, which aligns with the hierarchical structures needed for reasoning and abstraction in communication.

In view of the development of SemCom, this thesis identifies several architectural and methodological design paradigms discussed in the literature that significantly influenced the selection of the models’ design paradigms presented in the subsequent technical chapters. Mainly, the literature review has analysed the following design paradigms illustrated in figure 2.3.

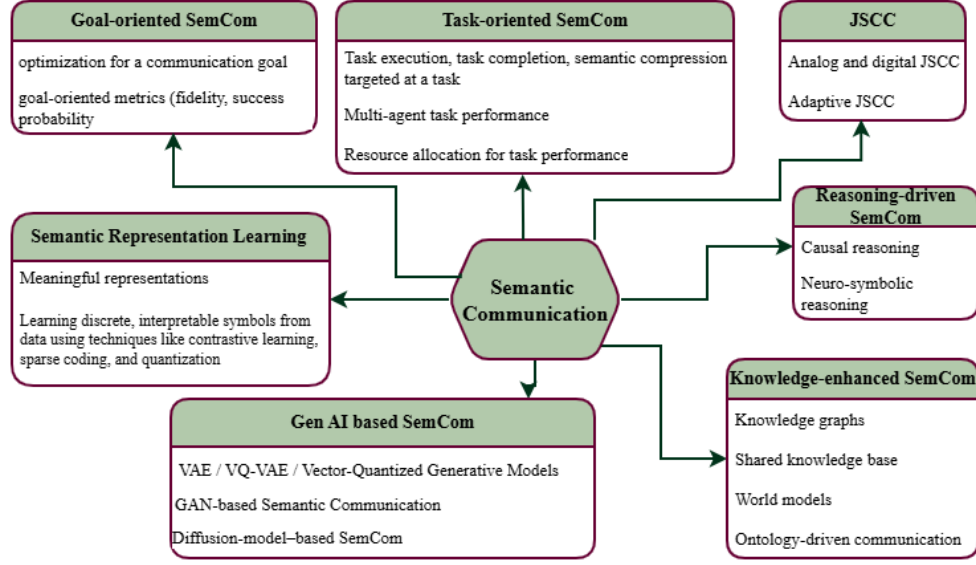


Figure 2.3: Literature review flow

2.2.1 Knowledge enhanced semantic communication

Knowledge-driven semantic communication leverages structured or learned knowledge, such as knowledge graph (KG), shared knowledge bases, domain ontologies, or large AI models, to guide semantic extraction, transmission, and recovery. Early conceptual and survey articles in this area of research [4, 34–36] clarify that knowledge is a core resource for moving from traditional bit-level reliability to meaning-based reliability. The work in [37] uses a KG of semantic triples at the base station, where users employ an attention-based reinforcement learning to select which triples to transmit under resource constraints. Similarly, the authors in [38] encode sentences into importance-ordered triples and adjust transmission based on channel conditions to enhance reliability in low-SNR regimes. In [39], the authors propose an incremental knowledge-based hybrid automatic repeat request (HARQ) mechanism with adaptive bit lengths and a denoiser, jointly reducing semantic errors and communication cost [39]. These knowledge-centric schemes provide interpretable semantic units and explicit control over what is transmitted. However, they depend on carefully constructed, task-specific knowledge and may struggle to scale to highly diverse or general domains.

A separate research direction is knowledge-enhanced transceivers. Accordingly,

in [40], where the authors design a text SemCom system with a shared, human-readable knowledge base (KB) and introduce semantic self-information and entropy measures, enabling the transmitter to send only residual information relative to the KB, reducing transmitted symbols without degrading semantic quality. The article [41] develops a cognitive SemCom system driven by KGs for single- and multi-user scenarios, with interpretable semantic alignment and correction algorithms and a software-defined radio prototype. Considering the receiver side knowledge enhancement, the research in [42] uses a transformer-based knowledge extractor that significantly improves decoding performance while leaving the transmitter architecture unchanged. For image transmission, the authors of [43] propose a domain knowledge-driven SemCom system, which is a dual-path encoder–decoder that incorporates medical domain knowledge to better preserve clinically relevant concepts under channel variations. The study [44] uses a shared semantic knowledge libraries and a diffusion generator to reconstruct images from semantic labels, together with a semantic fidelity index that measures mutual information and feature similarity.

These designs achieve strong semantic robustness and interpretability in their target domains. However, they often rely on domain-specific knowledge or labeled libraries, which limits the generalizability of context and increases the cost of KB maintenance in systems. More recently, development in this area has focused on the use of large models and a generative AI based knowledge prior as a more flexible and general alternative. The reference [45] proposes a large-scale AI model–based SemCom framework for images that uses a KB of semantic segments generated by the segment anything model (SAM). Using this KB, the model integrates image segments via attention-based semantic aggregation and applies adaptive semantic compression to reduce redundancy in image transmission. The authors of [46] extended this idea to multimodal data by combining multimodal language models for cross-modal alignment, large language model (LLM)-based personalized knowledge bases to resolve semantic ambiguity, and GAN-based channel estimation to combat fading. However, the knowledge provided by large-model approaches faces challenges related to computational cost, model hallucinations, security [47], and adoption in resource-constrained or secure environments [48]. Overall,

knowledge-driven semantic communication has demonstrated clear gains in semantic fidelity, interpretability, and robustness, but current systems remain constrained by the quality, scalability, generalization, and trustworthiness of their underlying knowledge sources.

2.2.2 Goal oriented semantic communication

The main intention of goal-oriented semantic communication (GO-SemCom) is to transmit the relevant information to achieve an ultimate goal, rather than reconstructing or generating the source information or preserving the perpetual fidelity of the transmitted data. A deep analysis of this is done in the conceptual works and foundational survey articles in [4, 5, 49–53]. The design is optimized to achieve a particular goal by embedding task utility, decision impact, and effectiveness metrics into the transceiver design. In other words, the training objective is shaped by the goal being achieved, with the promise of achieving bandwidth-constrained and latency-critical application-specific transmission. We identified the absence of unified mathematical frameworks, standardized metrics, and scalable modeling tools for goal-driven tasks in the literature.

The work on task-aware rate–distortion and fidelity modeling described in [54] is a theoretical approach to studying goal-oriented rate–distortion under intrinsic and extrinsic semantic constraints. They show that the optimal encoding is highly sensitive to task-dependent fidelity choices. The reference [55] integrates goal-oriented metrics with distributed edge learning. They introduce a mathematical formulation of goal-oriented source entropy, treating semantic relevance as an optimization variable intertwined with learning dynamics. These works remain largely conceptual and do not specify how such metrics should be instantiated in practical environments.

A separate research direction of GO-SemCom has introduced application-oriented solutions for real-time control and immersive applications. Accordingly, [56] introduces an age and value of information [57] framework for robotics. The authors use deep reinforcement learning to proactively schedule control-critical packets. The design captures the temporal utility of semantic updates and improves task success, but it

depends heavily on accurate models of task dynamics and environmental variability. In a similar direction, the research in [58] uses GO-SemCom for robotic-arm reconstruction in digital twins with goal-driven feature selection. Although this demonstrates improved control accuracy, it requires domain-expert supervision.

In visual communications, GO-SemCom has advanced toward task-aware generative and motion-guided encoding. The work in [59] employs an optical-flow-based autoencoder to extract goal-relevant motion cues. It reduces redundancy while maintaining action-level fidelity, but the video transmission performance degrades under highly irregular motion patterns. For image transmission, the study in [60] applies stable diffusion models to reconstruct perceptually and task-relevant outputs from sparse semantic representations. While enabling high semantic fidelity, diffusion-based systems are computationally intensive. Furthermore, the authors in [61] address avatar-centric augmented reality by integrating point-cloud semantics into end-to-end transceivers, achieving strong perceptual fidelity but relying on domain-specific semantic priors tied to the specific metaverse environments. Finally, the research in [50] presents a GO-SemCom framework that represents data types, semantic extraction strategies, recovery methods, and task-effectiveness metrics across XR, unmanned aerial vehicle (UAV) control, and IoT systems. The framework remains conceptual and lacks standardized quantitative measures.

Overall, existing GO-SemCom demonstrates substantial gains in communication efficiency, task reliability, and robustness by exploiting goal-oriented relevance, rather than raw information fidelity. Theoretical contributions define the limits of semantic effectiveness, while practical systems validate these ideas in robotics, XR, and autonomous control. However, current designs often rely on domain-specific priors, handcrafted semantic features, or computationally heavy generative models, limiting scalability, generalizability, and adoption in real-world 6G deployments.

2.2.3 Task oriented semantic communication

Task-oriented semantic communication (TOSC) differs from goal-oriented approaches by optimizing transmission for a specific downstream task such as classification, retrieval,

visual question answering (VQA). Although both look similar, goal-oriented schemes emphasize final objectives such as decision accuracy, reward, or task value, irrespective of task structure, and may abstract away even the representation itself. In contrast, TOSC systems retain the task model and align semantic feature extraction, compression, and resource allocation to preserve task-relevant information under channel and other transmission constraints.

One research direction is task-aware JSCC architectures, where the study [62] introduces a channel-adaptive digital JSCC framework for image-based tasks. They have modeled demodulation uncertainty in their robust training strategy and have supported diverse channel conditions and modulation orders. The methodology achieved strong performance in classification, reconstruction, and retrieval, but at the cost of increased training complexity and reliance on an accurate channel model. The reference [63] also uses deep JSCC, using a joint loss and a gated module for channel-aware feature pruning. This has increased complexity in design, but achieves multi-task robustness. To reduce overhead across diverse tasks, the novel study in [64] proposes a unified multi-task semantic system with dynamic feature selection and a shared codebook.

A complementary line of research focuses on resource and user level optimization for TOSC. In their research [65], the authors propose adaptable semantic compression with joint resource allocation. This methodology results in reduced data and increased task success, relying on accurate task-performance models. In the model defined in [66], the methodology dynamically allocates power bandwidth and compression ratio based on semantic importance, improving task accuracy at the cost of increased training time. The study in [67] integrates user-interest saliency and game-theoretic resource allocation for UAV image tasks in SemCom to improve personalization, but this increases model complexity. Further, models such as [68, 69], which are multi-user, multi-task semantic models, demonstrate strong robustness and task performance but rely on large, computationally intensive models.

Overall, TOSC advances semantic communication by tightly coupling and aligning the transmitted features with task requirements. Its main limitations arise from task-specific training demands, model complexity, and limited generalization across tasks, in

contrast to the broader adaptability sought by SemCom.

2.2.4 Joint source channel coding (JSCC)

JSCC based SemCom optimizes the source coding and channel conditions jointly. The early implementations of JSCC [70, 71] using deep learning techniques consider the joint optimization of source coding and channel optimization, where they directly map the source symbols to channel symbols. These techniques have the disadvantage of difficulty in adapting to heterogeneous channel conditions. However, adoption of the deep JSCC into SemCom is extensively studied in the literature, trying to address the problem of dynamically adapting to channel conditions, avoiding retraining to various channel conditions. Recently, deep JSCC techniques have been extensively used in studies to optimize SemCom pipelines due to their ability to avoid “clif effect” and adaptability to channel conditions. In [72], the authors have used a transformer-based model that embeds the SNR estimates into attention layers to enable lightweight channel adaptation. However, this is also suboptimal when channel statistics deviate significantly from the estimates used in the training or when the feedback is unreliable. The authors of [73] design an asymmetric decoder to adapt to channel conditions and heterogeneous downstream tasks with an SNR adaptive unit to enhance the adaptability to SNR ratios. The model is a masking-based transformer to combat channel distortions, but it increases implementation complexity in resource-constrained environments. Further, the results are obtained from low-resolution datasets, and the scalability is not assessed. The study in [74] explores adaptive encoder-decoder architectures that adapt to channel conditions, such as channel SNR, compression ratio, device capabilities, and power constraints. They used a sequential randomized layer selection for training. However, the experimental results are mainly based on pixel-centric evaluations, based on PSNR, and real-time layer reconfiguration may introduce system coordination challenges and risks of instability in highly dynamic environments. Specifically, in task-oriented settings, the study in [75] proposed an adaptive compression and attention-based fusion for efficient multi-image processing in vehicular networks. The models are designed to adopt compression based on the importance of semantics, which is specific to the task.

Although the model is designed for vehicular networks, the evaluation dataset is the CIFAR-10 dataset, which is primarily developed for classification tasks and is relatively low in diversity. The evaluation is suboptimal given the complexity of the task scenario. The research in [76] proposes multi-channel feature fusion to enhance robustness by adapting to various wireless channel conditions and compression ratios. Specifically, the methodology involves a feature enhancement module, which combines f-channel attention and spatial attention mechanisms to adaptively adjust the f-channel weights of the semantic feature map under various SNR conditions, but they often introduce additional architectural burdens, or the assumptions on the channel conditions and feature importance may limit generalizability in practical deployments. A separate research direction explores digital JSCC, in contrast to the traditional analog JSCC, which limits the compatibility with digital communication systems. To bridge this gap, the authors of [77] propose the D²-JSCC framework. It introduces a digital deep JSCC architecture that jointly optimizes deep source coding and channel coding via end-to-end distortion modeling. The technique mitigates the “cliff effect” and “leveling-off effect” observed in classical separation-based pipelines. The system benefits from explicit rate control and digital reliability, yet trades off the graceful degradation property of analog JSCC.

JSCC provides a promising direction for SemCom based systems, advancing from fixed analog deep JSCC architectures to flexible, digital-compatible, and adaptive transmission schemes. While recent works offer improved channel adaptivity and robustness under resource constraints, challenges remain in computational efficiency, real-time adaptivity, and integration with practical digital communication standards. These limitations highlight the need for efficient architectures capable of operating across diverse semantic tasks and dynamic channel environments in practical settings.

2.2.5 Generative models in semantic communication

Visual generative models have recently emerged in SemCom, enabling systems to move beyond reconstruction and generate the transmitted meaning with minimal symbols. The systems presented in the current literature leverage vector quantization, adversarial

learning, and transformer-based generative architectures to encode compact, semantically rich representations that remain robust across noisy channels. This section reviews key current work on visual generative AI driven SemCom, assessing their strengths, limitations, and possible design cues for improvements.

Several recent works have explored vector-quantized semantic communication. Authors of [78] apply VQ-VAE in a noisy channel setting with limited reconstruction tasks using the MINIST dataset with limited domain-level generalization and modeled using a theoretical discrete memoryless channel (DMC) channel. The methodology presented in the research of [79] employs a PatchGAN discriminator to focus on local texture realism, such that patch-level training lacks global semantic awareness and may lead to unstable training or poor generalization. A transformer-based model is used in the work proposed in [80], which is comparatively highly complex. The pixel-level and semantic-level results are comparatively low compared to the much less complex dataset FFHQ, where generalization across domains is not considered. Using LLM to fuse the text-based symbols with the visual feature symbols to generate meaning at the receiver, mitigating channel noise, is being employed in recent works. The work in [81] quantifies the semantic importance of messages using pre-trained LLMs and introduces a semantic error correction mechanism based on bidirectional encoder representations from transformers (BERT), which effectively mitigates semantic loss by reconstructing missing or corrupted content. However, the accuracy of these methodologies is not being evaluated in relation to the LLM prompt accuracy and inherent limitations of LLMs, such as hallucinations.

Generative AI techniques, particularly those based on deep neural networks and GANs [82], have significantly advanced the field of SemCom. Integrating generative models, such as GANs, into SemCom frameworks has improved data quality by preserving high-level semantic features, even under low SNR conditions. These models can synthesize or recover missing or distorted parts of transmitted signals, enhancing the perceptual quality and semantic integrity. Table 2.1 reviews the latest advancements in generative AI-based SemCom frameworks, highlighting their methodologies, strengths, and challenges.

Table 2.1: Summary of recent advancements in generative AI-based SemCom systems. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

Ref	Data type	Methodology	Strengths	Limitations
[83]	Image	Employs semantic segmentation to classify pixels by prioritizing regions of interest (ROI) to reduce overhead	Significantly reduces communication overhead, efficiently uses available bandwidth	Increased computational demands, requires fine-tuning for optimal operation across various image types
[84]	Image	Inverse JSCC and Generative JSCC	Enhances perceptual quality, outperforms traditional Deep JSCC	JSCC adaptation to practical channel characteristics
[85]	Image	Lightweight Generative Joint Source-Channel Coding (LGJSCC)	Low computational demands using compressed generator	Reliance on conditional GANs
[86]	Image	Uses GANs to prioritize ROI, reduces region of non-interest (RONI) quantization SNR during transmission	Reduces data volume by 60.53%	Increased complexity, challenging real-time implementation
[87]	Text	GAN-based Ti-GSC framework integrating autoencoder-based modules to suppress signal distortion without CSI	Boosts BLEU scores by 40%–62% in fading channels	High complexity, performance sensitive to fading types
[88]	Image	Uses denoising diffusion model	Lower complexity compared to traditional channel codes	Study limited to low-resolution MNIST dataset
[89]	Image	Vector Quantization improved with GAN and spatial conditional normalization	Enhances image reconstruction quality using multi-channel indexing and spatial normalization	Limited dataset variation (animal faces), potential overfitting
[79]	Image	Convolutional encoder-decoder with PatchGAN discriminator	Compatible with digital communication systems	Training limited to Cars196 and Kodak datasets, risks overfitting, domain specificity
[90]	Image	Masked VQ-VAE with ViT blocks, Feature Importance Module (FIM) to filter task-relevant features	robustness against semantic noise, compatible with digital systems	Task-oriented training, added complexity due to transformer blocks

2.3 Semantic representation in next-generation communication systems

Representation learning aims to create meaningful data representations, which can be applied to tasks such as classification, segmentation, and prediction [91, 92]. Its key advantage lies in data compression, which improves the generalization of underlying information, allowing SemCom systems to be independent of the dataset in which

they are trained. The authors in [93] propose a framework using identical vectors for nodes in a parse tree, demonstrating efficient parsing of images with varied structures. Furthermore, recent studies have emphasized the role of disentangled representations in achieving robust feature extraction, as shown in works [94] and [95]. These studies leverage VAEs to generate structured latent spaces that effectively capture essential data features while minimizing redundancy. Additionally, the proposed methodology in [96] uses deep convolutional GANs, demonstrating how learned representations can be leveraged for downstream tasks such as image synthesis and classification. In SemCom, the frameworks above can offer valuable insights into optimizing latent space structures to minimize communication overhead while maintaining data fidelity.

2.3.1 Representation learning in semantic communication

Semantic representation is the cornerstone of effective and robust SemCom, as it determines which information is extracted, compressed, and transmitted by semantic encoders. The paradigm shift from bit-level recovery to meaning-level semantics is highlighted in early research on 6G architectures [97] and a more advanced development in recent causal-reasoning perspectives that frame semantics as structured, intervention-aware representations rather than opaque features [98, 99].

In the first paradigm, the design focuses on learned latent representations from the source message. The study in [100] proposed a representation framework. The model is self-supervised, using covariance-based contrastive learning to extract goal-invariant dense semantic representations for image transmission. This model is not dependent on labeled data or joint transmitter-receiver training, but it improves the downstream classification accuracy under varying SNRs. However, it may suppress fine-grained features for specialized tasks when contrastive learning is applied at a single level. Further, the study in [101] proposes a classification-driven architecture that combines contrastive learning, sparse coding, and neural index quantization to produce both continuous and symbolic discrete representations. This model is tailored for classification and offers interpretability and strong compression while maintaining classification performance. Since it is tightly classification-driven, it may require

fine-tuning or retraining for different tasks or domains. In [102], the authors propose hierarchical variational autoencoders for RL, with the intention of balancing compactness and sufficiency in SemCom in partially observable Markov decision process (POMDP) based control frameworks. The model effectively encodes semantics in observations related to the control process’s decisions, demonstrating good sample efficiency and control performance. However, hierarchical latent design increases modeling and training complexity.

The second line of work in the current literature exploits structured and reasoning-aware representations. This paradigm encourages the SemCom to shift from data or knowledge-driven to reasoning-driven. Knowledge graph based semantic encoding is introduced in [103]. The semantic encoder represents the messages as node–edge–node triples and node embeddings, and the decoder reconstructs them, achieving high compression. However, it depends on accurate KG construction and may be brittle when the underlying knowledge is incomplete or misaligned, therefore, a lack of generalization. The work in [97] represents the message as symbolic units grounded in an application-specific knowledge base, which enables semantic compression and cross-lingual robustness. The next frontier of representation learning for next-generation wireless networks is causality-driven frameworks. The study in [99] proposes causal representation learning and signaling games to obtain representations that support reasoning, counterfactuals, and semantic capacity notions. The work provides a strong conceptual framework, but remains largely theoretical, with no implementation details. The authors in [104] illustrate a more engineering-oriented perspective, where lightweight encoder–decoder blocks are tailored for semantic representation while jointly supporting encryption and adversarial detection. The model addressed the robustness and deployability issues in mobile environments.

The representation methodologies developed for SemCom fall into neural codes, emphasizing task efficiency and compression, knowledge- and causal-based methods prioritize structure and reasoning, and lightweight designs focus on practical robustness. Open challenges remain in integrating these strategies to develop practical, generalizable semantic representation methods that provide compact, task-general, and deployable

solutions in realistic wireless environments.

2.3.2 World models

In the current literature, the concept of world models has advanced its conceptual definitions. The authors of [105] provide a comprehensive survey that divides world models into two categories. First, construct implicit representations to understand the external world; in technical terms, represent the environment as latent variables. The second is to predict the future state of the external world; in technical terms, a computational framework that simulates and predicts future world states by analyzing past observations and potential actions within a visual context. They also highlight possible application domains such as image reconstruction, autonomous driving, and video synthesis. The survey presents a taxonomy of contextual understanding of spatial relationships, in the first category, using CNNs [106] and vision transformer (ViT) [107,108]. They present video prediction models as a plausible way to reason and predict about the future states of the visual world. While the survey is thorough in its scope, it primarily emphasizes classification, with limited attention to empirical comparisons, unified benchmarking, or practical implementations.

The study on world models presented in [109], the authors introduced a three-component architecture. A VAE used for latent-state representation, a recurrent network to learn the environment dynamics, and a controller acting purely in latent space, to demonstrate that controlling the internal environment can be effectively done within an internal generative model of the world. Their next advancement of this model [110] further shows that recurrent world models accelerate policy evolution by enabling agents to learn behaviors entirely within the learned latent environment, illustrating the power of abstract predictive modeling for efficient decision-making. These works established the core principle that compact semantic abstractions, capturing the structure and dynamics of the environment, can be used to build models that accurately predict behaviours without using any sensory data collected from the environment. Recent research extends world models toward robustness, multi-view perception, and embodied intelligence. Masked reconstruction is used in the study [111] to develop multi-view

masked world models, using input from multiple camera viewpoints, by learning a latent representation invariant to viewpoint. This closely resembles how humans learn to model the world from input from the eye’s sensory system. In a further development, the authors of [112] consider integrating physical simulators and world models for embodied intelligence, emphasizing that learned world dynamics can generalize across tasks, modalities, and environments.

In summary, these developments in world model research highlight two main characteristics that are important to adopt in future SemCom systems: transmitting only the meaningful aspects of information needed for downstream tasks rather than reconstructing raw data. World models can produce such compressed, semantically rich latent representations to achieve:

1. **Representation alignment** The emergence of compact, task-relevant latent states
2. **Predictive semantics** The ability to model environment dynamics at a high level of abstraction.

Furthermore, multi-view and embodied world models illustrate how diverse sensory modalities can be fused into a unified latent space mirroring multimodal semantic communication in which devices exchange abstractions rather than raw signals. Thus, world-model based research offers a conceptual and technical foundation for designing future SemCom systems in which communication, perception, and prediction are jointly optimized within a shared latent semantic space, as illustrated in figure 2.4.

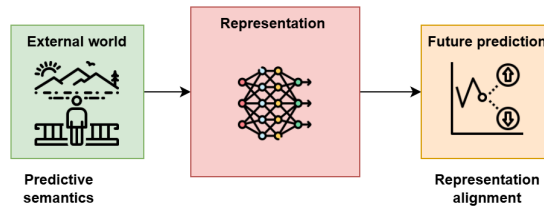


Figure 2.4: World-model pipeline illustrating the construction of implicit latent representations from visual observations and the prediction of future states through learned dynamics. In a practical sense, the semantic encoder compresses raw image frames into compact latent variables that capture semantics, while the dynamics model simulates prospective environment evolution conditioned on past states or actions.

2.3.3 World models and next generation communication

In the context of communication, such predictive internal models can transform semantic encoders into reasoning-aware transmitters. By modeling temporal dependencies, a semantic encoder may anticipate the evolution of visual or symbolic semantics and transmit only latent updates that are causally relevant to downstream tasks. In [113], authors introduce a causal reasoning and conceptual learning architecture that enables transmitters and receivers to communicate abstract concepts and infer latent causal relations. Authors of [98] extend this idea to system-level design, positioning causal reasoning as a paradigm for next-generation AI-native wireless networks and emphasizing its role in explainability and robust decision-making. Time-series joint embedding predictive architecture (JEPA) [114] advances this direction in a practical setting for remote control under constrained bandwidth, demonstrating how predictive latent representations can emulate world-model behavior in communication, but with a simulated dataset, which has motivated the last chapter of this thesis to develop world-model-inspired SemCom with a real-world dataset to conceptualize the idea. Together, these studies illustrate the practical movement of semantic communication research toward causality- and reasoning-aware paradigms.

In an overall evaluation on the SemCom design strategies, JSCC-based schemes provide robust transmission but remain pixel-centric, whereas object and task-driven designs enhance semantic efficiency at the cost of generality. Generative and multimodal frameworks enhance perceptual reconstruction; however, in video-based acSemCom, the current literature primarily relies on interpolation techniques [115], which lack causal dynamics. A remaining gap lies in modeling procedural and causal dynamics within semantic video transmission to support reasoning-aware intervention-based applications in communication.

2.4 Semantic noise mitigation

Semantic noise, arising from mismatches in meaning rather than symbol errors, has emerged as a central challenge in SemCom systems. Recent research has

focused on designing architectures that can detect, model, and suppress semantic distortions stemming from channel impairments, adversarial perturbations, multimodal inconsistency, knowledge mismatch, or ambiguity in learned semantic representations. Early studies explored codebook-based robustness, while more recent systems leverage adversarial training, knowledge graphs, uncertainty modelling, and attention-driven correction.

A prominent research direction investigates codebook-driven semantic robustness. The study in [90] introduced a masked VQ-VAE model with a codebook and adversarial training, with a feature importance module to suppress noise-related features. This approach provides strong robustness and efficient transmission via discrete symbols but requires accurate noise modelling and incurs training complexity. The results are primarily based on classification accuracy. In [116], the authors extend VQ-based designs for semantic security, adding semantic noise intentionally to confuse eavesdroppers while improving robustness through channel state information (CSI)-aware adaptive noise injection. While effective for secure transmission, this method introduces additional uncertainty at the receiver and depends heavily on reliable codebook coordination. Another stream of work employs knowledge-driven semantic restoration. Accordingly, [117] proposes a knowledge-graph-assisted system for text transmission, where shared symbolic relations guide semantic reconstruction under noise. The research in [118] further enhances this paradigm by combining knowledge graphs with adversarial BERT training and incorporating semantic noise datasets. The model achieves robust restoration against linguistic, adversarial, and channel noise. These knowledge-centric systems demonstrate strong robustness against noise but rely on high-quality, domain-specific knowledge graphs and can suffer from scalability limitations in dynamic environments. A study of image and video-based semantic transmission is to be conducted.

Several works target multimodal systems that can address semantic degradation across modalities. A graph neural network (GNN) based multimodal semantic communication framework is used in [119], combining graph convolution, graph attention, cross-attention, and long short-term memory (LSTM)-based correction to

mitigate multimodal semantic and physical noise. This approach excels in complex multimodal inference tasks but is computationally intensive due to stacked graph and attention layers. Uncertainty-aware approaches have also been investigated. The study [120] models semantic ambiguity using probabilistic uncertainty coding, where semantic features are treated as Gaussian distributions and optimized via the information bottleneck principle. Although it supports multi-view semantic predictions and improves robustness, it requires careful tuning of variance–information tradeoffs and may complicate downstream tasks. System-level noise characterization has been examined in [121], which studies the performance limits of DeepSC under multi-interferer RF interference and shows that semantic corruption grows with interference power. They propose a lifelong learning-based interference-resistant SemCom design, but further architectural refinement is needed for practical deployment. A broader perspective is provided in [122], which surveys semantic noise types and mitigation strategies, such as deep semantic restoration, knowledge-driven detection, and semantic HARQ, and offers a taxonomy but does not introduce new architectures. Further, the study in [123] proposed an autoencoder-based semantic noise metric that compares latent vector spaces and achieves high correlation with subjective assessments, while valuable for benchmarking semantic noise, it does not inherently mitigate noise.

Domain-specific semantic noise mitigation has also been studied. In [124], the authors proposed a video semantic communication system integrating channel/spatial attention, a feature refinement module, and a U-Net-based semantic correction module. Although effective against combined physical and semantic noise, its multi-stage pipeline increases computational load. The work in [125] presented a modular denoising approach that combines GAN-based source denoising with channel-level autoencoder correction, improving robustness at low SNR but relying on separate modules rather than a joint optimisation strategy. The authors of [126] addressed semantic noise caused by LLM hallucinations by proposing a cross-domain fusion codec with hallucination filtering for task-oriented semantic communication. However, its effectiveness depends on reliable LLM generated knowledge.

Overall, the literature demonstrates significant progress in combating semantic noise

across security-aware, knowledge-driven, multimodal, probabilistic, and attention-based SemCom systems. However, existing strategies exhibit limitations, including high computational cost (GNN/Transformer-based systems), reliance on specialized knowledge resources (knowledge graph-supported designs), sensitivity to codebook accuracy (VQ-VAE systems), and limited generalization under unseen semantic perturbations. These gaps underscore the need for novel, efficient semantic noise mitigation frameworks that can handle complex real-world semantic distortions across diverse communication scenarios.

2.5 Deep learning techniques

This subsection outlines the deep learning models and techniques used in this thesis, specifically the applications of the tool across various application domains and its applicability to SemCom. The technical details of the techniques are discussed in detail, along with the features of these deep learning techniques that align with the design considerations presented in this thesis. These are essentially efficient representations, generalizability, robustness, and bandwidth efficiency.

2.5.1 Generative adversarial network

One of the most prominent deep generative models in the literature is the GAN, first introduced in [127]. GANs consist of two components, the generator and the discriminator. The generator tries to match the ground-truth distribution by minimizing a divergence between the generated image distribution and the true data distribution. The discriminator is designed to distinguish between the real and generated samples. The gradients are then backpropagated into the generator during training to adjust its weights, guiding it to generate samples that closely resemble the real distribution. As a result, it is difficult for the discriminator to classify the generated samples as fake. This adversarial training process enables the GAN to produce high-quality image synthesis. GANs are used in several areas, such as image synthesis [127, 128], super-resolution [129, 130], image-to-image domain translation [131, 132], and data

augmentation [133, 134]

GANs have various application domains due to their ability to generate visually realistic samples from even noisy content. The robust generative priors can fill in missing details to ensure perpetual quality even in highly noisy content, which makes GANs a powerful tool in SemCom. However, they have limitations, such as mode collapse and training instabilities, which motivate the direction of research in SemCom architectures that offer more interpretable representation learning frameworks.

2.5.2 Contrastive learning

Contrastive learning is a self-supervised representation-learning paradigm that trains models by pulling semantically similar samples closer together in the latent space while pushing dissimilar samples apart. Instead of relying on labeled data, contrastive methods create positive pairs from different views or augmentations of the same input and negative pairs from different inputs, enabling models to learn discriminative, task-relevant features from raw data. The paradigm of contrastive learning was initiated by contrastive predictive coding introduced in [135], which demonstrated how mutual information based objectives could extract high-level representations from audio and visual data. The SimCLR framework [136] is a contrastive self-supervised learning framework designed to learn image representations by maximizing agreement between differently augmented views of the same image. Further advances, such as MoCoV3 [137], present a contrastive learning framework designed for self-supervised training of vision transformers using instance discrimination. Its architecture focuses on learning feature embeddings for downstream classification tasks by maximizing agreement between augmented views of the same image and minimizing agreement with others. The method operates under clean, static offline conditions and does not address communication-specific challenges. The methodology involves two augmented views of the same image, which are passed through an encoder and a momentum encoder. A contrastive loss is then minimized to align positive pairs and repel negatives. All the above contrastive learning methodologies are designed for representation learning in clean, static environments using continuous embeddings and do not address issues related to quantization, channel noise,

or semantic degradation in communication systems. There is a growing requirement to develop robust SemCom systems, and contrastive learning is being identified as a potential solution. Figure 2.5 illustrates the contrastive learning concept.

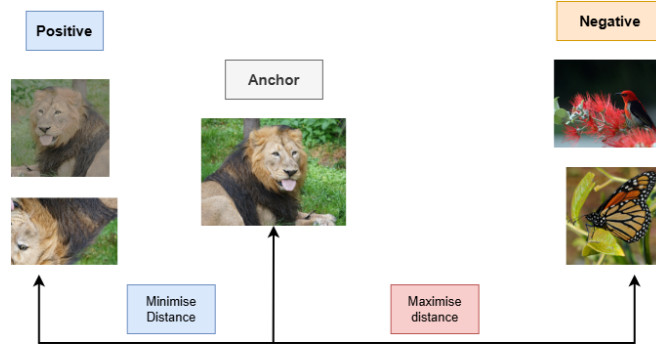


Figure 2.5: Contrastive learning: make the positive pairs closer and push negative pairs apart.

Contrastive learning has become a principal technique for obtaining robust and discriminative semantic representations in modern SemCom systems. Early work [138] uses contrastive instance and cluster discrimination to disentangle learnable from memorizable data, enabling minimalist semantic representations. Its strength lies in semantic clustering and representation. However, the method does not directly address physical-layer channel impairments or multimodal integration, limiting its applicability to real-world settings. For wireless image transmission, article [139] introduced a contrastive learning-based JSCC framework treating channel corruption as data augmentation. Their two-stage optimization improves both downstream task accuracy and reconstruction quality. However, the approach remains single-modality and task-specific, restricting generalization. The authors extended the idea [140] through semantic contrastive coding and semantic re-encoding, enabling encoder training without access to downstream models while balancing semantic fidelity and detailed reconstruction. Despite substantial empirical gains, especially under tight bandwidth constraints, the framework suffers from high training complexity and focuses primarily on vision tasks.

To support distributed and personalized SemCom, the authors in [141] proposed federated contrastive learning, which avoids model aggregation and uses a contrastively

trained semantic centroid generator to mitigate data heterogeneity. This achieves robustness across clients and low-SNR channels, but the reliance on federated infrastructure increases system overhead and may constrain deployment in lightweight devices. With growing interest in multimodal SemCom, the study in [142] introduced a system that integrates a multimodal cross-attention module and dual-contrastive learning to handle redundancy and conflicts in heterogeneous inputs. The model is expandable and performs well across multiple tasks, but its architecture is comparatively complex and relies heavily on accurate modeling of modality interactions. Similarly, authors of [143] proposed a unified multimodality-oriented SemCom framework that first harmonizes modalities via symmetric encoder–decoder structures and then aligns them using contrastive learning. This design enhances flexibility and supports a wide range of modality–task combinations. However, the reliance on a unified representation may obscure fine-grained modality-specific features and impose substantial computational overhead.

Overall, contrastive learning has shown substantial benefits in SemCom, including semantic separability, noise robustness, and improved downstream task performance. However, existing approaches differ significantly in focus, ranging from semantic minimalism to robust unimodal transmission, distributed personalization, and multimodal unification. Common limitations include sensitivity to the construction of positive/negative pairs, high training complexity, limited domain generalization, and challenges in aligning contrastive objectives across diverse semantic tasks. These gaps motivate the development of more efficient, adaptive, and task-aware contrastive SemCom systems.

2.5.3 Vector quantized variational autoencoder

Probabilistic generative models are powerful tools that generate data by sampling from learned prior distributions. The VAE introduced in [144] is a probabilistic generative model that learns a latent representation by jointly optimizing an encoder, a decoder, and a variational objective. Following the training objective of maximizing the evidence lower bound (ELBO), which encourages the latent space to follow a prior distribution,

the encoder approximates a posterior distribution over the continuous latent variables, and the decoder models the conditional likelihoods that generate new samples. The limitations identified in the existing body of knowledge are the Gaussian assumption of the prior and posterior collapse, which result in VAEs struggling to capture high-frequency details, leading to blurry reconstructions. Posterior collapse occurs when the encoder’s approximate posterior distribution becomes nearly identical to the prior, regardless of the input, and the decoder ignores the latent details.

To address these limitations, VQ-VAE [145] provides a discrete latent space parameterized by a learned codebook of embedding vectors. The encoder maps each input to a continuous latent vector, and the quantization process selects the nearest codebook entry, resulting in the quantized latent vector. This methodology replaces the probabilistic latent distribution with a categorical distribution over discrete indices, providing a symbolic, more structured representation of the input distribution. The decoder models the conditional likelihood of the image given the quantized latent vectors. The model is trained with loss terms for reconstruction, codebook, and commitment, enabling stable end-to-end learning of discrete latent structures that capture meaningful semantic patterns. VQ-VAEs have demonstrated their strong usability across several domains, including image compression [146], speech synthesis [147], neural audio codecs [148], and generative modelling [149].

The representation learned by VQ-VAE is highly advantageous for semantic communication. In accordance with the SemCom conceptual framework in 1.1.1, the quantized latents form a compact symbolic vocabulary that enables efficient transmission of semantics as indices rather than full continuous feature representations, which significantly reduces the bandwidth requirements. The discretization provides inherent robustness to channel noise, as small perturbations in the transmitted bits typically affect only a limited number of codebook indices. The shared codebook between transmitter and receiver, which is the shared knowledge-base, a representative of the data distribution, ensures that even lossy or corrupted latent representations can be decoded into coherent and semantically meaningful reconstructions. The compatibility of discrete latents with transformer-based or autoregressive priors allows

us to redesign powerful generative models to restore missing tokens or refine noisy representations, which enhances robustness in challenging channel conditions. These properties make VQ-VAEs a promising foundation for semantic communication systems where the goal is to transmit minimal symbols that convey meaning rather than raw signal fidelity. However, challenges such as the usage of VQ-VAEs in communication contexts, such as scalability of the learned representations to unseen distributions, task performance, and robustness to noise, are yet to be explored.

The initial SemCom models used VQ-VAE [90] with the objective of robustness against semantic noise. However, the model was evaluated only for classification and retrieval tasks and was tuned towards them; hence, it may not generalize well to other tasks or contexts. In [150], the authors propose a variant of the VQ-VAE that uses federated learning to personalize codebooks for different vehicles. The model improves performance in varied environments, but at a cost of computation due to increased coordination. In JSCC inspired design [151], introduced VQ-VAE as a unified backbone to achieve graceful degradation in noisy channels, with limited coding rate flexibility compared to current channel coding methodologies, which leads to retraining the model for varying channel conditions. The security aspect of SemCom is explored in [152], where the authors have added adaptive semantic noise to confuse eavesdroppers, with the intention of improving privacy, but this will cause semantic drift in the underlying receiver tasks. A recent study by [153] explores task-adaptive quantization by combining imitation learning with transformer-based semantic extraction, which offers a promising future direction. A limitation could be the task-driven training with limited generalizability. VQ-VAE based SemCom systems have recently emerged in large numbers across different aspects, indicating strong backing for SemCom.

2.5.4 Transformers

Transformers were first introduced by Google research [154] as a novel deep-learning architecture. Initially, transformers were studied in natural language processing, and subsequently their use in deep learning expanded to time series forecasting, computer vision, multimodal learning, and speech processing, and has recently been used in

generative architectures. The core of transformers is the self-attention mechanism, which enables each token to contextualize itself relative to all others in a sequence, thereby capturing long-range dependencies and outperforming recurrent neural network (RNN) and LSTM. In the vision domain, the ViT [155] demonstrates that images can be processed effectively by splitting them into fixed-size patches, treating them as tokens. This enables transformers to perform global semantic feature extraction [156] and model global spatial relationships. For modeling temporal dynamics in video sequences, transformers can be used to capture long-range dependencies across frames, as the authors of [157] did for the video classification task. Predicting mission information or tokens from learned images or videos, enabling generative modeling of video sequences with extracted spatial feature extractions, was explored in [149], where the authors generated high-resolution long-range video. In recent literature, transformers are investigated by Google for predicting future states in the context of world models [158, 159].

These capabilities make transformers particularly advantageous in semantic communication systems, where global context understanding, robust feature extraction, and the ability to operate on discrete latent tokens are essential for maintaining semantic integrity under channel distortions. As a result, transformer-based architectures are increasingly being integrated into encoder-decoder pipelines, predictive world models, and generative reconstruction frameworks within next-generation SemCom research.

2.6 Semantic communication and edge intelligence

By leveraging task-oriented encoding and knowledge-aware feature extraction, SemCom systems [160, 161] enable efficient communication in bandwidth-constrained environments typical of edge computing. Recent studies have emphasized the potential of semantic-based transmission to support intelligent applications such as the internet of vehicles (IoV) [162] and edge-assisted image and video systems [163], where contextual understanding is prioritized over bit-level accuracy [164]. Integrating SemCom with edge intelligence [165] allows distributed devices to exchange compact semantic representations instead of entire data streams. This significantly reduces communication

overhead while maintaining semantic fidelity [166], making it a promising foundation for scalable and intelligent consumer electronic ecosystems [167].

Resource optimization at the edge is studied in recent literature. The model in [168] extracts continuous semantic features for collaborative processing at edge servers, optimizing edge server density and semantic compression ratios to enhance energy efficiency. The study in [169], a lightweight ViT model is deployed at the edge. In contrast, a more complex ViT model is used on the server in [170], enhancing efficiency via attention-aware patch selection and entropy-aware image transmission, but with increased computational intensity when combined with ViT, VQ-VAE models, introduced in [15], effectively discretize data features. The authors in [90] employ VQ in a transformer-based SemCom system to mitigate semantic noise, thereby enhancing classification performance. However, this approach incurs higher resource consumption, making it unsuitable for edge computing systems. Furthermore, the authors in [79] achieve better transmission quality by sending discrete vectors, in contrast to more resource-efficient index-based systems.

2.7 Video semantic communication

Recent research in video generation and prediction focuses primarily on accurate frame synthesis and temporal coherence, often without considering communication constraints. The model developed in [171] is a latent prediction framework for procedural video representation learning, while the model in [172] employs LLM-guided scene planning for consistent multi-scene generation. Authors of [173] model videos as continuous, multi-dimensional processes for fine-grained frame prediction, and the model in [174] utilizes flow-guided deformable networks to capture local motion cues. Authors in [175] propose spatial-temporal motion generation using joint modeling, while [176] designs a hybrid Convolution LSTM-3D CNN for frame forecasting, and [177] combines VQGAN and temporal transformers for long video generation. While these methods concentrate on visual fidelity and temporal smoothness, they primarily address frame prediction in isolation, without integrating the predictive process into a semantic or communication

framework. Most aim to reconstruct future frames rather than explicitly modeling the underlying dynamics or trajectories of the observed world. Consequently, a gap remains in developing latent prediction architectures that couple world-model-based developments with efficient and communication-aware representations.

Recent advances in semantic communications for video have begun to shift from bit-level fidelity to goal-oriented and semantically relevant transmission [178], while aiming to minimize bandwidth while maintaining actionable or interpretable information. A foundation for learning-based semantic video transmission beyond conventional codec pipelines is considered. DeepWiVe model [179], which jointly learns compression and channel mapping to enable end-to-end deep wireless video delivery, demonstrates adaptability to channel variations. Deep video semantic transmission [180] further extended this paradigm to semantic spaces by proposing a nonlinear JSCC transform that embeds task-relevant features directly into channel symbols for improved robustness and a rate–distortion trade-off. The work in [181] considers major objects as semantic entities with contextual encoding to allocate bitrate to salient regions, and [182] designs a surveillance-specific semantic compressor that integrates object detection and language summarization.

In video transmission, channel noise is an important factor because even little distortions can have a big impact on the content’s visual quality and semantic interpretation. The study in [183] proposes a U-Net–based deep learning framework for robust video semantic transmission that compensates multi-dimensional noise using attention-guided correction layers. The model focuses on pixel-level recovery and lacks interpretability or task-level semantic optimization. Recent advancements in generative models, such as GANs [184] and diffusion models [185], have contributed significantly to the development of deep video SemCom. The reference [186] employs a diffusion-based goal-oriented framework, achieving superior visual fidelity under noise. The authors of [187] fuse multimodal semantics for cross-modal alignment. Edge-based video SemCom systems are a potential application in this area of research, and the implementation in [188] integrates computer-vision sensing for edge feature extraction. Beyond reconstruction, the model in [189] demonstrates adaptive bandwidth control for

video question answering, which is a goal-oriented approach.

2.8 Semantic communication applications

Semantic communication is being considered for real-world deployment across a variety of application domains, not only to compress data but also to support intelligent services under stringent latency, reliability, and resource constraints. The surveys increasingly consider practical deployments that jointly consider semantic extraction, task performance, and network efficiency across domains such as XR, edge computing, and healthcare. As the foundation of this high-level perspective, recent research on practically driven SemCom systems can be categorized into several main streams as shown in figure 2.6.

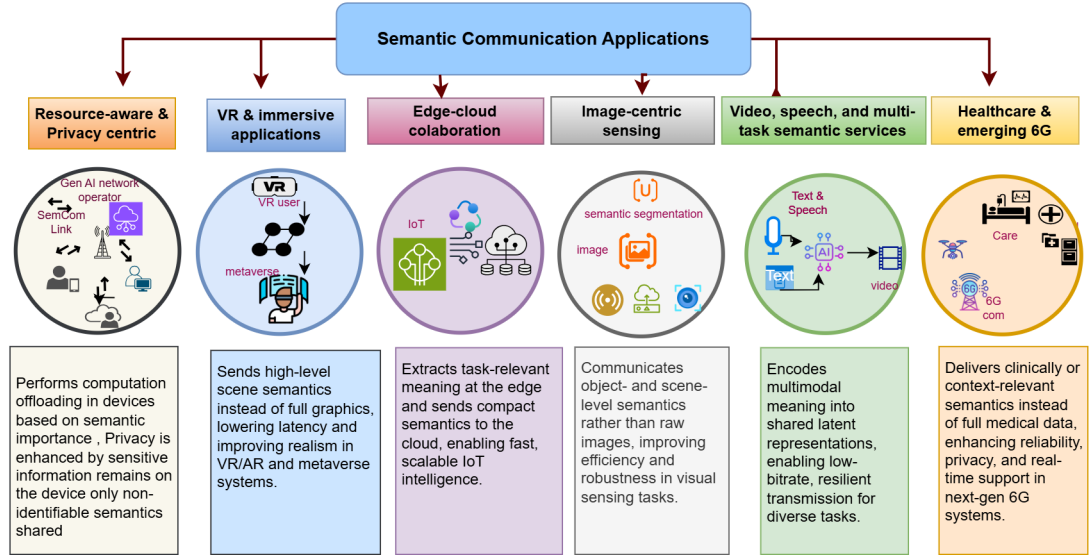


Figure 2.6: Semantic communication applications

Resource-aware and privacy-centric semantic networking In this application area, SemCom is used in 6G network design to use as a resource allocation tool. The survey in [190] demonstrates how SemCom can be embedded into future 6G architectures, emphasizing semantic information extraction, knowledge graph support, and resource-aware transmission policies. However, it remains largely conceptual and lacks detailed implementation guidelines. In [75], the authors design a computation

offloading framework using SemCom in edge networks and solve a delay-minimization problem via deep RL. In integrated sensing–computing–communication with fluid antennas [191], the system jointly optimizes beamforming, antenna positions, and semantic extraction ratios in near-field integration of sensing and communication (ISAC) systems. In [192], the authors designed a SemCom framework for privacy-oriented applications that uses SemCom and strong coordination for randomized distributed function computation under differential privacy constraints.

Virtual reality, metaverse, and immersive applications These applications demand tight coupling between semantics and temporal consistency. Accordingly, [193] proposes a age of incorrect information (AoII) as a joint metric for semantic error and timeliness in SemCom systems in XR aided by non-orthogonal multiple access (NOMA). The model provides a practically meaningful metric for XR, but focuses on text-based pipelines. Metaverse in 6G networks is another application that can benefit from SemCom. The authors in [194] propose caching domain-specific and user-specific models at the edge to improve responsiveness and reduce the cost of building individual knowledge bases, but the work remains conceptual and lacks quantitative analysis. The study in [195] presents two typical application scenarios in SemCom from the education and disaster tracking domains, which are relevant for India, and, importantly, standardization efforts are also studied.

Edge–cloud collaboration and intelligence for time-critical IoT Several works use SemCom to coordinate edge devices and servers for time-critical inference. The study in [196] proposes a collaboration framework to prioritize the transmission of the most relevant semantic data using “effectiveness encoding”. With augmentation, they have converted the CNN into a robust, channel-agnostic model. However, results are not generalizable across heterogeneous tasks. The proposed methodology in [197] combines semantic extraction with computation offloading in edge computing networks. However, they treat semantics primarily as an adjustable compression factor and do not fully utilize the task structure. In [194], the authors highlight semantic caching at the edge to accelerate knowledge-base construction. In contrast, article [198] outlines how SemCom

can support data reduction and reliability in healthcare systems with distributed sensors and medical devices. These efforts demonstrate the promise of SemCom enabled edge intelligence, yet most still lack application-layer control over semantic efficiency.

Image-centric sensing and reconstruction for ITS and IoT In perception-driven applications, image semantics are explicitly shaped by task relevance. The model proposed in [199] is an offline knowledge-base and attention-driven SemCom framework for ITS, in which segmentation masks of dynamic objects, such as vehicles, pedestrians, and traffic signals, are stored and fused via sparse, bi-level routing attention at runtime. This design achieves better reconstruction quality but may struggle with unseen object categories. Survey work presented in [8] further explores the applications in UAV remote sensing and intelligent transportation, reinforcing that semantic segmentation and task-driven feature selection are key to practical deployments. Notably, semantic segmentation based on SemCom is widely used in ITS based solutions, while they suffer from a lack of generalization due to a fixed number of label classes.

Speech, and multi-task semantic services Speech and multi-modal services constitute another important class of real-world applications. A universal speech SemCom framework presented in [200] is developed that combines wav2vec 2.0 for unsupervised speech representations, enabling both recognition and high-quality reconstruction over noisy channels. The system maintains robust performance even at high BER, but its reliance on large pre-trained models may challenge deployment on constrained IoT devices. The concept of SemCom can be used in agentic frameworks in non-wireless settings, as [201] uses to model financial markets as complex systems where heterogeneous agents assign meaning to information. Using wavelet-based analysis, they show how semantic information flows shape price dynamics. While not aligned with the traditional SemCom stack or the common objectives of SemCom, this work illustrates how semantic concepts can inform time-series modeling in socio-economic systems. The authors of [202] investigate extensions of canonical correlation analysis and latent semantic indexing for resolving semantic incompatibilities in data interpretation, while providing statistical tools for aligning heterogeneous semantic spaces. The drawback is

that this method remains largely at the algorithmic level.

Healthcare and emerging 6G domains Medical imaging is a field that can greatly benefit from SemCom with the emergence of telemedicine, remote healthcare monitoring/elderly Care, remote rehabilitation, physiotherapy, remote surgery, and telesurgery, where every spatial and temporal detail may not be needed to make decisions. Such an application, as surveyed by [198], can leverage SemCom for bandwidth-efficient transmission of medical images and sensor data. However, most proposals are still at the conceptual or early-evaluation stage. In broader terms, the articles [8, 190] synthesize application trends across UAVs, transportation, healthcare, metaverse, and 6G-native services, identifying open challenges in knowledge-base management, reliability, and cross-layer standardization.

In summary, existing real-world deployments demonstrate substantial gains in bandwidth, latency, and robustness. Nevertheless, they often rely on task- or domain-specific architectures, and many do not fully address long-term maintainability, interoperability, and security in large-scale operational processes.

2.9 Communication channel models and channel coding

AWGN channel The AWGN channel assumes that transmitted signals are corrupted by a noise component (n), which is an independent and identically distributed Gaussian random variable with zero mean and variance (σ^2). The received signal is expressed as, $y = x + n$, $n \sim \mathcal{N}(0, \sigma^2)$ because the noise is “white”, its power is evenly distributed across all frequencies, and it does not introduce any frequency-selective distortion. The AWGN channel represents an idealized environment and is often used as a performance upper bound for wireless systems. AWGN plays an essential role in research due to its analytical tractability and its relevance for systems dominated by thermal noise, such as short-range or highly controlled wireless links [24].

Rayleigh fading channel While AWGN captures noise, it does not account for one of the defining characteristics of real wireless environments, which is multipath

propagation. Rayleigh fading models the scenario where the transmitted signal reaches the receiver through numerous scattered, non-line-of-sight paths. The channel coefficient (h) is modeled as a complex Gaussian random variable, $h \sim \mathcal{CN}(0, 1)$. The received signal becomes $y = hx + n$. Since the magnitude ($|h|$) follows a Rayleigh distribution, this model captures rapid amplitude fluctuations, deep fades, and the stochastic nature of wireless propagation. Rayleigh fading is widely used to model urban, indoor, and dense multipath environments [20], representing real-world conditions where signal blockage and scattering occur frequently.

In summary, both AWGN and Rayleigh fading channels approximate distinct aspects of communication environments, where AWGN represents thermal noise, front-end noise, and scenarios with minimal multipath distortion, and Rayleigh fading approximates typical non-line-of-sight propagation found in mobile, urban, or indoor wireless networks [20]. Together, they offer complementary perspectives for evaluating the robustness of communication systems to noise, interference, and fading.

Channel coding The choice of channel coding in this thesis for chapters 4-7 prioritizes realism, deployability, and alignment with modern wireless standards, making LDPC codes the most suitable option for SemCom-image and video transmission. LDPC codes constitute the forward error correction (FEC) mechanism in fifth generation (5G) NR for the data channel with large payload such as images and video, while polar codes are widely used for the control channel in 5G wireless communication systems [203], with increased complexity for very long blocks but efficient for short ones [204, 205]. Therefore, LDPC [206] offers direct compatibility with practical communication pipelines and integrates seamlessly into the 5G-compliant Sionna simulation environment adopted in this work. In modern 5G systems, polar codes protect the control channel carrying short signaling messages, whereas LDPC codes protect the data channel responsible for large payload transmission. Semantic communication involves high-dimensional user data (images, video, feature vectors), which LDPC coding provides the standard-aligned and scalable choice for this work. This ensures that the experiments reflect operational channel conditions, including iterative decoding effects, rate adaptation, and realistic

error-floor behavior.

NVIDIA Sionna Sionna [207–209] is a graphics processing unit (GPU)-accelerated, open-source library for research in communication systems, which has recently gained attention for its more realistic outputs and capability in evaluating wireless models. To explore a more realistic simulation environment, compatible with practical implementation, we consider the NVIDIA Sionna simulator in this thesis for the simulations in chapters 4-7. The significant aspects of Sionna that motivated the use in this research are as follows:

- **Realistic physical layer outcomes:** Sionna enables us to explore advanced channel models (For example, 3GPP-compliant MIMO, spatial consistency, mobility, and delay spread), which are more closely related to real-world conditions, leading to more realistic performance evaluations of channel coding techniques.
- **FEC models:** Sionna offers 3GPP-compliant channel and FEC models [210], including LDPC encoders and decoders specified for 5G NR, the standardized 5G radio access technology defined by 3GPP. These allow system-level evaluation under AWGN and Rayleigh fading channels that closely resemble real deployment scenarios.
- **End-to-end learnable simulations:** The Sionna environment allows the integration of channel coding and neural network-based components within an end-to-end simulation pipeline because its components are fully differentiable. This is specifically important in our work when we start exploring the deployment of the proposed algorithm. The TensorFlow-based APIs can be directly integrated into the SemCom pipelines in the experiments.
- **Performance and scalability:** Sionna is a GPU-accelerated library built on TensorFlow, which enables fast, large-scale simulations. This allows us to train and evaluate models over larger datasets and more complex scenarios compared to CPU-based simulators.
- **Reproducibility and industry relevance:** Sionna is an open-source, industry-based framework; therefore, our results are easier to reproduce and align more

closely with practical deployment characteristics, compared to simplified, custom-built simulators that may not generalize as well.

2.10 Summary

This chapter establishes the theoretical and technical foundations for subsequent developments in this thesis by offering a structured review of semantic communication research, its evolution, and its open challenges. The literature review first explores the information theoretical concepts. The review explores the classical Shannon-based concepts and their shift to semantic-based concepts. It further explores the main practical concepts underlying the theories, including semantic extraction, semantic representation, and the development of semantic languages and knowledge bases. The chapter reviews key strands of the SemCom design standards that currently prevail in the literature, including knowledge-enhanced SemCom, goal-oriented and task-oriented frameworks, JSCC-based systems, and generative-model-driven architectures, highlighting their operational principles, strengths, and limitations. An important aspect in developing this research was the appropriate selection of deep learning tools, such as discrete neural representations, contrastive learning, knowledge graph embeddings, and reasoning-aware world models, which enable the SemCom designs to evolve from raw latent feature vectors to structured, interpretable, and predictive semantic representations. Furthermore, the chapter surveys major application domains, including ITS, XR, distributed sensing, and edge intelligence, and identifies their constraints and emerging requirements. Collectively, this review identified critical gaps in SemCom system development with standard design principles based on the evolving SemCom theories, such as the instability of shared representations, efficient knowledge base designs, limited robustness to semantic noise, a lack of standardized semantic metrics, and the absence of predictive temporal modeling in current video based SemCom systems, thereby motivating the design choices and problem formulations addressed in the subsequent chapters of the thesis.

Chapter 3

Segmentation guided semantic symbol encoding for generative semantic communication

This chapter presents the first experimental design and implementation of a Generative AI-based SemCom system for wireless image transmission. It provides the foundational study upon which the later generative and vector-quantized frameworks developed in this thesis are built. The work introduces an end-to-end communication system that integrates a semantic encoder based on image segmentation and a GAN-based decoder, operating over a practical simulated wireless channel. The study explores how semantic segmentation maps can be transmitted as compact representations of image semantics, which are then used to generate the image at the receiver, leveraging the shared knowledge base, and evaluated under realistic channel conditions.

The primary contribution of this study is to identify the fundamental design considerations and challenges associated with implementing practical generative AI-based SemCom systems. Specifically, the chapter analyses the relationship between semantic feature preservation under physical channel impairments, the role of a shared common knowledge base in maintaining consistency between the transmitter and receiver, and the trade-offs between bandwidth efficiency and reconstruction fidelity. The experiments assess the feasibility of using semantic maps as semantic symbols and derive quantitative

Chapter 3. Segmentation guided semantic symbol encoding for generative semantic communication

insights into semantic robustness, edge preservation, and the impact of quantization and channel noise on generative reconstruction.

By establishing these empirical foundations, this study provides a practical framework for understanding design considerations for generative AI-based SemCom systems. It serves as the first step towards the development of more advanced generalizable and scalable generative AI-based SemCom architectures introduced in subsequent chapters. The findings of this chapter highlight both the promise and the open challenges of SemCom bridging the theoretical constructs of semantic-based transmission with the practical realities of physical wireless channels.

In summary, this chapter presents the following novel contributions.

- A SemCom based image end-to-end transmission system for limited bandwidth wireless communication systems that leverage a pre-trained GAN network as the semantic decoder. The semantic decoder generates the intended image at the receiver using the semantics (segmented semantic map) of the ground truth image transmitted through a noisy channel.
- We study the effect of physical channel noise on the design of semantic communication systems using existing physical communication channels and derive important insights for scalable, generalizable SemCom systems. The study also evaluates the effect of channel noise for three scenarios.
 - (a) Semantic communication of the images under varying channel noise.
 - (b) Quantization effect on semantic communication with zero channel noise.
 - (c) The joint channel noise and quantization effect on semantic image transmission.

The structure of this chapter is as follows. Section 3.1 presents the problem formulation, including the concept and the theoretical justifications that explain how the key components jointly contribute to semantic compression and preservation under channel noise. Section 3.2 presents the system model, including the main components and the training strategy. Section 3.3 presents the simulation results, and finally, section 3.5 concludes this chapter.

Chapter 3. Segmentation guided semantic symbol encoding for generative semantic communication

This chapter is based on the following open-access article published in the IEEE Access journal under a Creative Commons Attribution (CC-BY 4.0) license: M. U. Lokumarambage, V. S. S. Gowrisetty, H. Rezaei, T. Sivalingham, N. Rajatheva and A. Fernando, "Wireless End-to-End Image Transmission System Using Semantic Communications." vol. 11, pp. 37149-37163, 2023, doi: 10.1109/ACCESS.2023.3266656.

3.1 Problem formulation

In modern intelligent multimedia communication systems like surveillance and autonomous systems, the pixel-level accuracy is less important because the objective is no longer to reconstruct the transmitted image with the exact pixel-level accuracy for human visual consumption but to extract task-relevant semantic entities, including objects, actions, or scene context, to get a meaningful outcome. Let the ground truth image be represented as $\mathbf{x} \in \mathbb{R}^{h \times w \times c}$, where h , w , and c denote height, width, and number of channels, respectively. The semantic encoder $f_\theta(\cdot)$ maps the source image into a semantic representation S_{sem} with object-level semantic features:

$$S_{sem} = f_\theta(\mathbf{x}), \quad S_{sem} \in \mathcal{S}, \quad (3.1)$$

Where S_{sem} is a finite set of semantic symbols that is extracted from the image as semantics, including segmentation labels. The mapping f_θ extracts object semantics that represent "what" the image depicts, rather than "how" it appears. The objective of our SemCom system is to preserve the object level semantics in the received image $\hat{\mathbf{x}}$, using the shared knowledge base $\mathcal{K}$, despite the presence of channel noise and can be expressed as

$$\min_{f_\theta, g_\phi} \mathbb{E}_{X \sim P_X} [d_{sem}(f_\theta(X), g_\phi(h(f_\theta(X))))], \quad (3.2)$$

where $g_\phi(\cdot)$ denotes the semantic decoder, $h(\cdot)$ models the physical channel. We define $d_{sem}(\cdot, \cdot)$ as a semantic distortion function that quantifies the loss of meaning rather than pixel-level differences, in our SemCom system, for the generation of objects.

The physical channel introduces additive noise $N \sim \mathcal{N}(0, \sigma^2)$, such that the

transmitted semantic features S_{sem} is received as

$$Y_{sem} = S_{sem} + N. \quad (3.3)$$

We introduce channel encoding and decoding functions $e_c(\cdot)$ and $d_c(\cdot)$ that are applied to improve the robustness of the transmitted semantic symbols:

$$\hat{S}_{sem} = g_\phi(d_c(Y_{sem})) = g_\phi(d_c(e_c(S_{sem}) + N)). \quad (3.4)$$

The optimization objective for our SemCom system is a joint minimization of semantic distortion (segmentation accuracy) and channel-induced degradation (minimization of physical channel distortions), and can be illustrated as

$$\min_{f_\theta, g_\phi, e_c, d_c} \left(\mathbb{E}[d_{sem}(S_{sem}, \hat{S}_{sem})] + \lambda \mathbb{E}[\|N\|^2] \right), \quad (3.5)$$

Where λ is a trade-off coefficient balancing semantic fidelity and channel efficiency.

In the context of this study, semantic segmentation maps serve as the transmitted semantic representation S_{sem} . The receiver employs a GAN to reconstruct $\hat{\mathbf{x}}$ from the received semantic symbols, $\hat{S}_{sem}$ using the shared knowledge base $\mathcal{K}$. In our study, we used the COCO dataset as the shared knowledge base. The transmission objective is to minimize the loss between the ground-truth semantics and the reconstructed image, subject to bandwidth limitations and the noise of the physical channel. The problem formulation serves as the basis for the succeeding architecture, where object-level semantics form the core of the transmitted information rather than raw pixel-level data, enabling the receiver to generate semantic content even under noisy conditions.

3.2 System model

We have designed the SemCom system based on the theoretical framework derived in figure 3.1. This three layered model is based on Shanon's theory and is initially optimized to illustrate the Generative AI based SemCom. The semantic communication framework comprises two principal components: semantic feature encoding and decoding. The

Chapter 3. Segmentation guided semantic symbol encoding for generative semantic communication

semantic encoder extracts high-level semantic features from the source message, which can be in the form of text, images, video, or sound. This extraction process is based on the wisdom shared between the transmitter and receiver to interpret the underlying meaning of the message, the shared knowledge base. The semantic decoder reconstructs the intended semantics from the received symbolic representations, utilizing the same shared knowledge base to ensure accurate interpretation. Beneath the semantic layer lies the physical layer, which handles the bit-level transmission of the encoded semantic information through the physical channel, governed by Shannon’s theories. This layer focuses on optimizing data transmission by applying channel-level techniques, specifically channel coding, to reduce bit errors introduced during channel transmission. Above the semantic layer is the application layer, which addresses task-specific operations on the received message, such as classification, object detection, or scene understanding. The SemCom system described in the following sections is grounded in this layered theoretical model, integrating semantic understanding with efficient transmission and task-oriented processing.

Unlike in traditional communication systems, noise can affect information at multiple levels of the communication process in SemCom. Channel noise refers to random disturbances introduced by the physical communication medium during transmission, which distort the transmitted signal and lead to errors in the received data. Traditional communication systems primarily consider channel noise at the physical layer. Semantic communication introduces the concept of semantic noise, which occurs at the semantic layer during the process of meaning extraction and interpretation. In the architecture proposed in this chapter, semantic noise may arise from inaccuracies in the segmentation process, including incorrect object labels, limited labels interpreted, missing regions, or boundary distortions between semantic classes. These errors may alter the semantic structure of the transmitted representation even when the underlying bit transmission is reliable. Consequently, distortions introduced in the communication pipeline can propagate to the receiver in the form of boundary corruptions in the segmentation maps and affect the generation process at the receiver. Semantic noise is therefore identified as arising in two forms: ambiguities introduced in the semantic extraction process and

pixel-level boundary corruptions caused by channel noise, which is propagated to the decoding process, causing failures in the generation process. A more detailed theoretical and experimental analysis of semantic noise and its impact on semantic communication systems is quantitatively assessed and modeled in the chapter 6 of this thesis.

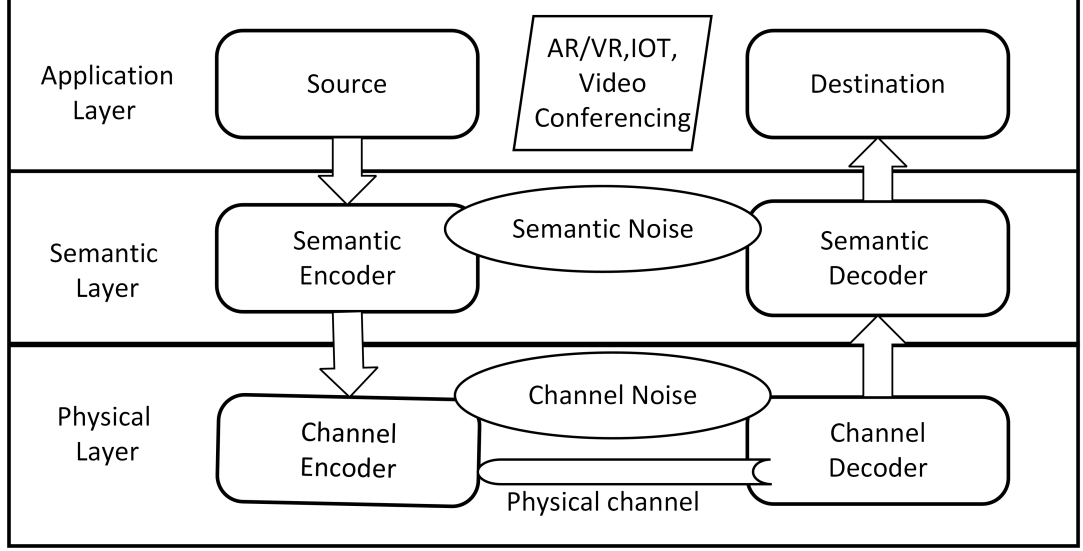


Figure 3.1: Conceptual three-layer model of semantic communication, comprising the technical layer for signal transmission, the semantic layer for meaning extraction, and the effectiveness layer for task-driven interpretation.(Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)

Figure 3.2 presents the architectural overview of the system, with its constituent modules according to the layered theoretical framework described above. The semantic layer comprises the semantic feature extraction module (semantic segmentation), the shared knowledge base, and the semantic decoder, a GAN, all of which contribute to the interpretation and reconstruction of semantic content. The physical layer comprises the channel encoder, transmission channel, channel decoder, and an error concealment module. The channel encoder adds controlled redundancy to the transmitted data to provide forward error correction, which the channel decoder subsequently leverages to identify and correct bit-level errors introduced during transmission. We used polar coding as the channel coding methodology in this design. The error concealment module is designed to mitigate the impact of channel noise, thereby enhancing the fidelity of the

received semantic representation. The following sections describe the model in detail.

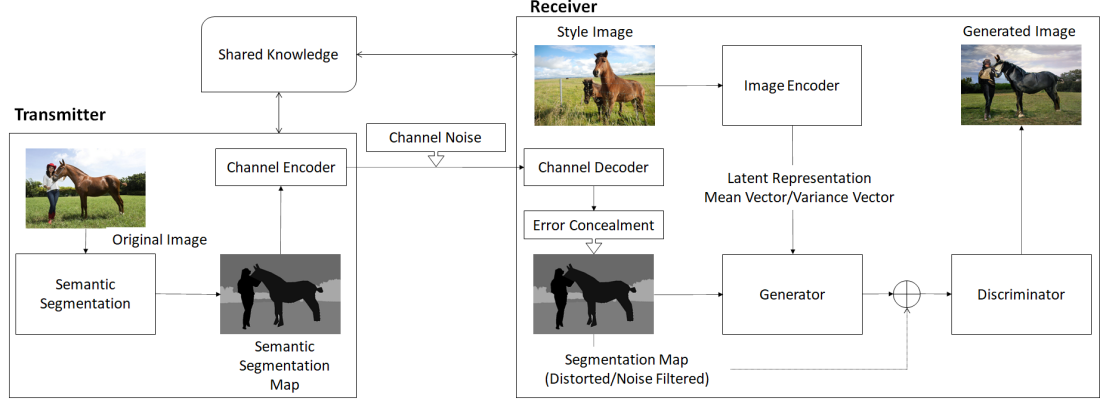


Figure 3.2: Semantic communication system for image transmission, illustrating semantic extraction, channel transmission, and semantic-driven generation. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)

3.2.1 Semantic encoder (transmitter)

The semantic encoder extracts semantics and represents the high-level semantics of the source image, capturing its meaning, in terms of object classes and their position. The transmitter (the semantic encoder) performs semantic segmentation to generate a structured representation of the image for transmission over a bandwidth-limited wireless channel. The encoder applies a semantic segmentation function $\psi_{\theta}(\cdot)$, trained on the COCO–Stuff dataset [211], to assign a semantic label $l_{i,j}$ from the predefined label set $\mathcal{L} = \{l_1, l_2, \dots, l_{|\mathcal{L}|}\}$ to each pixel (i, j) in the image as

$$S_{sem}(i, j) = \psi_{\theta}(X(i, j)) \in \mathcal{L}. \quad (3.6)$$

The output S_{Sem} is a semantic segmentation map as shown in figure 3.3 in which each pixel encodes the class identity of the object region it belongs to. The symbolic representation reduces the data dimensionality from continuous pixel intensities to discrete semantic indices based on class labels, thereby saving bandwidth while retaining the essential object-level information needed for reconstruction and task inference at the receiver. The segmentation network employed in this study is trained on the COCO–Stuff dataset [211], which provides 182 semantic object and background classes

Chapter 3. Segmentation guided semantic symbol encoding for generative semantic communication

containing a broad range of natural scenes. We selected the COCO stuff dataset to extend the concept of semantic symbols into a generalized SemCom system, as discussed in the next chapter. The network thus learns to assign consistent and interpretable labels for example “person”, “car”, “tree”, “road”, “sky”, or “background”, forming a compact symbolic minimalist vocabulary shared by both the transmitter and the receiver according to the theories discussed in section 2.1. The use of this common semantic vocabulary ensures that the meaning extracted by the transmitter is interpretable in the same context at the receiver, establishing a shared knowledge base.

3.2.2 Common knowledge base

The COCO dataset and its knowledge of 182 object classes serve as the common knowledge base for the developed semantic communication-based image transmission system. The semantic segmentation (transmitter) and the GAN (receiver) are both trained on this single dataset. The Common Objects in Context (COCO) dataset is designed for object detection, segmentation, and captioning. The dataset consists of 118,000 training images and 5000 validation images of day-to-day objects captured in everyday scenes, which aligns with the thesis objective of designing generalized SemCom systems that are not limited to a specific domain.

Chapter 3. Segmentation guided semantic symbol encoding for generative semantic communication



Figure 3.3: Visualization of segmentation masks and the associated transmitted images, showing how semantic regions are encoded and preserved in semantic communication. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)

3.2.3 Semantic decoder (receiver)

We used a GAN pre-trained on the COCO dataset as the semantic decoder in this system, specifically SPADE [212]. The training process of this GAN is illustrated in figure 3.4. The generator G produces images $G(x)$ from a random input x , and the generated images are fed to the discriminator along with the ground truth images. The generated image is then classified as “real” or “fake” by the discriminator, and the loss function is updated.

The GAN aligns the generated data distribution with the desired probability distribution. The $V(D, G)$ denotes the adversarial objective in vanilla GAN that defines the two-player minimax game between the discriminator and the generator.

$$\min_G \max_D V(D, G) = E_{s \sim p_{data}(s)}[\log D(s)] + E_{x \sim p_x(x)}[\log(1 - D(G(x)))], \quad (3.7)$$

where $p_{data}(s)$ is the probability distribution of real images, and $p_x(x)$ is the probability distribution of generated images. $E_{s \sim p_{data}(s)}[\log D(s)]$ is the average log probability of the discriminator with a real image, and $E_{x \sim p_x(x)}[\log(1 - D(G(x)))]$ is the average log probability of the discriminator with a generated image. The objective of the discriminator is to accurately categorize real images as “true” in order to bring $D(s)$ close to 1 by maximizing the term $E_{s \sim p_{data}(s)}[\log D(s)]$, and to maximize $E_{x \sim p_x(x)}[\log(1 - D(G(x)))]$ by accurately categorizing generated images as “false” and bringing $D(G(x))$ close to 0. The generator’s objective is to trick the discriminator by producing images that closely resemble real images. The generator aims to minimize the term $E_{x \sim p_x(x)}[\log(1 - D(G(x)))]$ by making $D(G(x))$ close to 1. To generate the image based on the semantics S_{Sem} transferred from the transmitter in the form of a semantic segmentation map, we have used a conditional GAN, as

$$\min_G \max_D V(D, G) = E_{s \sim p_{data}(s)}[\log D(s|c)] + E_{x \sim p_x(x)}[\log(1 - D(G(xc)))], \quad (3.8)$$

where c is the condition .

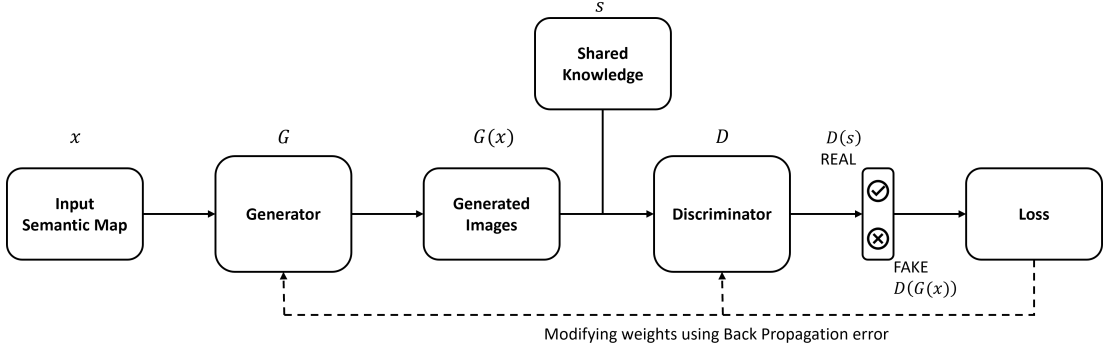


Figure 3.4: Adversarial training loop of a GAN, illustrating generator–discriminator updates. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)

3.2.4 Communication framework

We used polar coding as the channel coding scheme in this architecture due to its extremely reliable FEC. Polar codes, implemented in article [204], were adopted as the channel coding scheme in this architecture because they provide extremely reliable FEC. The block length, denoted by N_c , refers to the total number of bits in each encoded codeword produced by the channel encoder. In a linear block coding scheme, K_c number of message bits are mapped into N_c coded bits, resulting in a code rate $\mathcal{R}_c = K_c/N_c$. In polar codes, the code length is restricted to powers of 2. Channel polarization [213] is a technique that takes $N_c = 2^n$ identical copies of a binary-input discrete memoryless channel (B-DMC) W as $I(W)$ and transforms them into N_c virtual channels:

$$W_{N_c}^{(1)}, W_{N_c}^{(2)}, \dots, W_{N_c}^{(N_c)}.$$

As N_c increases (i.e., as the code length grows) each individual channel $W_i^{N_c}$ ($1 \leq i \leq N_c$) become extremely reliable (almost noiseless), meaning the reliability approaches to one (highly reliable ($I(W_i^{N_c}) \rightarrow 1$)), while the remaining channels become extremely unreliable (almost useless for data transmission), meaning the reliability becomes zero (highly unreliable ($I(W_i^{N_c}) \rightarrow 0$)). The reliable channels can be computed using Bhattacharya parameters [213], where the information bits are located. Figure 3.5 depicts the error-correction performance of polar codes for blocklengths in the range of

512 to 8192 with $\mathcal{R} = 1/2$. The performance of polar codes improves as the code length grows.

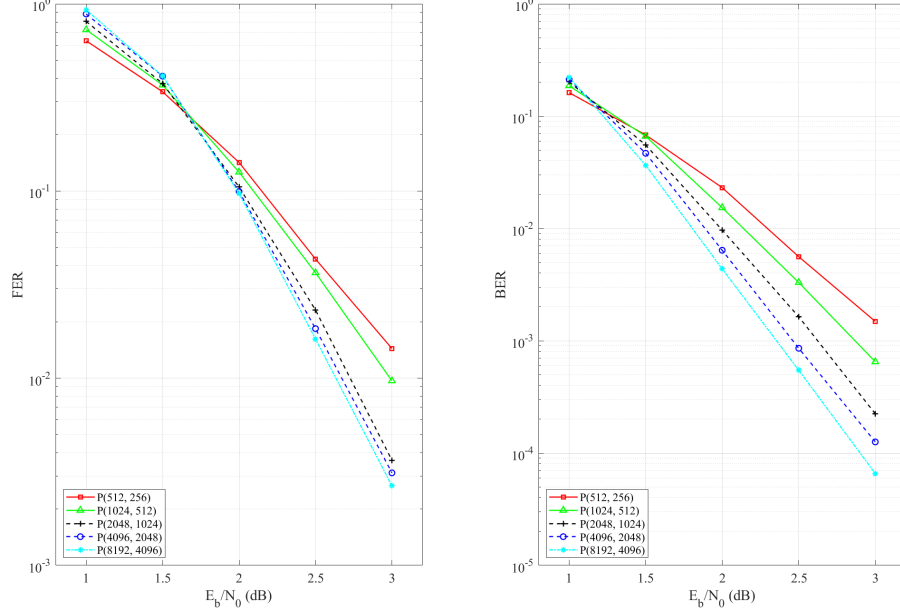


Figure 3.5: Error-correction performance of polar codes under different block lengths.(Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)

To verify the proposed semantic communication system using practical codes, a polar code of size $N_c = 4096$ carrying $K_c = 2048$ information bits is selected. Similar to [214], the chosen polar code is optimized for $E_b/N_o = 2.5$ dB. The codewords are modulated using binary phase-shift keying (BPSK) over an AWGN channel. At the receiver end, the noisy log-likelihood ratio (LLR)s are fed to the decoder, which computes the estimated codewords. The code rate was 0.5, and the demapping method is log-likelihood ratios.

3.2.5 Error concealment block

The transmitted semantic segmentation maps are susceptible to channel noise, which may alter pixel-level label values and degrade semantic consistency at the receiver. To mitigate these effects, an error-concealment block based on median filtering is applied after channel decoding. The median filter operates as a non-linear spatial

filter, replacing each pixel value in the received semantic map with the median of its local neighborhood. This process mitigates isolated bit errors of the type known as salt-and-pepper noise [215], caused by pixel values in the transmitted segmentation map due to channel noise, while preserving the boundaries and structural integrity of the segmented regions, preventing label errors during decoding. The median filtering module was implemented using the “medfilt2” function from the MATLAB Image Processing Toolbox [216]. It performed two-dimensional median filtering optimized for matrix-based image data. This mechanism ensures that the input semantics provided to the semantic decoder, which is the GAN, remain stable and representative of the original transmitted semantics.

3.3 Simulation results

This section presents the simulation results of the proposed communication framework for wireless image transmission. The primary objectives of these experiments are to assess how the design considerations should be aligned in designing future SemCom systems in practice,

- We evaluate the feasibility of transmitting images using only their semantic representations as minimal symbol representations derived from segmentation maps;
- We quantify the effects of quantization and channel noise on the preservation of semantic fidelity
- We compare the system’s performance with conventional image transmission methods.
- We evaluate the human perceived fidelity of the transmitted images, which provides a robust evaluation of the proposed direction of designing SemCom systems using minimally defined symbols and their potential usage in M2M and machine-to-human (M2H)

To achieve these objectives, controlled experiments were conducted using semantic segmentation maps extracted from the COCO–Stuff test set, transmitted through an AWGN channel with polar code encoding. The objective of this model is to design

Chapter 3. Segmentation guided semantic symbol encoding for generative semantic communication

a preliminary Gen-AI-based SemCom system to identify design considerations for future implementations; hence, we have only used an AWGN channel. Later chapters investigate the effect of more complex channels, specifically fading channels. We selected ten sample images representing different categories, including human figures, natural scenery, animals, vehicles, and household items. The intended task of the application layer at the receiver is object classification, in which the received images are categorized according to the set of object classes on which the GAN was initially trained. Performance was assessed both objectively and subjectively through measures of PSNR, compression ratio, and human-perceived semantic similarity. The results presented in this section demonstrate the proposed system's ability to convey the essential semantics of visual content over noisy channels with significantly reduced bandwidth, providing a foundation for the generative, reasoning-aware semantic communication architectures explored in subsequent chapters.

3.3.1 Compression efficiency

The ultimate objective of semantic communication is to reduce the strain on wireless communication channels caused by multimedia transmission. To compare traditional image compression with the proposed communication system, we have considered JPEG image compression and semantic images at different stages of the communication process. The typical compression ratio of JPEG [217] is around 1 : 10 with a slight quality degradation. Achieving higher compression results in a loss of image quality. Figure 3.6 illustrates the image size comparison at different stages of the communication path for the proposed transmission system with a sample image.

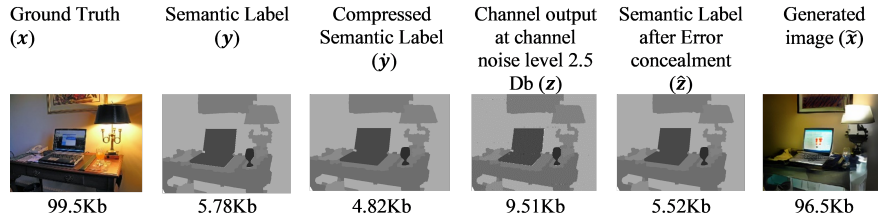


Figure 3.6: Visual illustration of the transmitted image size comparison for a sample image. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)

With JPEG, the transmitted image was 99.5 Kb in size, and the semantic segmentation map (semantic symbols) was only 5.7 Kb, resulting in a compression ratio of 17.2 (x/y). To further reduce the data rate, we compress the semantic segmentation map using traditional compression, gaining further compression of 20.6 ($x/\hat{y}$). The segmentation map size at the channel decoder input is 9.51 Kb at a noise level of 2.5 dB. The segmentation becomes 5.52 Kb after error concealment at this noise level, while the transmitted image is 96.5 Kb.

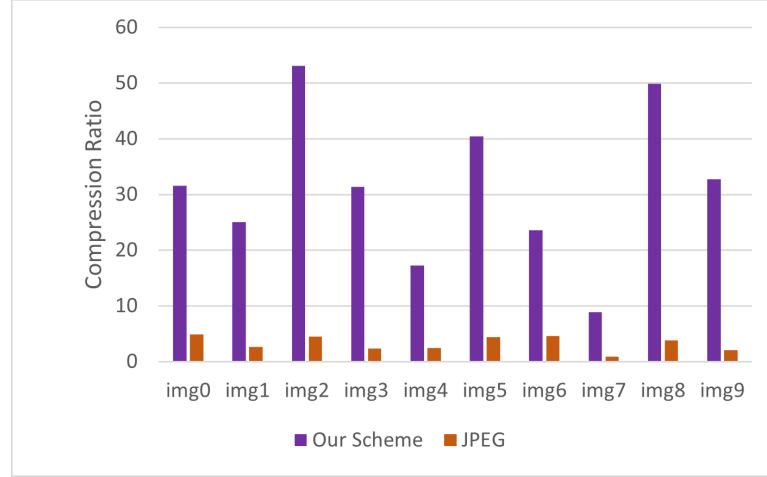


Figure 3.7: Comparison of compression ratios for sample images with the proposed model and traditional compression scheme. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)

The ground truth image was compressed using JPEG with zero channel noise, and the comparison with the proposed model is illustrated in figure 3.7. Noise is added to the compressed JPEG bit stream, and the bit stream is attempted to be decoded at the receiver. However, the JPEG decoder fails to decode any bit stream with added noise. Figure 3.8 illustrates the comparison of JPEG compression with noise. The proposed scheme outperforms JPEG-compressed image transmission in low-bit-rate and low-SNR scenarios. Even in high-SNR scenarios, the proposed compression outperforms JPEG compression in terms of bit rate, though JPEG-compressed images provide better quality for human perception. Since the intended main application of the proposed scheme is M2M communications, it outperforms JPEG-compressed transmission at all SNR values, as it consumes only 5% of the bandwidth required by JPEG.

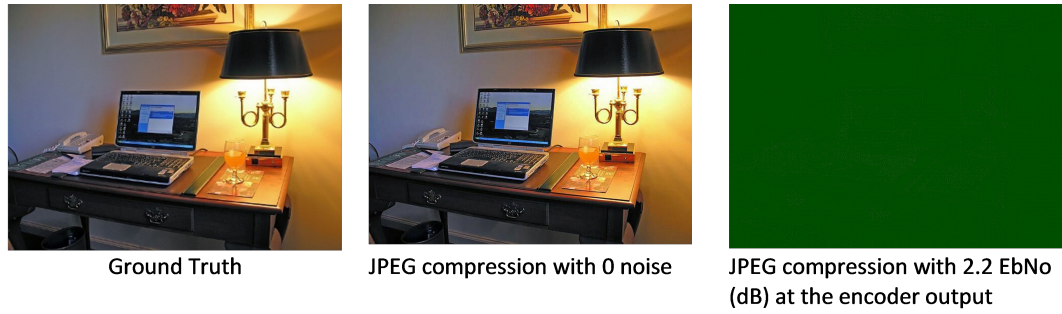


Figure 3.8: Illustration of JPEG image compression with added noise. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)

3.3.2 Impact of channel noise on semantic symbol accuracy and edge preservation

In semantic segmentation, each object in an image is represented by a distinct pixel value, and the differences in these values define the boundaries between objects. When the semantic maps are transmitted through a noisy physical channel, distortions occurring along object edges can disrupt the reconstruction of the original image at the receiver, as shown in figure 3.9.

This behavior was examined by gradually increasing the channel noise from 2.0 dB to 3.0 dB. At higher noise levels, a considerable number of semantic segmentation maps failed to produce valid outputs from the GAN, primarily due to severe edge distortion. The generator has learned to interpret object boundaries based on a finite set of predefined label classes. Consequently, when the edges become distorted by noise, creating “hybrid” or unexpected edge patterns, the generator encounters classifications that fall outside its learned manifold. This causes it to misinterpret the corrupted segmentation map, which appears to contain more label classes than those available in the training dataset, leading to incorrect or failed generations. Experiments in the following sections validate that maintaining edge integrity is essential for reliable semantic communication of image data. Figure 3.10 shows the received semantic map at a noise level of 2.3 dB before and after applying the error concealment, highlighting the improvement in edge preservation after processing.

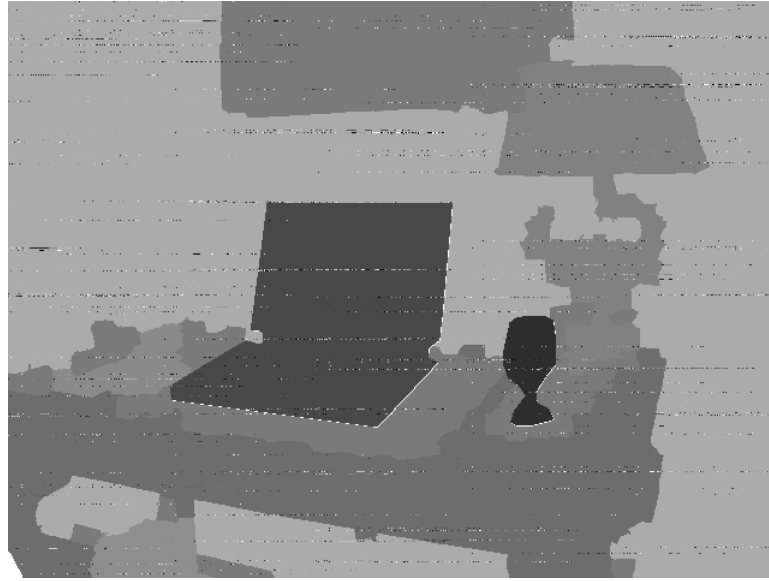


Figure 3.9: Effect of channel noise on the semantic segmentation map, illustrating how increased noise levels degrade semantic consistency and object boundary clarity. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)

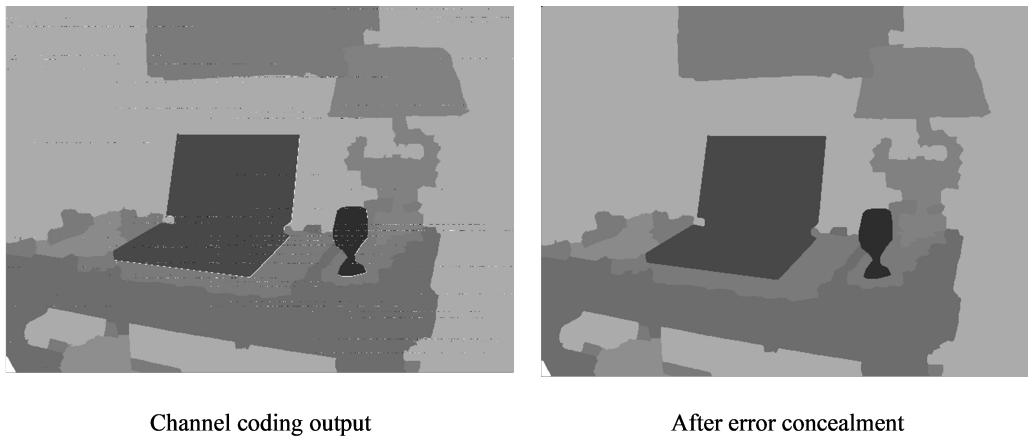


Figure 3.10: Effect of median filtering on segmentation maps corrupted with channel noise, demonstrating effective noise removal while preserving sharp edges and structural details. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)

3.3.3 Semantic symbol transmission under varying channel noise levels.

The experimentation on the proposed system is based on the AWGN channel. At low SNR levels, noise adds distortion to the semantic segmentation images; hence, the label data is erroneous, with distorted pixel values for the object classes. To mitigate this effect, we used median filtering as discussed in section 3.2.5. After error concealment, the PSNR value for a particular segmentation map is the same across all channel noise levels from 2.1 dB to 3.0 dB, as illustrated in figure 3.11. When the channel noise level increases, some semantic segmentation map images fail to generate an image at the GAN output. We have identified segmentation maps that the GAN fails to decode at some random SNR levels, and we observe that this is random. We concluded that the GAN required to be trained on a more diversified dataset with more class labels. The SemCom system can withstand up to 2.1 dB of noise to generate realistic images at the GAN level and, hence, is capable of performing semantic transmission. Figure 3.12 illustrates the PSNR variation for the generated images.

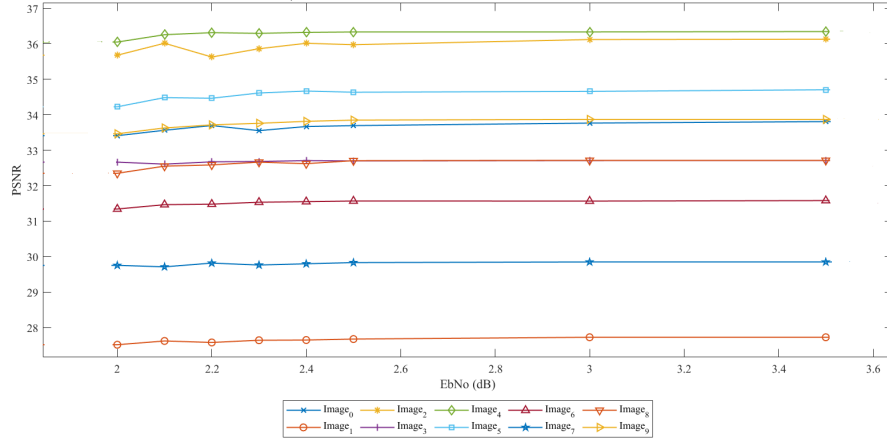


Figure 3.11: PSNR versus SNR for a sample of 10 semantic segmentation map images after noise concealment. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)

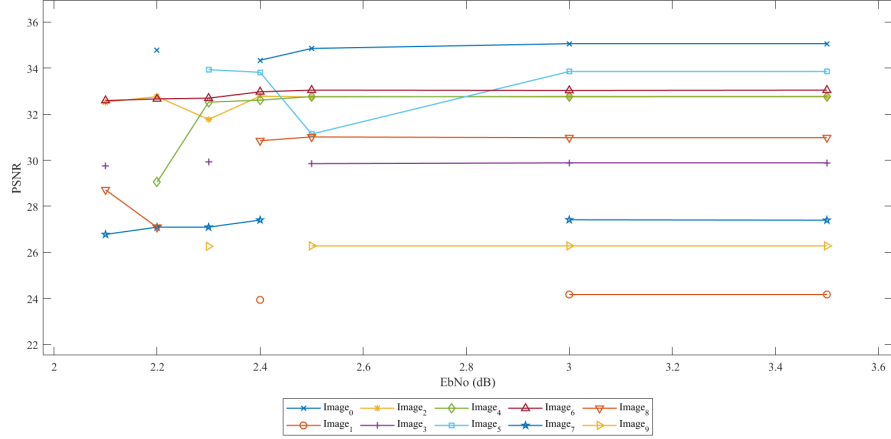


Figure 3.12: PSNR versus SNR transmitted image sample of 10. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)

As an ablation study, we have removed the error concealment block at the receiver and tested the semantic transmission. However, the semantic decoder fails at all the SNR levels. The intuition is that the developed system is susceptible to channel noise, and even if a single bit error may alter the pixel value of a particular label value of the transmitted semantics (S_{sem}), and hence alter the label definition that the semantic decoder can understand. Figure 3.13 illustrates the PSNR values for the selected 10 semantic segmentation map images without error concealment. Thus, the system is highly dependent on the dataset it is trained on, which has led us to design more generalizable SemCom systems, as in chapter 4.

3.3.4 Quantization effect with zero channel noise

Understanding the effect of quantization noise is essential because it provides insight into how compression-induced distortions alter the meaning represented by the semantic maps, thereby impacting the overall reliability of the semantic communication system. The compressed semantics were transmitted through the physical channel under severe noise. The PSNR for all the compressed semantic maps converged to infinity. However, the GAN fails to generate images because the number of label classes exceeds 182, which is the number of label classes the GAN was pre-trained on. Therefore, it was identified

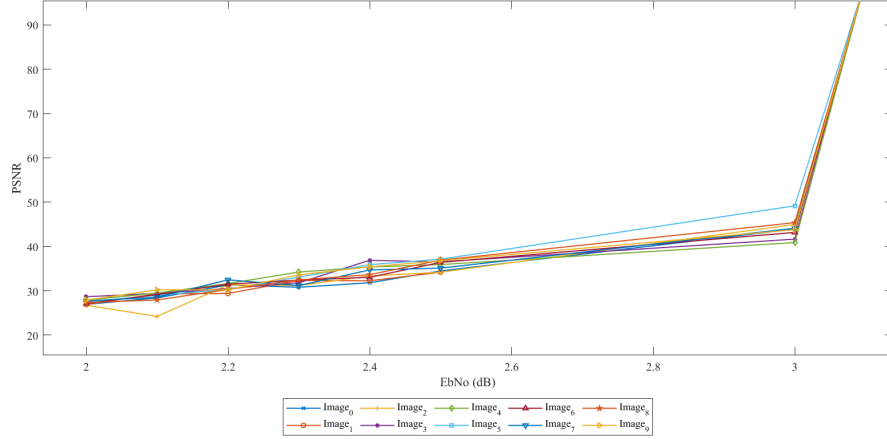


Figure 3.13: PSNR versus SNR without noise concealment for transmitted semantic segmentation map image sample of 10. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)

that, with lossy compression, the semantics transmitted in the form of segmentation maps cannot be handled by the semantic decoder. This observation has led us to develop a more generalizable SemCom system as defined in the later chapters of this thesis.

3.3.5 Joint quantization and channel noise effects

In this experiment, the combined impact of quantization and channel noise on the proposed semantic communication system was investigated. The semantic segmentation maps were first losslessly compressed at the transmitter before transmission. As shown in figure 3.14, the PSNR values computed for ten test images remained constant across all noise levels for a given semantic label image.

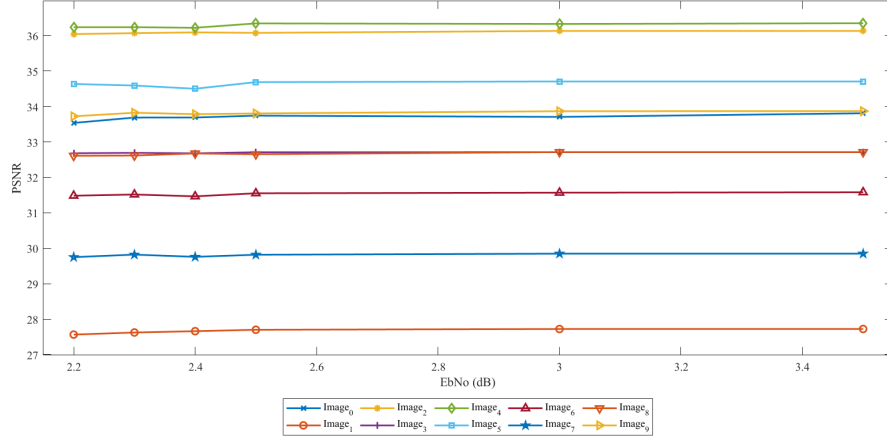


Figure 3.14: Impact of combined quantization effects and channel noise on transmitted compressed semantic segmentation map images, with noise concealment applied to mitigate distortions and preserve semantic regions. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)

Although the PSNR values of the semantic maps were unchanged after applying the median filter for error concealment, the decoder’s performance deteriorated as the channel noise increased. The system maintained acceptable generative performance up to a noise level of approximately (2.2 dB), beyond which the GAN failed to reconstruct realistic images. In several cases within this range, the GAN was unable to produce valid outputs because the received segmentation maps contained a higher number of label classes than those defined in the training dataset. This phenomenon is attributed to severe edge distortions in the semantics being transmitted that could not be effectively mitigated by the median filter. The PSNR variations of the GAN-generated images are illustrated in figure 3.15, demonstrating that edge distortion is the primary factor leading to reconstruction failure at higher noise conditions, which led us to develop the Sobel-based attention module for the semantic encoder discussed in chapter 4, in section 5.2.1

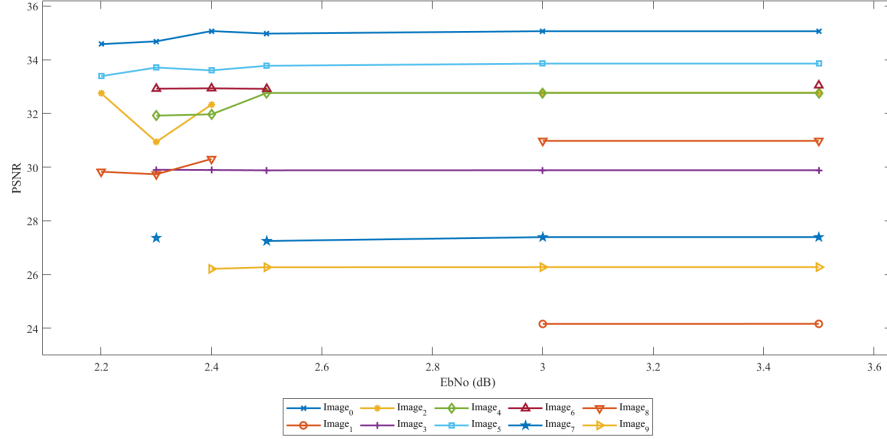


Figure 3.15: Image quality at different noise levels for generated images with joint quantization and channel noise. (Reprinted with permission from [1] under CC BY 4.0 license ©2023 IEEE)

3.3.6 Human perceived semantic fidelity: A qualitative evaluation

To complement the distortion metric, a subjective experiment was conducted to assess semantic fidelity from a human-perception perspective. Since semantic communication aims to preserve meaning rather than pixel-level similarity, traditional signal-based metrics are insufficient to fully capture whether reconstructed images convey correct object-level semantics. Human evaluation provides a necessary validation of meaning preservation.

In this experiment, we evaluate the human-perceived fidelity [218] of transmitted images using semantic symbols at zero SNR, involving 30 participants selected at random, representing various educational and social backgrounds. The questionnaire presented a paired comparison between the original images and the reconstructions. Participants were instructed to evaluate semantic resemblance rather than visual quality, explicitly focusing on object presence, position, size, and contextual consistency. This instruction isolates semantic fidelity from low-level pixel-based artifacts, including blur or texture distortion, thereby aligning the evaluation directly with the objectives of semantic communication. A five-point scale (1 = “Quite Contrast”, 5 = “Quite Similar”)

was employed to quantify perceived semantic similarity. The scale was selected because it provides an ordinal measurement suitable for perceptual similarity judgments.

Images were sourced from the COCO dataset to ensure semantic diversity across object categories and scene types. Multiple comparison instances (10 sample images) were included to mitigate single-sample bias and to evaluate robustness across varied semantic contexts. The selected images have fallen into natural scenes, human figures, animals, vehicles, day-to-day equipment, and indoor scenes.

Ratings were aggregated across participants to compute the average score, which reflects human-perceived fidelity of the transmitted images. The results showed an average rating of 4.19 out of 5, indicating that human perception of the transmitted image’s semantic fidelity was high.

3.4 Limitations

The main limitation of this initial architecture is its generalizability across different applications and domains. Since the semantic symbols are based on the class labels of the COCO dataset, domain knowledge is limited to these symbols. Hence, the meaning generated at the receiver is limited. In technical terms, the distortion in the pixel space of the segmentation maps causes the pretrained GAN to fail in generation, even with a fraction of an error in the segmentation edge. Further, the experiments consider the AWGN channel, which limits the realism of the real-world simulations. The polar coding simulation [204] used in the experiments is a bespoke Python-based simulation, optimized for hardware platforms and particularly efficient field-programmable gate array (FPGA) implementation, which limits the model’s evaluation in practical settings and end-to-end SemCom pipeline. These limitations have led us to consider the design paradigms in the subsequent chapters and more realistic channel conditions.

3.5 Conclusion

The work presented in this chapter marks the first experimental step toward realizing a generative AI-based semantic communication system for wireless image transmission.

Chapter 3. Segmentation guided semantic symbol encoding for generative semantic communication

The motivation is to show how semantic segmentation maps can serve as a minimal symbolic representation of visual data and how these symbols can be transmitted through a noisy physical channel to regenerate meaningful images at the receiver using a shared knowledge base. The experiments confirmed that semantic maps extracted from the COCO–Stuff dataset, when decoded by a GAN, can convey an image’s essential meaning with only a fraction of the bandwidth required by traditional methods, specifically JPEG. Through this preliminary investigation, several practical observations were made. The system demonstrated that transmitting semantic information, rather than raw pixel data, can significantly reduce bandwidth requirements. However, the experiments also showed that the reconstructed output is highly dependent on the accuracy of the transmitted semantic boundaries. These results underscore the importance of preserving edge information for maintaining semantic accuracy in generative reconstruction. The analysis of the joint effects of quantization and channel noise showed that the proposed system remained stable up to around 2.2 dB of channel noise. Beyond this level, the generative decoder failed to reconstruct realistic images, mainly due to severe edge deformation that could not be recovered through filtering. Furthermore, the comparison between lossless and lossy compression revealed that the GAN could not handle semantic maps containing label distributions outside the 182-class vocabulary on which it was trained. This finding indicates that the generalization capability of a semantic communication system depends strongly on the diversity and coverage of its training data. The outcomes of this study laid the groundwork for the design decisions made in the following chapters. The sensitivity of the GAN to boundary noise and edge distortion motivated the introduction of an edge-aware attention module in the Generative AI–Based VQ SemCom framework presented in Chapter 4. The limitations of fixed-label representations led to the development of contrastively optimized codebook-based quantization strategies discussed in chapters 4, 5, and 6. The need for lightweight processing at the transmitter side informed the edge-based implementation in chapter 4, while the importance of maintaining semantic consistency over time guided the extension of these methods to video transmission in chapter 7. We choose Sionna-based LDPC coding in the later chapters for experimental validation, motivated by the need for

Chapter 3. Segmentation guided semantic symbol encoding for generative semantic communication

standards-compliant LDPC codes, differentiable and modular components, and flexible block length support. LDPC is considered the widely used coding method in 5G data channels.

In conclusion, this preliminary study established both the promise and the limitations of using generative AI for semantic communication. It demonstrated that transmitting minimal semantic symbols rather than data is a viable approach. Still, its reliability depends on how well semantic information is represented, protected against channel impairments, and aligned with the shared knowledge between the transmitter and receiver. The findings served as the foundation for the generalized, scalable, and predictive SemCom frameworks developed in the subsequent chapters.

Chapter 4

Edge computing based semantic communication with vector quantization

This chapter introduces a novel VQ-based semantic encoding for edge-based implementation, and this is the first experiment conducted in this research to address the limitations identified in chapter 3. The empirical validations in chapter 3 revealed that semantic encoding based on a data-centric approach, specifically semantic segmentation, limits system performance in the presence of channel noise. Through literature review, VQ is identified as a methodology to satisfy the practical implementation objectives as well as compliance with the conceptual framework developed related to modern theories for SemCom. Typically, edge devices have limited computational and memory resources and operate with limited bandwidth. Hence, the direct transmission of full-resolution images or high-dimensional continuous latent vectors, including autoencoder-based or JSCC systems, may be inefficient and costly. These limitations underscore the importance of developing compact, discrete, and channel-robust semantic representations that reduce communication overhead without compromising the semantic utility of the transmitted content.

The primary objective of this chapter is to demonstrate that VQ-based semantic representations constitute a feasible and effective path toward practical SemCom

deployment in edge computing scenarios. Based on the empirical results, we identified design-related improvements to further develop the architecture. The chapter demonstrates the full pipeline from feature extraction and quantization at the edge to reconstruction and multilevel task inference at the receiver. The system uses a VQ-VAE as the base model for the semantic extraction and transmits the corresponding indices of the quantized vectors through the simulated physical communication channel. The initial design outlines how the proposed architecture, based on VQ can be used in practical settings, including the transmitter and receiver designs and the deployment of the model, including a one-time transfer of the model weights. This serves as the foundation for the further developments introduced in chapter 5. The main contributions of this chapter are summarized below.

- **Edge-oriented VQ-based SemCom architecture:** In view of SemCom design considerations identified in chapter 3, we implement a lightweight SemCom framework using a VQ-VAE model that performs semantic extraction and quantization at the edge and transmits only the codebook indices to a central processing location for the underlying task, significantly reducing symbol overhead compared to JSCC-based methods, as a initial step in designing robust SemCom frameworks.
- **Channel-agnostic design:** The proposed system incorporates LDPC [206] for transmission over AWGN and Rayleigh fading channels, ensuring robustness to noise without SNR-specific retraining, unlike most JSCC systems.
- **Quantified efficiency and complexity analysis:** We provide a detailed computation and latency comparison with JSCC, demonstrating a 25% reduction in transmitted symbols at a LDPC code rate of 0.5 and lower inference latency. With LDPC coding, the system surpasses BPG in traditional metrics, including PSNR and MS-SSIM, as well as in perceptual metrics like LPIPS. These experiments validate that the proposed methodology works well with existing channel coding methods required by modern SemCom systems, as joint coding methods are impractical under heterogeneous channel conditions.
- **Generalization and multi-task capability:** The learned quantized representations

support both image reconstruction and classification tasks with the CIFAR-10 [219] dataset. The classifier is trained on the quantized latent vectors. The results show strong generalization to out-of-distribution datasets, including reconstruction on Tiny ImageNet [220], validating the adaptability to various data distributions.

- **Edge relevance and scalability:** By jointly optimizing semantic compression and quantization to realize the underlying task accuracy, the proposed framework aligns with the computing limitations of modern edge devices, making it scalable for future edge-based semantic-driven IoT applications, empirically validated through high-resolution datasets.

The chapter is organized as follows. Section 4.1 formulates the problem addressed in this practical SemCom implementation. Section 4.2 presents the system model and describes the encoder, quantizer, and decoder components in detail. Section 4.3 provides simulation results illustrating reconstruction quality, noise robustness, and transmission efficiency. Section 4.4 discusses limitations, and section 4.5 concludes the chapter.

This chapter is based on the following article submitted for publication in IEEE Transactions on Consumer Electronics, which is currently under its second round of review. “Edge Computing based Vector Quantized Semantic Communication System for Wireless Image Transmission.” M.U. Lokumarambage, T. Sivalingam, H. Rezaei, N. Rajatheva, and A. Fernando. This was earlier accepted as a poster, at EUCNC 2024 (November).

4.1 Problem formulation

Building upon the theoretical three-layered model developed in chapter 3, the thesis advances it by defining common semantic symbols with semantic coding. As shown in the advanced theoretical model for SemCom figure 4.1, the “Goal” is the intended task of the receiver, including classification, reconstruction, or captioning, related to the “Level C: the effectiveness problem” of Shannon’s communication theory [3]. Extraction of the meaning from the source and semantic coding is associated with “Level B: the semantic problem”. Semantic coding is referred to as the mapping of the retrieved

meaning of the intended message into semantic symbols [221, 222]. Then these minimal semantic symbols are mapped into channel symbols in “Level A: the technical problem”. Noise can be introduced at both levels. In level B, the semantic noise, and in level A, the channel or technical noise.

In level B (semantic problem), two distinct forms of semantic noise are introduced. The first is caused during semantic extraction, where the intended message is mapped into semantic representations. This first form of semantic noise stems from model limitations, contextual ambiguity, or insufficient feature abstraction. The second occurs during semantic coding, where continuous semantic representations are discretized into codebook symbols. This second form of semantic noise related to semantic coding results from quantization error, shared codebook (knowledge base) limitations, or ambiguous nearest-neighbor assignment. Unlike channel noise, both forms operate within the semantic domain and affect meaning integrity prior to physical transmission. Beyond the transmitter-side semantic distortions, channel impairments may corrupt the transmitted discretized symbol mappings, resulting in additional semantic distortion at the receiver.

The common language that both the transmitter and receiver understand shall be a shared semantic symbolic language that effectively translates the images into symbols. In practice, developing a common language of semantic symbols for image-based SemCom systems is challenging because it involves minimizing semantic noise.

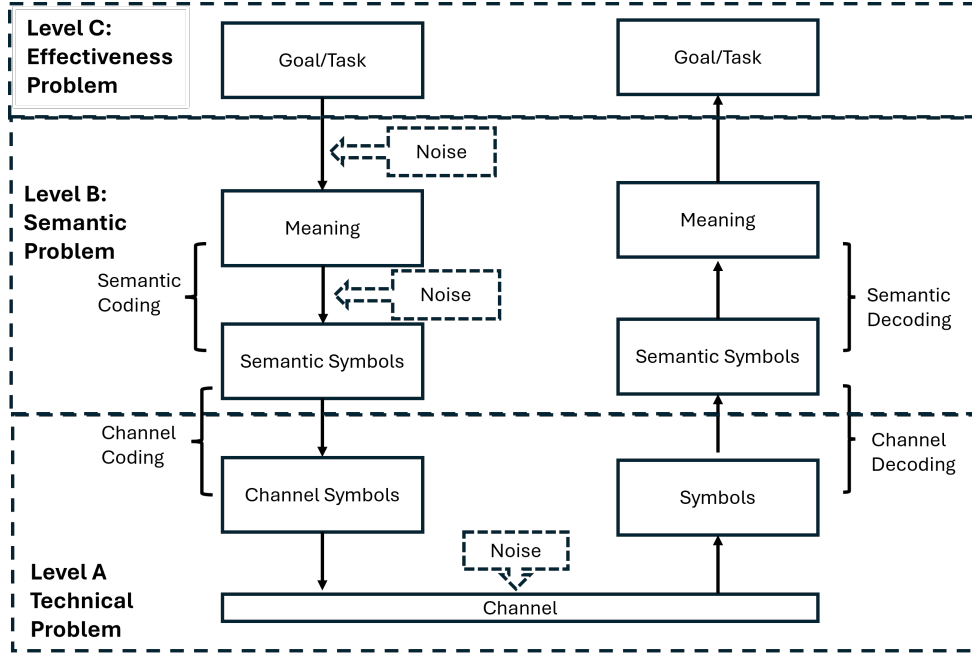


Figure 4.1: Theoretical SemCom system model illustrating semantic noise at the extraction and coding stages and channel noise during physical transmission. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

The problem addressed in this chapter originates from the practical constraints of edge computing environments, where devices must transmit visual information under tight bandwidth budgets and unpredictable wireless channel conditions. The device must communicate only the most essential semantic content to a central processing unit, which in turn performs reconstruction or task-level inference. However, transmitting raw pixel data or high-dimensional continuous features is neither bandwidth-efficient nor scalable. Therefore, the central question guiding this chapter is, “How the resource-constrained edge device encodes and transmits compact, discrete semantic representations of visual data, using minimal bandwidth, while preserving semantic relevance and ensuring robustness to channel noise.”

To address this problem statement, we consider a semantic encoder $f_e(\cdot)$ deployed at the edge, which maps the input image $\mathbf{x}$ to a continuous latent vector $\mathbf{z}_e$. A vector quantization module subsequently maps this latent vector to the nearest codebook element referring to a learned set of embeddings, $\mathbf{E} \in \mathbb{R}^{K \times D}$. This quantization

process produces a discrete latent index i that represents the relevant codebook entry ($\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_K\}$). In contrast to transmitting the full latent vector, only the index i is transmitted through the wireless channel, significantly reducing the number of transmitted symbols. The objective is to ensure that the received quantized vector $\hat{\mathbf{z}}_q$, derived from the channel-decoded index ($\hat{i}$) transmitted through LDPC coding, remains sufficiently representative of the original semantic representation such that the reconstructed image $\hat{\mathbf{x}} = f_d(\hat{\mathbf{z}}_q)$ preserves both perceptual and semantic integrity, where $f_d(\cdot)$ is the semantic decoder.

This chapter formulates a practical SemCom problem in which edge devices perform lightweight semantic extraction and quantization, transmit only discrete symbolic indices, and rely on a generative decoder at the receiver to reconstruct semantically meaningful outputs. This problem lays the foundation for chapters 5 and 6 to design scalable and robust SemCom systems, advancing the SemCom system design knowledge area by bridging theoretical developments and practical implementations in SemCom.

4.2 System model

The proposed SemCom architecture, illustrated in figure 4.2, involves end users transmitting original images to edge servers with limited processing capacity. The feature extraction is offloaded to edge devices because user devices in IoT environments are resource-constrained in processing power, memory, and energy. Performing complex feature extraction directly on these devices can be inefficient, leading to increased latency and reduced battery life. The connectivity from end users to edge servers is via high-bandwidth, low-latency wired intranet connectivity, and is assumed to be error-free. The edge servers extract the features from the images and transmit them through the limited bandwidth wireless channel. The model was trained on a central processing server with significantly higher computational power. Once the encoder-decoder and VQ module training is complete, the classification head is trained as a combined model using the quantized vectors as the training set. As a one-time transfer, the model weights are securely transmitted to the edge servers through a high-bandwidth secure connection. The edge servers initialize the model using the transferred weights. The

shared knowledge base is the codebook learned during the model’s training, which is shared between the edge server and the central server.

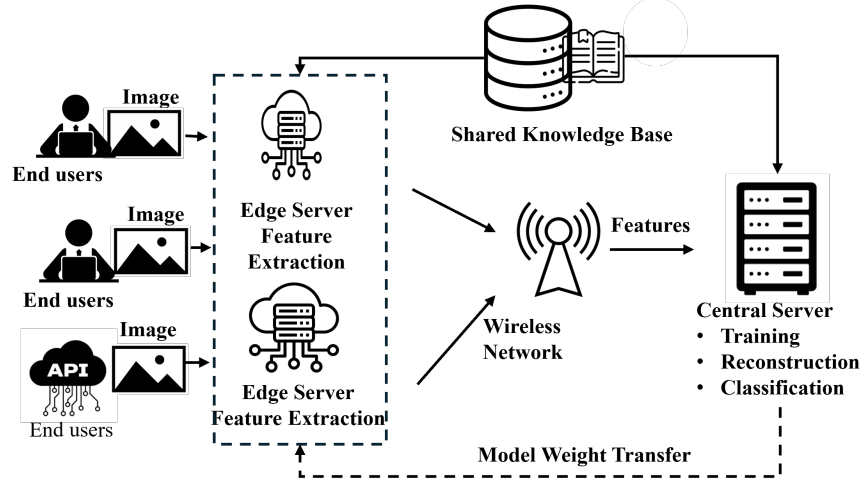


Figure 4.2: Illustration of the implementation of the proposed system .

This architecture is particularly suited for applications where devices like smart cameras, security monitoring systems, or industrial sensors that are connected to localized edge servers through wired connections. This setup is practical in scenarios including smart factories, where cameras monitor production lines and transmit data to nearby edge servers for initial processing. Similarly, autonomous vehicles in a controlled environment may rely on wired connections to edge servers positioned along a specific route or within a parking facility. Once the data is processed and compressed into semantic indices, it is wirelessly transmitted from the edge server to a central processing server or cloud infrastructure. Particularly in cloud infrastructures, where storage capacity is highly costly, this approach is most cost-effective as it reduces the raw data volume stored and processed in the cloud. This is achieved by replacing the continuous latent vector with a discrete embedding vector indices. The discretized representation captures essential features of the input, allowing the model to map important details into a compact code that the decoder can interpret. Each codebook element represents distinct semantic features, enabling effective semantic manipulation and meaningful generation at the decoder. The descriptive details of the main components of the architecture are as follows, and those are discussed in detail in the subsequent sections.

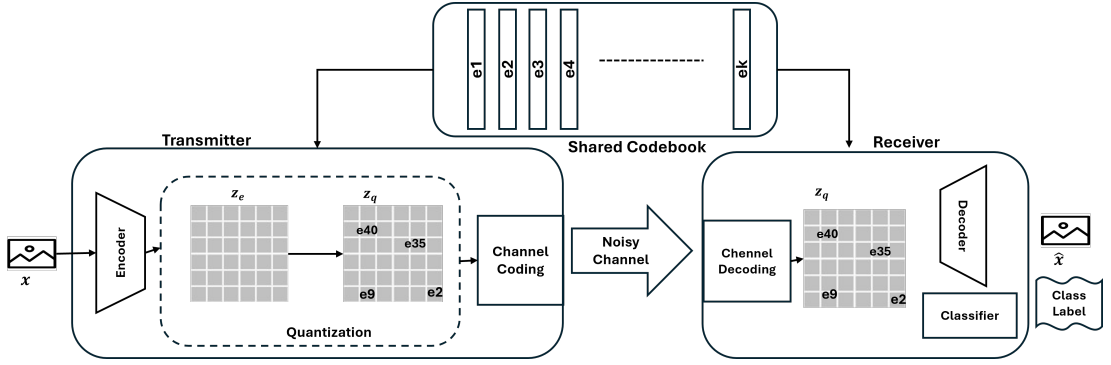


Figure 4.3: The VQ-based index transmission system includes the physical channel with channel coding.

4.2.1 Semantic encoder

Given an input image $\mathbf{x} \in \mathbb{R}^{h \times w \times c}$ where h , w , and c are height, width, and the number of channels respectively, the encoder $f_e(\mathbf{x})$ as illustrated in figure 4.3 maps the input image into a lower-dimensional latent space representation as

$$f_e(\mathbf{x}) : \mathbf{x} \rightarrow \mathbf{z}_e. \quad (4.1)$$

The encoder consists of three convolution layers. The initial convolutional layer extracts 32 feature maps using a 3×3 kernel and a rectified linear unit (ReLU) activation function, followed by a stride of 2 for downsampling. This is followed by a second convolutional layer with 64 feature maps, which further downsamples the feature maps. The final stage uses a 1×1 convolutional layer to project the features into the desired latent dimensionality.

4.2.2 Discrete representation of latent space

The codebook is the embedding vector space, $\mathbf{E} \in \mathbb{R}^{K \times L}$, where K is the number of embeddings and L is the dimensionality of each embedding vector. The codebook vectors, $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_K\}$ are learned during training. The continuous latent space $\mathbf{z}_e$, drawn from a distribution $\mathbb{P}(\mathbf{z}_e)$, is represented by this finite set of embeddings, complying with the minimal semantic representation as defined in the conceptual framework. The

Chapter 4. Edge computing based semantic communication with vector quantization model employs regularization techniques, of dropout and data augmentation, to prevent overfitting. By minimizing Euclidean distance, the continuous latent representations are mapped in to its nearest embedding vector e_i as:

$$\mathbf{z}_q = \arg \min_{e_i \in E} \|\mathbf{z}_e - \mathbf{e}_i\|^2, \quad (4.2)$$

where $\|\cdot\|^2$ is the Euclidean distance. This quantization discretizes the continuous latent space into a finite set of discrete codes as shown in figure 4.3. During backpropagation, the quantization operation is non-differentiable, which prevents direct gradient flow. To overcome this limitation during training, we used a straight-through estimator (STE) to allow gradients from the reconstruction loss to pass through $\mathbf{z}_q$ as if flowing directly through $\mathbf{z}_e$. This mechanism allows the joint optimization of both the encoder and the codebook. The embedding vectors are modeled as learnable parameters and are updated during backpropagation according to the VQ loss in (4.5). As a result, the embeddings are encouraged to align with the encoder outputs, thereby improving their ability to represent the latent space effectively.

4.2.3 Semantic decoder (generator)

At the receiver, the transmitted indices are one-hot encoded to map each index into its corresponding embedding vector in the codebook. The generator reconstructs the input image from the quantized latent representation $\mathbf{z}_q$ as:

$$f_d(\mathbf{z}_q) : \mathbf{z}_q \rightarrow \hat{\mathbf{x}}, \quad (4.3)$$

where $f_d(\mathbf{z}_q)$ denotes the decoder function, and $\hat{\mathbf{x}}$ is the received image. The decoder upsamples $\mathbf{z}_q$ using transposed convolutional layers. The first layer applies a 3×3 kernel with 64 filters, ReLU activation, and a stride of 2. The second layer upsamples further with a 3×3 kernel, 32 filters, ReLU activation, and a stride of 2. Using a 3×3 kernel, the final transposed convolution layer produces the three-channel output image.

4.2.4 Loss function

The training objective of the model comprises with reconstruction loss $\mathcal{L}_{\text{recon}}(\cdot)$ and the VQ loss $\mathcal{L}_{\text{vq}}(\cdot)$ and can be expressed as

$$\mathcal{L}_{\text{total}}(\hat{\mathbf{x}}) = \mathcal{L}_{\text{recon}}(\hat{\mathbf{x}}) + \mathcal{L}_{\text{vq}}(\hat{\mathbf{x}}). \quad (4.4)$$

The VQ loss is implemented as a layer in the VQ module and comprises commitment loss and the codebook loss. The balance between these two terms is controlled by the parameter β . The VQ loss is given as

$$\mathcal{L}_{\text{vq}}(\hat{\mathbf{x}}) = \beta \cdot \mathcal{L}_{\text{commit}}(\hat{\mathbf{x}}) + \mathcal{L}_{\text{codebook}}(\hat{\mathbf{x}}), \quad (4.5)$$

$$\mathcal{L}_{\text{vq}}(\hat{\mathbf{x}}) = \beta \|\text{sg}[\mathbf{z}_e] - \mathbf{z}_q\|_2^2 + \|\mathbf{z}_e - \text{sg}[\mathbf{z}_q]\|_2^2, \quad (4.6)$$

where $\text{sg}[\cdot]$ is the stop-gradient operator. The commitment loss enforces the encoder output to stay close to their corresponding codebook vectors, while the codebook loss adjusts the codebook vectors to better match the distribution of the encoder outputs. This joint optimization allows for capturing meaningful representations from the data distribution and improving the reconstruction quality.

The reconstruction loss quantifies the pixel wise difference between the input image $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$ as

$$\mathcal{L}_{\text{recon}} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2, \quad (4.7)$$

where $\|\cdot\|_2^2$ refers to L_2 norm.

In this initial implementation of the VQ based SemCom, we adopt a two-step training strategy to ensure stable and generalizable representation learning for multitask M2M objectives of reconstruction and classification. In the first stage, the model is trained with the loss function in (4.4) to learn a task-agnostic discrete codebook. This stage emphasizes the formation of the compact, discrete semantic manifold that captures the semantic structure of the data. In the second stage, the learned quantized latent representations are used to train a classification head $f_\phi(\mathbf{z}_q)$. During this phase, both reconstruction and classification losses are active; the codebook parameters remain fixed,

Chapter 4. Edge computing based semantic communication with vector quantization ensuring that learning is confined to the classifier layers. The overall loss is defined as

$$\mathcal{L}_{\text{combined}}(x, \hat{x}) = \mathcal{L}_{\text{recon}}(x, \hat{x}) + \lambda \mathcal{L}_{\text{ce}}(y, f_{\phi}(\mathbf{z}_q)), \quad (4.8)$$

where $\mathcal{L}_{\text{ce}}(\cdot)$ is cross-entropy loss weighted by a hyperparameter λ . This two-step optimization aims to learn a generalizable semantic manifold as the shared knowledge base, where the lower-level component learns a task-agnostic representation, and the upper-level component adopts these representations for task-specific inference. Jointly optimizing all losses from the beginning can cause gradient interference [223], leading to conflicting update directions from reconstruction, VQ, and task objectives, which may destabilize the codebook. By decoupling the representation and task learning stages, the proposed approach prevents codebook instability, maintaining a generalizable codebook that supports multiple tasks (reconstruction and classification) and channel conditions.

4.2.5 Communication framework

The wireless link from the edge server to the central server is the segment where channel noise is present. To simulate the channel characteristics of mobile scenarios like autonomous vehicles, the channel is modeled as a Rayleigh fading channel, characterized by multipath propagation due to user mobility. The proposed system employs LDPC coding to mitigate channel noise and enhance data integrity in such mobile scenarios. To simulate fixed or indoor deployments, including smart factory settings, the channel is approximated as an AWGN channel in this experimental validation, where noise is the dominant factor impacting signal quality. In contrast to the hardware-oriented multi-kernel polar coding architecture explored in chapter 3 [203–205], the experimental evaluation in this chapter requires a channel-coding scheme that is both compliant with standards and seamlessly integrated with deep learning-based SemCom pipelines. Sionna’s LDPC transmitter implementation provides these capabilities. It follows the 5G data-channel specifications and enables stable, reproducible simulations of Rayleigh fading and AWGN channels. We evaluated the system performance using code rates of $1/2$, $2/3$, $3/4$, and $5/6$, carrying 512 information bits with simulated AWGN and fading

Chapter 4. Edge computing based semantic communication with vector quantization channel with QPSK.

4.3 Simulation results

The training details and simulation results on the quality of transmission, generalization, bandwidth efficiency, and model complexity are illustrated in this section.

Implementation details

The model is trained with the CIFAR-10 [219] dataset with different codebook sizes of $K = 32, 64$, and 128 embeddings within a $L = 16$ dimensional discrete vector space in the experimental setup. In the initial phase, we trained the model using the CIFAR-10 dataset, comprising 50,000 training images and 10,000 testing images, over 200 epochs. To illustrate the concept, we evaluate the model's classification accuracy using a test set of 10,000 images at various SNR levels. The learning rate of 0.001 and Adam optimizer is used while setting β to 0.25 based on the original article on VQ-VAE [15]. For the combined model training we selected value 1 for $\mathcal{L}_{\text{recon}}$ and 0.5 for $\mathcal{L}_{\text{ce}}$. Further, to evaluate the model's scalability, it is trained on 8000 Flickr8K dataset images, resized to $256 \times 256 \times 3$. Data augmentation is applied with random geometric transformations, intensity variations, and horizontal flipping to improve model robustness and generalization, which increased the training dataset size to 26472.

4.3.1 Experimental results on CIFAR-10 dataset

The classification experimentation is conducted to initially identify the factors that contribute to the task achievement in SemCom and understand the design considerations. The classification head achieved a training accuracy of 85.52% with $K = 64$ when trained with clean images. We added AWGN noise to the test dataset at varying SNR levels and measured the classification accuracy, and the results are illustrated in figure 4.4 for different codebook sizes. The results indicated that the codebook size of $K = 32$ resulted in the best classification accuracy, suggesting that smaller embedding sizes lead to more effective representation and generalization. This is likely due to reduced latent

Chapter 4. Edge computing based semantic communication with vector quantization space complexity, which enhances robust feature extraction and minimizes overfitting. A further evaluation of this aspect is conducted experimentally in chapter 6, examining the effects of codebook size and its usability. To compare classification accuracy with traditional image transmission, a classification model is trained on JPEG, achieving 98.09% accuracy. To simulate the noisy channel conditions, AWGN noise is directly added to the test set images, and the accuracy is evaluated, dropping to approximately 10%.

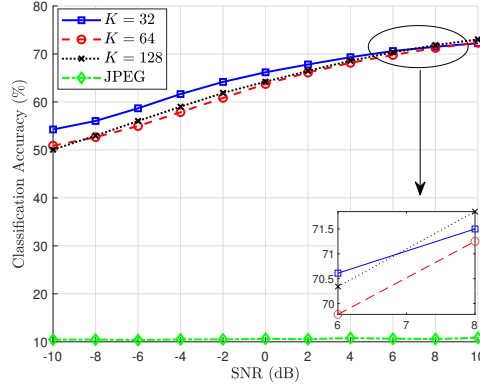


Figure 4.4: Classification accuracy vs. SNR on the CIFAR-10 test set under AWGN noise for $K = 32, 64, 128$, and $L = 16$

The LDPC coding-based channel simulation is developed with Sionna, and its BERs are plotted in figure 4.5 for different coding rates. Further, a JSCC scheme is implemented to compare the results, and both are used throughout this thesis for comparison. We reproduced the results for the CIFAR-10 dataset for the model implemented in [71]. We used the model trained at an SNR level of 10 and with a channel-per-pixel (CPP) value of 1/12. A limitation of this methodology is that the model requires training for a specific SNR, which may not be practical in a dynamic IoT environment where channel conditions are constantly changing.

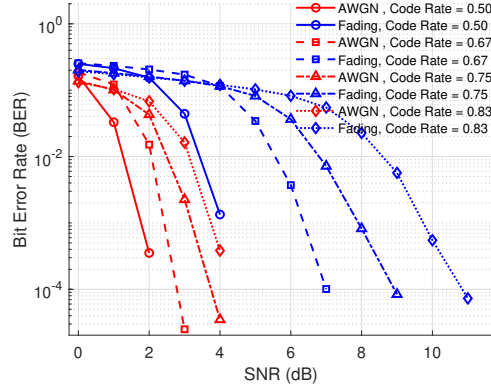
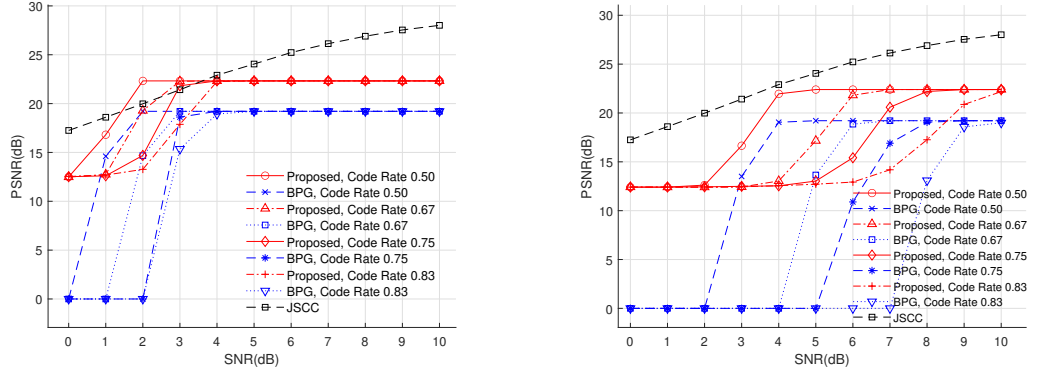


Figure 4.5: Error correction performance for simulated AWGN and fading channels with Sionna for information bit size 512 with different coding rates for LDPC using QPSK

Figure 4.6a illustrates the received image quality for varying channel coding rates over an AWGN channel for sample images from the CIFAR-10 test set. The proposed SemCom system achieves better reconstruction quality than BPG at similar compression rates. BPG was selected because of its high compression ratio, which is much smaller than that of JPEG for comparable quality. To ensure a fair comparison, we evaluate the number of transmitted complex symbols for both JSCC and the proposed method, considering that the JSCC scheme transmits the $8 \times 8 \times 8$ tensor as uncoded complex symbols, with 8 bits per symbol. For JSCC, the total number of transmitted complex symbols corresponds to the size of the source tensor: $N_{\text{symbols, JSCC}} = 8 \times 8 \times 8 = 512$ symbols. The proposed method, transmitting 64 indices requiring 6 bits per index, results in a total uncoded bit size of $64 \times 6 = 384$ bits. Using LDPC coding with a rate $R_c = 0.5$, the total number of coded bits becomes $\frac{384}{0.5} = 768$. With QPSK modulation, $N_{\text{symbols, proposed}} = \frac{768}{2} = 384$ symbols. The proposed method achieves a lower transmission overhead compared to JSCC, requiring only $\frac{384}{512} = 75\%$ of the transmitted symbols.



(a) AWGN channel

(b) Fading channel

Figure 4.6: (a) Image quality (CIFAR-10) at different SNR levels over AWGN channel with LDPC coding for QPSK, (b) Image quality (CIFAR-10) at different SNR levels for flat fading channel with LDPC coding for QPSK (the proposed only requires 75% of the transmitted symbols compared to JSCC with code rate = 0.5).

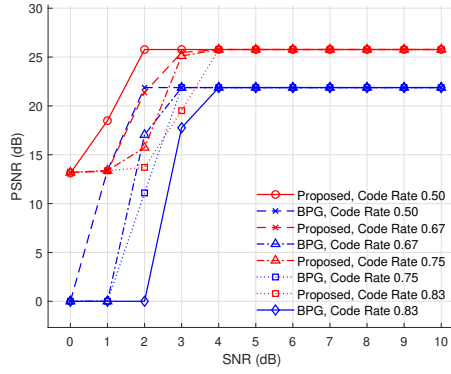
The comparative analysis of the proposed scheme, BPG, and JSCC across different code rates with a fading channel simulation is shown in figure 4.6b. The proposed scheme demonstrates superior performance in low SNR regions, particularly at code rates of 0.5 and 0.67, achieving higher PSNR than BPG. This makes VQ-based systems more suitable for noisy channels with limited bandwidth, where maintaining reconstruction quality is crucial. BPG's performance weakens at a code rate of 0.83, leading to significant PSNR degradation. Traditional image coding methods, including BPG, are more susceptible to channel conditions because they rely on block-based prediction, transform coding, and entropy coding methods. In contrast, the proposed scheme can generate the image even if part of the received features are corrupted. This empirical validation has made VQ based SemCom systems a better solution for further improvement.

4.3.2 Generalization

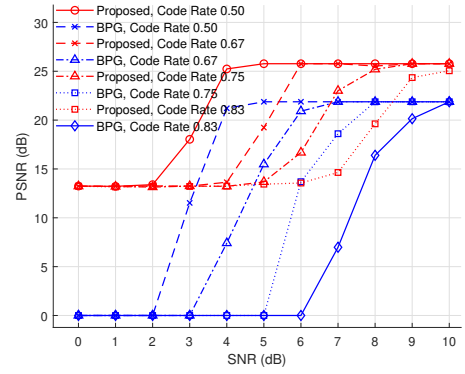
To assess the generalization capability of the VQ based systems, the system is evaluated on the subset of RGB images from the Tiny ImageNet dataset [220]. Tiny ImageNet exhibits substantially greater image complexity, diversity, and scale, which can provide a experimental validation of the hypothesis of the model's ability to adapt to novel data

distributions not encountered during training. The evaluation of PSNR figure 4.7a and figure 4.7b demonstrates that the proposed model generally outperforms the BPG in terms of reconstruction quality across various SNR values in both AWGN and fading channels. The structural similarity, as measured by MS-SSIM in figure 4.8a and figure 4.8b, highlights the ability of the system to retain structural details in reconstructed images more effectively than the traditional image coding schemes, including BPG. In AWGN channels, the proposed system consistently achieves higher MS-SSIM values than BPG. This suggests that a VQ-based SemCom system can capture and maintain important spatial structures. In fading channels, the proposed method shows similar robustness; the gap narrows at lower SNR values due to the higher channel-induced distortion, suggesting further improvements to withstand channel noise, which this research has introduced in the next chapter. This robustness across channel conditions implies that VQ-based SemCom systems could be particularly advantageous in scenarios requiring high structural fidelity, such as medical imaging or remote sensing, where fine details are crucial for interpretation.

The LPIPS metric results, illustrated in figure 4.9a and figure 4.9b, show the perceptual quality of reconstructed images, where lower scores indicate higher fidelity. Across all SNR values and coding rates in both AWGN and fading channels, the proposed SemCom system consistently outperforms BPG with lower LPIPS scores. Results suggest that the VQ-based system effectively preserves high-level semantic features crucial to semantic similarity. Despite low SNRs, the VQ-based system maintains perceptual quality over BPG, suggesting that its discrete codebook representation robustly captures essential features.

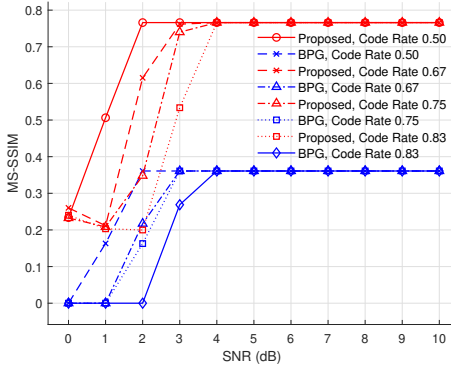


(a) AWGN channel

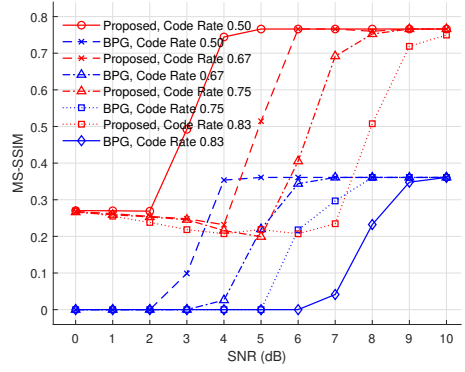


(b) Fading channel

Figure 4.7: Tiny ImageNet dataset : (a) PSNR at different SNR levels over AWGN channel with LDPC coding for QPSK, (b) PSNR at different SNR levels for flat fading channel with LDPC coding for QPSK.

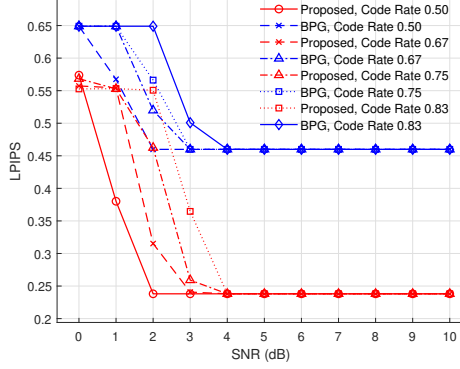


(a) AWGN channel

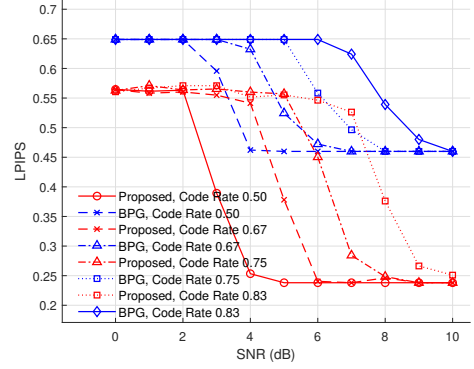


(b) Fading channel

Figure 4.8: Tiny ImageNet dataset : (a) MS-SSIM at different SNR levels over AWGN channel with LDPC coding for QPSK, (b) MS-SSIM at different SNR levels for flat fading channel with LDPC coding for QPSK.



(a) AWGN channel



(b) Fading channel

Figure 4.9: Tiny ImageNet dataset : (a) LPIPS at different SNR levels over AWGN channel with LDPC coding for QPSK, (b) LPIPS at different SNR levels for flat fading channel with LDPC coding for QPSK.

4.3.3 Scalability and qualitative illustration

The proposed model was further evaluated on Kodak and CelebA-HQ datasets to assess scalability with image dimension of $256 \times 256 \times 3$. As shown in figure 4.10, the system preserves perceptual and structural fidelity across diverse image domains and channel conditions, confirming its scalability for real-world edge multimedia applications. However, we have retrained the model on a higher-resolution dataset, which has led to the design decision in chapter 5 to develop a scalable model that can operate on up to 4K images.

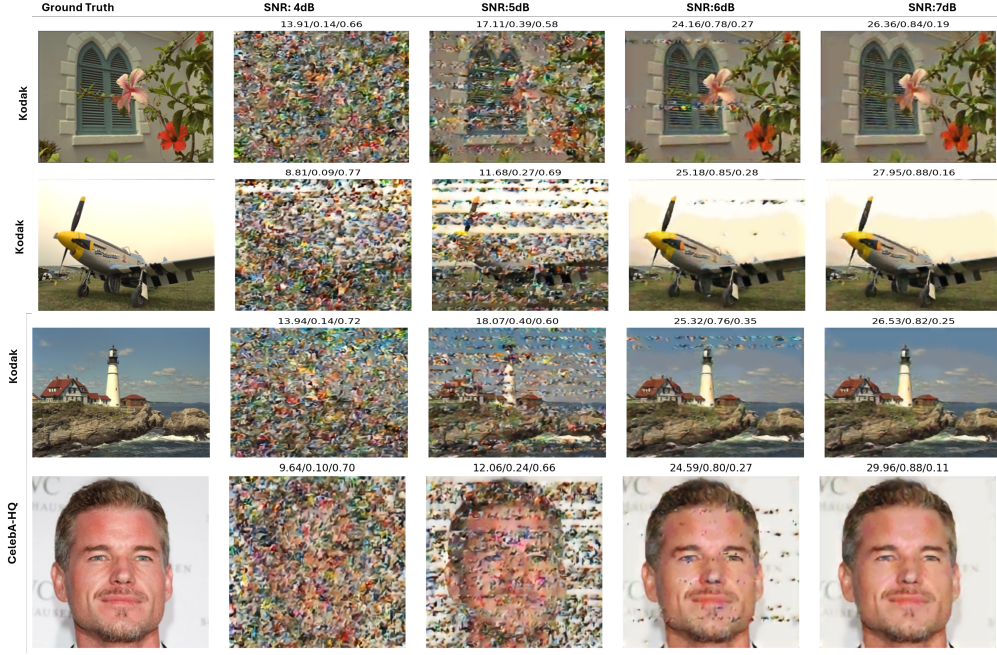


Figure 4.10: Illustration of image reconstruction quality for Kodak and CelebA-HQ dataset for Rayleigh flat fading channel with LDPC coding with code rate of 0.5. The top of each image shows the values for PSNR/MS-SSIM/LPIPS.

4.3.4 Computational complexity analysis

This section presents a detailed computational complexity analysis of the proposed system, the conventional BPG compression scheme, and JSCC. The objective is to quantify the computational workload in terms of million floating-point operations (MFLOPs). This assessment of overall computational complexity will evaluate the efficiency of the proposed system in edge scenarios. In modern edge devices, computational efficiency and data transmission are crucial, particularly in devices such as smart cameras, IoT sensors, and health monitoring systems. These devices operate under strict resource constraints, necessitating effective data compression techniques to balance processing power and transmission load. The proposed system minimizes transmitted data by quantizing inputs into compact indices, reducing the number of transmitted symbols.

From a computational perspective, the MFLOPs analysis in table 4.1 illustrates that the proposed model, while it incurs lower complexity than JSCC with a 25%

Table 4.1: Computational cost and inference latency for image reconstruction

Metric	Ours	JSCC	BPG
Model complexity (MFLOPs)	2.20	6.49	0.22
Inference latency (ms)	1.2–1.6	10–12	0.2–0.4

reduction in symbols (with LDPC coding with code rate of 0.5), is advantageous over both complexity and data rates. Despite the higher encoding complexity compared to BPG, the transmitted image fidelity of our model under noise is comparatively higher, striking a balance between complexity and fidelity for edge devices.

Inference latency was estimated based on the decoder parameter count and typical FLOP-to-time ratios measured on an NVIDIA T4 GPU, assuming a realistic effective throughput of approximately 2.4 TFLOPs at 30% utilization for single-image inference. The proposed method shows lower latency than the JSCC and shows a high latency due to its iterative refinement. BPG shows the lowest latency, due to discrete cosine transform (DCT), quantization, and entropy coding.

4.4 Limitations and future works

The limitations of this chapter are that the model uses the basic VQ-VAE to learn a quantized latent representation, without employing any optimization strategies. This was experimented with as the first step towards the research experiments, and later, the optimizations mentioned in chapters 5 and 6 are introduced to align the design more towards the robust communication objective. Despite the model performing well in reconstruction at the receiver, its classification accuracy drops significantly at low SNR (approximately 50% in accuracy), indicating sensitivity to noise. This motivated the need for a robust model with contrastive learning at multiple levels, as described in chapter 5, that models and handles the noise at the semantic level. Although reconstruction and classification are supported, the system is not evaluated in more complex downstream tasks, including object detection (machine-centric) and image captioning (human-centric). The model required retraining when scaled to higher-resolution datasets, increasing the initial implementation cost and limiting its practical

scalability. This constraint motivated the design choice in chapter 5 to develop a more flexible, resolution-agnostic architecture that operates efficiently on images up to 4K without necessitating full retraining. Further, future work will focus on enhancing the model architecture to improve the perpetual fidelity of the transmitted image and extend practical applications to computer vision tasks, such as indoor scene understanding with higher resolution images and videos.

4.5 Conclusion

This chapter proposed a novel SemCom architecture with edge inference, where edge devices extract semantic features. These devices map discrete vectors to relevant codebook indices, which are shared as a knowledge base between the transmitter and receiver. Transmitting the relevant indices to a central location significantly reduces the data transmission rate and conserves wireless resources, making it ideal for IoT environments. Leveraging a discrete codebook learned during training, we evaluated designs with various embedding dimensions, focusing on reconstruction and classification accuracy. The proposed system demonstrates superior performance compared to BPG when the indices are transmitted with LDPC coding over AWGN and Rayleigh fading channels, achieving consistently higher perceptual quality as evaluated by LPIPS. Using an out-of-distribution dataset, we tested the adaptability of the learned representations in the latent space and the robustness of the quantized codebook under new, diverse inputs. The system achieved a training accuracy of 85.52% for the classification task trained with latent vectors, with classification accuracies of 50.88% and 72.35% at low and high SNRs, respectively, while the JPEG model averaged only 10% across noise levels. The proposed system introduces more complexity than BPG, but it maintains higher fidelity in transmitted images and significantly reduces overall complexity and transmission overhead compared to JSCC, striking an effective balance between efficiency and semantic fidelity in edge environments, validating the VQ based SemCom systems applicability on practical scenarios for further research.

Chapter 5

Contrastive vector-quantized generative semantic representation for robust semantic communication

Based on the foundational work established in chapters 4 and 5 on using visual generative AI for decoding symbolic semantics in image transmission, this chapter explores developing a robust SemCom framework that abstracts symbolically encoded information into perceptually grounded visual representations. The studies in previous chapters revealed several constraints that limit the scalability and reliability of SemCom systems, including channel noise, limitations on the symbolic representations, and the dependency of semantic symbols on the training dataset. To address these limitations, this chapter presents a more generalized, scalable, and channel-agnostic system architecture that is governed by the modern SemCom theories defined in section 2.1. The development of a practical, scalable semantic communication system introduces several fundamental challenges. These challenges span semantic representation, compression, robustness, data diversity, and evaluation.

Semantic feature extraction involves obtaining representations that capture the

essential meaning and structure of visual information while remaining valid across different data distributions [9, 224]. Achieving generalizable feature representations is a fundamental challenge for ensuring that the system performs reliably when exposed to previously unseen images. The second technical challenge is semantic compression, which is reducing the dimensionality of semantic features without losing their meaning or introducing significant semantic noise [225]. Maintaining robustness to channel noise is another central issue. Current implementations of SemCom, which are JSCC or deterministic deep learning model based SemCom designs, struggle to develop channel-agnostic semantic communication systems that can perform reliably across diverse practical channel conditions and seamlessly integrate with established channel coding techniques. This limitation contrasts with JSCC, where model performance is often tightly bound to the SNR used during training [71]. The limitations of available datasets further constrain system generalization [9, 226], thereby restricting the diversity of semantic information required to train a communication system capable of representing a wide range of visual contexts. Finally, the evaluation of semantic communication remains an open problem. Conventional image quality metrics fail to measure how well meaning is preserved after transmission, particularly when channel distortions are present. Therefore, the challenge is the need for metrics that jointly assess semantic preservation, perceptual quality, robustness to noise, and task-level performance so that the overall effectiveness of a semantic communication system can be properly understood. The SemCom framework proposed in this chapter, which is the main contribution of this thesis, is designed to address these challenges, and the novel technical contributions are summarized below.

- We formulate a theoretical SemCom model aligned with the semantic information theory, which is aligned with the defined conceptual framework in section 1.1.1. The theoretical framework builds on the three-layered framework introduced in chapter 3, and it satisfies all the features of the defined conceptual framework.
- In accordance with the developed theoretical framework, a novel generative AI-based practical SemCom system called SVQ-WGAN is designed. Consequently, we propose

semantic coding to achieve higher semantic compression and to represent image semantics using minimal symbols (a semantic language). The transmitter and the receiver agree on a learned codebook, representing the shared knowledge base, to interpret image semantics. Extracted image semantics are mapped into codebook vectors, and the transmitter conveys the relevant codebook indices of the quantized latent over the communication channel.

- Traditional SemCom systems, including autoencoders, use neural-based compression, where the representation’s efficiency largely depends on the training dataset. We use a Wasserstein distance [227] based discriminator to quantify the lowest cost required to convert real data distribution to the generated data, so the generated data closely represent the data distributions, not the individual image semantics, which largely contributes to the generalization of the system model. This enhances the language grammar for better reconstruction quality and semantic fidelity, and provides a more meaningful quantitative measure of the difference between the generated and real images, leading to a more stable training process than traditional GANs.
- Motivated by the findings in chapter 3, where object semantics are primarily based on accurately capturing edges, we designed a novel attention mechanism based on the Sobel operator to emphasize low-level object edge details.
- The SVQ-WGAN framework is compatible with standard digital communication systems, as it operates entirely on discrete latent representations. The vector-quantized output consists of index sequences that can be directly mapped to symbol streams, making the system compatible with conventional channel coding schemes, including LDPC, and modulation methods, including quadrature amplitude modulation (QAM). To explore a more realistic simulation environment, we use the NVIDIA Sionna simulator in this implementation, enabling us to evaluate channel coding techniques under conditions closer to the real world and resulting in more realistic performance evaluations.
- This chapter introduces a novel application of a contrastive learning strategy tailored for semantic communication with multi-level contrast mechanisms across the pixel

level, latent space, perceptual features, and semantic task outputs. We achieve this through spatial augmentation of training data, injecting noise into the quantized latent vector, index masking, and Wasserstein distance-based discrimination. This contrastive mechanism is explicitly tailored for SemCom, optimized for noisy wireless communication environments, compared to the traditional contrastive models SimCLR [136] and MoCo v3 [137], which operate in clean, noise-free environments. This further strengthens robust generalization to noisy channels without requiring explicit error correction at the receiver, which would add complexity and latency.

- To address the limitations identified by the literature review, this chapter proposes a novel metric called RI quantifying the human perceived fidelity and the channel noise characteristics, which is tailored specifically for SemCom systems, measures the robustness quantitatively, and illustrates the proposed system’s semantic and channel noise robustness. SVQ-WGAN supports both machine-centric and human-centric task performance, validated through a set of classification and image captioning experiments. Further, the model is validated using a novel task-oriented semantic distortion measurement to provide a meaningful evaluation of the usability of the semantic symbols towards the task fulfillment, and it empirically demonstrates that the proposed model outperforms BPG in an experimental setting. This matrix quantifies object identification and semantic placement within the image.

The structure of this chapter is as follows. Section 5.1 presents the problem formulation, including the concept and the theoretical justifications that explain how the key components, quantized latent representation, index transmission, adversarial training, and contrastive regularization, jointly contribute to semantic compression, preservation, and robustness under channel noise. Section 5.2 presents the system model, including the main components and the training strategy. Section 5.3 provides a theoretical justification for the design decisions and implementation. Section 5.4 presents the simulation results, and section 5.5 analyzes the system’s limitations, followed by section 5.6, which concludes the chapter.

This chapter is based on the following open-access article published in the IEEE

Transactions on Machine Learning in Communications and Networking journal under a Creative Commons Attribution (CC-BY 4.0) license: M. Lokumarambage, T. Sivalingam, F. Dong, N. Rajatheva and A. Fernando, "Generative AI-Based Vector Quantized End-to-End Semantic Communication System for Wireless Image Transmission." vol. 3, pp. 1050-1074 doi: 10.1109/TMLCN.2025.3607891'

5.1 Problem formulation

Along with these theoretical developments with the three-layered communication problem presented in chapter 4, in this chapter, we reformulate our image transmission problem in resource-constrained environments as a world model-inspired SemCom approach. Semantic information theory defines semantic entropy [221] $H_{\text{sem}}(s)$ as $-\log_2 \sum_{w \models s} p(w)$, where w represents a world model as defined in section 2.3.2, $\models$ denotes the propositional satisfaction relation, and $\sum_{w \models s} p(w)$ is the probability of all world models in which s holds "true". Within a probabilistic framework, world models correspond to possible system states or interpretations. Similarly, each discrete code in the codebook represents a distinct feature or pattern in the data, allowing different codes to capture different aspects of an image. Image reconstruction can therefore be conceptualized as identifying the "world model" that most accurately represents the high-level features of the data distributions. Further, semantic rate distortion is defined as follows:

$$R_{\text{sem}}(D) = \min_{p(\hat{s}|s)} [H_{\text{sem}}(s|\hat{s})] \quad (5.1)$$

$$\text{subject to } \mathbb{E}[D_{\text{sem}}(s, \hat{s})] \leq D$$

where s denotes the original message, $\hat{s}$ is the reconstructed message after encoding and decoding, $p(\hat{s}|s)$ represents the conditional probability distribution of the reconstructed message given the original, $H_{\text{sem}}(s|\hat{s})$ is the conditional semantic entropy indicating the remaining uncertainty about s given $\hat{s}$, and $D_{\text{sem}}(s, \hat{s})$ is a semantic distortion measure that quantifies the difference in meaning between s and $\hat{s}$, such as the bilingual evaluation understudy (BLEU) [228] score or perceptual losses in images. The SemCom

system encodes images into symbolic representations. Subsequently, it decodes or generates them back into a visual format, ensuring that semantic rate distortion does not compromise the intended objective. This methodology is essential for applications in computer vision, image search, automated captioning, and content-based retrieval. However, addressing semantic rate distortion introduces several challenges in practical scenarios.

The fundamental theoretical underpinning of the SemCom is the minimal semantic language according to the theories defined in 2.1. The common shared semantic symbol language in this architecture is the learned codebook, and each codebook vector represents a specific pattern or feature the model has learned to recognize. The model incorporates a Wasserstein discriminator to guide the generation process. The underlying intuition of the codebook-based SemCom is to merge semantically equivalent features of an image into a single codebook entry, which is similar to matching similar contexts in natural language mapped into a single message. This approach yields a compact subset of messages that preserves the intended meaning with minimal semantic loss. The resulting reduced subset achieves the lowest entropy and the shortest expected coding length.

Mapping the above theoretical grounds to an implementation perspective, the discriminator ensures the language symbols are expressive and better reflect the knowledge within the defined language model, which results in the symbols being connected, semantically expressive, consistent, and meaningful to enhance the “Grammar” during training. Thus, it reduces semantic noise by making symbol combinations more closely resemble features, thereby achieving the goal of defining accurate semantic language or symbols for images.

5.2 System model

The system model consists of five main components: encoder, quantizer, noisy communication channel, critic, and generator. The learned codebook acts as the shared knowledge base. The image data is represented as quantized vectors, and the Wasserstein generative adversarial network (WGAN)s stable training process provides

improved control and interoperability for the generated images, with an efficient representation of information. The model employs quantized vectors as conditioned inputs, which provides a more controlled generative process than the random noise used in traditional GANs. Each vector can correspond to specific attributes or high-level features in the generated images, enabling a precise generative process when the knowledge is efficiently represented in the generator network, producing higher-quality images with reduced semantic noise. It is beneficial in applications where specific characteristics of the output need to be controlled or analyzed. The training phase of the proposed architecture is depicted in figure 5.1 and the inference phase in 5.2.

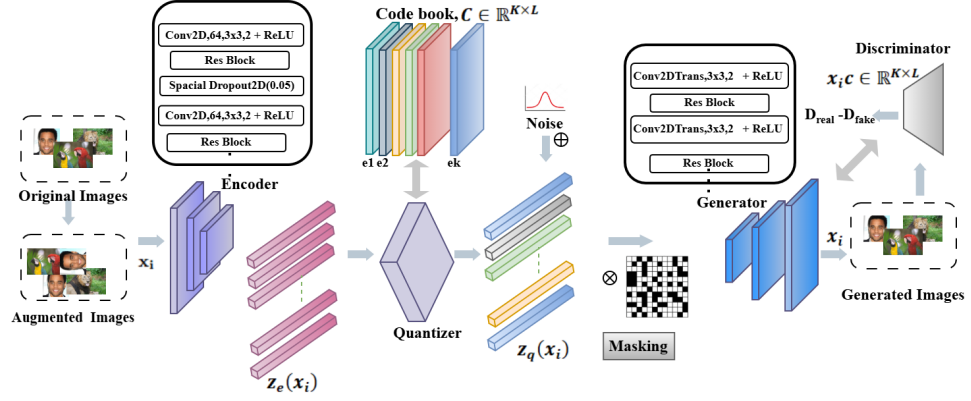


Figure 5.1: Training phase of SVQ-WGAN: quantization of continuous features using the learned codebook, noise injection in the quantized latent space, and random element masking with probability $\mathbb{P}_{mask}$. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

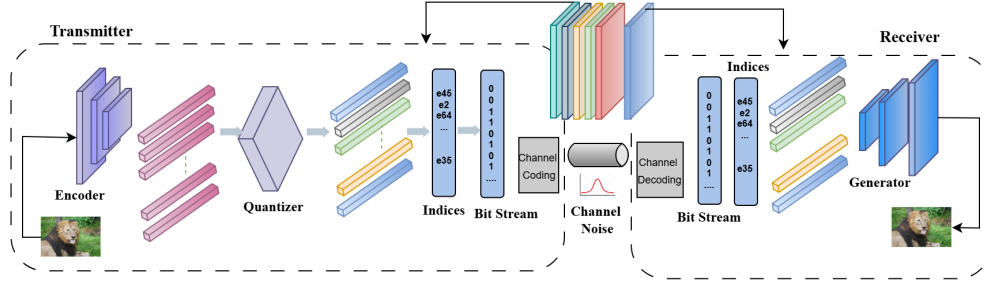


Figure 5.2: Image transmission over a noisy channel: codebook indices from the semantic-encoded and quantized image are mapped to a bitstream and transmitted over AWGN and fading channels with LDPC channel coding. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

5.2.1 Semantic encoder

Given an i^{th} input image $\mathbf{x}_i \in \mathbb{R}^{h \times w \times c}$, the encoder E parameterized by θ_E captures the semantic features of the input and maps it to a continuous latent representation $\mathbf{z}_e(\mathbf{x}_i) \in \mathbb{R}^F$ as

$$\mathbf{z}_e(\mathbf{x}_i) = E(\mathbf{x}_i; \theta_E), \quad (5.2)$$

where F is the number of features. The encoder consists of convolution and residual blocks, as illustrated in figure 5.1.

Spatial attention based residual blocks

The intuition of the design of the residual block is to highlight object edges using a lightweight Sobel-based attention mechanism, thereby preserving high-frequency features critical to semantic fidelity. The energy of an image, $E_n(x, y)$, is defined as the magnitude of its gradient. Image energy highlights the rapid intensity changes at object edges [229]. To underline this concept, we introduce a spatial energy-based attention mechanism to highlight the object edges in the image. Traditionally, the Sobel operator is used to detect edges of objects. We use this to enhance the object's edge information. The gradient magnitude is computed along the x and y directions using Sobel kernels P_x and P_y . The gradient magnitude S_g is then given by $S_g = \sqrt{S_x^2 + S_y^2}$, where $S_x = x * P_x$ and $S_y = x * P_y$. We used a sigmoid normalization function on

input features to get the attention weights, which are represented by $F' = A \odot F$, where $\odot$ is element-wise multiplication. The attention map is multiplied by the feature representation to emphasize object edge regions. A convolutional layer is used to ensure dimensional consistency. The final residual output is produced by adding the attention-modulated feature map to the skip connection. This mechanism directs the model's focus toward high-frequency areas, thereby preserving object boundaries under high levels of compression. One argument against this methodology is that it may under-represent smooth or textured areas. However, this limitation is mitigated through residual learning and convolutional integration across layers. To address this, we also use multilevel contrastive learning to train the end-to-end model, which reinforces semantic consistency across scales. The proposed attention mechanism captures low-level structural cues, including edges and gradients, and integrates them into higher-level semantic representations via residual attention connections. The block is designed with three convolution layers, the first applies a convolution with ReLU activation and same padding to preserve spatial dimensions, followed by batch normalization for stable, efficient training, as illustrated in figure 5.3.

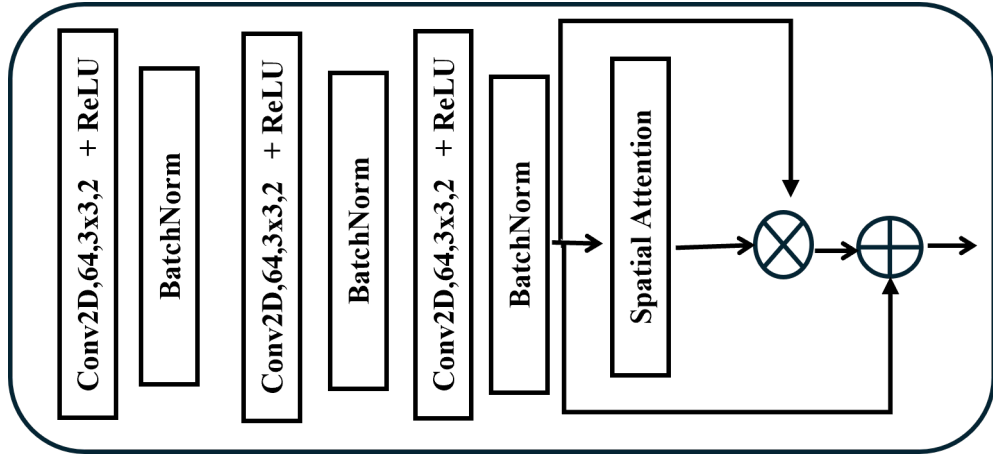


Figure 5.3: Residual block. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

Semantic coding through quantization and codebook

In this architecture, the codebook, $\mathbf{C} \in \mathbb{R}^{K \times L}$ is the shared knowledge base which contains K discrete code vectors (embeddings) with dimension L . These embeddings are trainable. The encoder maps the image to a continuous latent vector $\mathbf{z}_e(\mathbf{x}_i)$, which is then compared with the codebook entries. The quantization process calculates the minimum distance vector in the codebook $\mathbf{c}_k$ to represent each vector of the encoded image, $\mathbf{z}_e(\mathbf{x}_i)$ by minimizing the Euclidean distance. The quantized representation is given by :

$$\mathbf{z}_q(\mathbf{x}_i) = VQ(\mathbf{z}_e(\mathbf{x}_i); \theta_{VQ}). \quad (5.3)$$

The embeddings are trainable and updated during training. The continuous latent vector $\mathbf{z}_e(\mathbf{x}_i)$ is compared to each code vector in $\mathbf{C}$. The quantization process selects the code vector $\mathbf{c}_k$ that is closest to $\mathbf{z}_e(\mathbf{x}_i)$ by minimizing the Euclidean distance.

$$\mathbf{z}_q(\mathbf{x}_i) = \arg \min_j \|\mathbf{z}_e(\mathbf{x}_i) - \mathbf{c}_j\|_2^2 \quad \text{for } j \in \{1, 2, \dots, K\}, \quad (5.4)$$

This selected vector $\mathbf{c}_k$ is then used as the quantized representation of the image, where $\|\cdot\|^2$ is the Euclidean distance.

To further compress the image while preserving the wireless channel's bandwidth, we use semantic coding. The derived index sequence j of the quantized latent of the image is the semantically coded image transmitted via the physical channel. There are two loss terms associated with this process: loss term $\mathcal{L}_{VQ}$. The intuition behind these loss terms is to encourage the codebook vectors to align with the encoded output (represented by the codebook loss) and to encourage the encoder to encode images close to the codebook vectors (represented by the commitment loss). The total VQ loss is given by

$$\mathcal{L}_{vq} = \mathcal{L}_{\text{codebook}} + \mathcal{L}_{\text{commit}}. \quad (5.5)$$

Codebook loss is given by

$$\mathcal{L}_{\text{codebook}} = \|\mathbf{z}_q(\mathbf{x}_i) - \text{sg}[\mathbf{z}_e(\mathbf{x}_i)]\|_2^2, \quad (5.6)$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation, and the commitment loss is given by

$$\mathcal{L}_{\text{commit}} = \alpha \cdot \|\mathbf{z}_e(\mathbf{x}_i) - \text{sg}[\mathbf{z}_q(\mathbf{x}_i)]\|_2^2, \quad (5.7)$$

where α is a hyperparameter that controls the weight of the commitment loss. The quantization step introduces a non-differentiable operation, preventing gradients from propagating directly during backpropagation. To address this issue, a STE is employed during training. The STE allows the gradients from the reconstruction loss to flow through $z_q(x_i)$ as if they were passing through $z_e(x_i)$, thereby enabling the joint optimization of the encoder and the codebook. The embedding vectors are treated as trainable parameters and updated via the vector quantization loss, enabling the learned codebook to progressively better represent the latent feature space.

5.2.2 Semantic decoder

In this proposed architecture, the generator serves as the semantic decoder, generating the image from the transmitted semantic representations. Together with the semantic encoder, it forms an end-to-end system that learns to convey the meaning of visual information rather than its raw pixel content. Unlike deterministic reconstruction SemCom systems that use autoencoders, which primarily minimize pixel-wise reconstruction loss, the generator in the proposed SVQ-WGAN framework focuses on restoring semantically meaningful and perceptually coherent images. It leverages adversarial feedback and vector-quantized latent features to ensure that the reconstructed outputs align with the underlying semantics and visual realism of the original content. In this way, the generator acts as a learned interpreter, translating discrete semantic codes back into a meaningful visual representation aligned with human and task-level perception. Let the Wasserstein distance-based critic be (D) , which provides the adversarial feedback, to guide the decoder towards the original data distribution. The transmitted image is given by

$$\hat{x} = G(\mathbf{z}_q(\mathbf{x}_i); \theta_G), \quad (5.8)$$

where G is the generator.

The generator architecture uses a series of transposed convolutional layers to progressively reconstruct the image from its quantized latent representation. The first layer applies a two-dimensional transposed convolution with 64 filters, a kernel size of three, ReLU activation, and a stride of two to perform the initial upsampling of latent features. This is followed by a residual block designed to refine and enhance the intermediate feature representations. A subsequent transposed convolutional layer with 32 filters and similar hyperparameters performs further upsampling, after which another residual block is applied to preserve structural consistency and fine details. The final transposed convolutional layer, configured with three filters and a kernel size of three, produces the reconstructed image that matches the spatial dimensions of the original input. The generator produces high-quality, semantically coherent images from discrete latent vectors while maintaining computational efficiency, making it well-suited for practical SemCom implementations.

5.2.3 Wasserstein distance and critic design

The Wasserstein distance quantifies the lowest cost required to convert one probability distribution into another. In the context of generative modeling, the Wasserstein distance can be viewed as a measure of the energy required to transform the generated data distribution $\mathbb{P}_{gen}$ into the real data distribution $\mathbb{P}_{data}$. This formulation offers a more stable and meaningful objective for optimization compared to traditional divergence measures, including the Kullback–Leibler (KL) divergences, which often lead to vanishing gradients and unstable training [230]. By grounding adversarial learning in optimal transport theory, the Wasserstein distance enables smoother gradient flow, facilitating convergence and improving control over the generative process.

Formally, the Wasserstein distance between the real and generated data distributions is defined as [18]:

$$W(\mathbb{P}_{data}, \mathbb{P}_{gen}) = \inf_{\gamma \in \Pi(\mathbb{P}_{data}, \mathbb{P}_{gen})} \mathbb{E}_{(x, \hat{x}) \sim \gamma} [\|x - \hat{x}\|], \quad (5.9)$$

where $\Pi(\mathbb{P}_{data}, \mathbb{P}_{gen})$ represents the set of all possible joint distributions $\gamma(x, \hat{x})$ with $\mathbb{P}_{data}$ and $\mathbb{P}_{gen}$ as their marginals. The expectation $\mathbb{E}_{(x, \hat{x}) \sim \gamma}[\|x - \hat{x}\|]$ denotes the average distance between real and generated samples under the optimal transport plan γ , representing the most efficient redistribution of probability mass from the generated to the real data distribution. Minimizing this expected cost encourages the generated distribution to align more closely with the real one, thereby improving perceptual and semantic fidelity in reconstructed images.

Critic network and design rationale

The critic in SVQ-WGAN architecture replaces the conventional discriminator used in standard GAN frameworks. Rather than producing a binary classification output (“real” or “fake”), the critic estimates the Wasserstein distance between the generated and real data distributions. This scalar output provides a continuous and differentiable measure of realism, leading to smoother gradient updates and more stable training dynamics. By leveraging the advantages of the Wasserstein distance, the critic can better control the distribution of generated outputs and mitigate issues including mode collapse [17]. The discrete latent vectors serve as compact representations of the input data, while the critic enforces semantic and perceptual coherence by evaluating how closely the generated distribution approximates the true data distribution. Through this mechanism, the WGAN critic ensures high-quality reconstruction, preventing the generator from converging to a limited set of outputs.

Adversarial objective function

The training process of the proposed model is governed by the adversarial loss, denoted as $\mathcal{L}_{adv}$ [127], which jointly optimizes the generator and critic. This loss quantifies the discrepancy between the real and generated data distributions, guiding both networks toward equilibrium. Let the Wasserstein distance-based critic be (D) , which provides the adversarial feedback, to guide the decoder towards the original data distribution. The formulation of the adversarial loss is given as:

$$\begin{aligned} \mathcal{L}_{\text{adv}} = \min_G \max_D \big(& \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\text{data}}} [D(\mathbf{x})] - \mathbb{E}_{\mathbf{z}_q(\mathbf{x}) \sim \mathbb{P}_{\mathbf{z}_q(\mathbf{x})}} [D(G(\mathbf{z}_q(\mathbf{x})))] \big) \\ & + \lambda \mathbb{E}_{\hat{x} \sim \mathbb{P}_{\text{interp}}} [(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2]. \end{aligned} \quad (5.10)$$

The first expectation term, $\mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\text{data}}} [D(\mathbf{x})]$, represents the average critic score for real samples drawn from $\mathbb{P}_{\text{data}}$. The second term, $\mathbb{E}_{\mathbf{z}_q(\mathbf{x}) \sim \mathbb{P}_{\mathbf{z}_q(\mathbf{x})}} [D(G(\mathbf{z}_q(\mathbf{x})))]$, corresponds to the critic's assessment of generated samples produced by the generator G , conditioned on the quantized latent vectors $\mathbf{z}_q(\mathbf{x})$. The generator aims to minimize this divergence by producing reconstructions that are indistinguishable from real data, effectively reducing the Wasserstein distance between the two distributions.

To ensure the critic satisfies the 1-Lipschitz continuity condition [231, 232] required for a valid Wasserstein metric, a gradient penalty term is introduced. The last term in equation 5.10, $\lambda \mathbb{E}_{\hat{x} \sim \mathbb{P}_{\text{interp}}} [(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2]$, penalizes deviations of the gradient norm from unity. Here, $\mathbb{P}_{\text{interp}}$ represents the distribution of points linearly interpolated between real and generated samples, and λ denotes the gradient penalty coefficient. This regularization stabilizes training by constraining the critic's gradient behavior, ensuring that the Wasserstein distance remains a valid and reliable objective function throughout optimization.

In summary, we integrate a Wasserstein distance based critic that provides a stable and theoretically grounded framework for adversarial learning. This design ensures the generator receives smooth, informative gradients, enabling it to produce semantically accurate and perceptually realistic reconstructions, even under channel-induced noise.

5.2.4 Model loss function

The total loss function for the model is expressed as a combination of several loss terms: the reconstruction loss $\mathcal{L}_{\text{rec}}$, the vector quantization loss $\mathcal{L}_{\text{vq}}$, the adversarial loss $\mathcal{L}_{\text{adv}}$, and the perceptual loss $\mathcal{L}_{\text{perceptual}}$ as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{vq}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}} + \lambda_{\text{perceptual}} \mathcal{L}_{\text{perceptual}}. \quad (5.11)$$

Randomness is introduced to the model at the training phase, which randomly drops elements from the quantized latent vector $\mathbf{z}_q(\mathbf{x}_i)$ to enable contrastive learning in the latent space. It prevents the model from becoming dependent on specific dimensions of the latent representation and enables it to learn more robust, general features, making it less sensitive to noise in the input data. The binary mask applied during index dropping $\mathbf{m}_i$, where $\mathbf{m}_i \sim \text{Bernoulli}(\mathbb{P}_{mask})$ represents the randomness of dropping elements with probability $(\mathbb{P}_{mask})$ and represented as,

$$\mathbf{z}_{q,d}(\mathbf{x}_i) = \mathbf{m}_i \odot \mathbf{z}_q(\mathbf{x}_i).$$

The generated image is given by

$$\hat{\mathbf{x}}_i = G(\mathbf{z}_{q,d}(\mathbf{x}_i); \theta_G).$$

The reconstruction loss is an expectation over possible masks:

$$\mathcal{L}_{\text{rec}}(\theta_E, \theta_G, \theta_{VQ}) = \mathbb{E}_{\mathbf{m}_i \sim \text{Bernoulli}(\mathbb{P}_{mask})} \left[\frac{1}{N} \sum_{i=1}^N \|\mathbf{x}_i - G(\mathbf{m}_i \odot VQ(E(\mathbf{x}_i; \theta_E); \theta_{VQ}); \theta_G)\|^2 \right]. \quad (5.12)$$

Perceptual loss improves the visual fidelity of reconstructed images, which are more aligned with human visual perception. Perceptual loss is calculated using the RESNET50 [233] network, comparing the feature maps of the true and predicted images as

$$\mathcal{L}_{\text{perceptual}}(\mathbf{x}_i, \hat{\mathbf{x}}_i) = \frac{1}{N} \sum_{u=1}^N \|\Phi(\mathbf{x}_i)_u - \Phi(\hat{\mathbf{x}}_i)_u\|^2, \quad (5.13)$$

where $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$ are the ground truth and generated images, respectively, $\Phi(\cdot)$ represents the feature extraction function using a specific layer of the RESNET50 network, and N is the number of elements in the feature maps. Algorithm 1 shows the training

procedure for the model using these loss terms.

Algorithm 1 Training algorithm (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

```

1: Input: Training dataset  $\mathcal{D}$ , batch size  $B$ , number of epochs  $M_e$ , latent dimension
    $L$ , number of embeddings  $K$ , number of discriminator steps  $n_D$ , gradient penalty
   weight  $\lambda$ , drop probability  $p_d$ , learning rates  $\eta_G, \eta_D$ , training variance  $\tau^2$ 
2: Output: Trained model parameters  $\theta_E, \theta_G, \theta_{VQ}$ , and  $\theta_D$ 
3: Initialize: Optimizer for discriminator  $\text{Adam}_D$  with  $\eta_D$ 
4: Initialize: Optimizer for generator  $\text{Adam}_G$  with  $\eta_G$ 
5: for  $m \leftarrow 1$  to  $M_e$  do
6:   Sample mini-batch  $\{\mathbf{x}_i\}_{i=1}^B$  from  $\mathcal{D}$ 
7:   Train Discriminator:
8:     for  $j \leftarrow 1$  to  $n_D$  do
9:       Encode  $\mathbf{x}_i$  to latent space:  $\mathbf{z}_e^i = E(\mathbf{x}_i)$ 
10:      Quantize latent vectors:  $\mathbf{z}_q^i = VQ(\mathbf{z}_e^i)$ 
11:      Add AWGN noise:  $\mathbf{z}_q^i = \mathbf{z}_q^i + \mathcal{N}(0, \tau^2)$ 
12:      Drop indices with probability  $p_d$ 
13:      Decode to generate fake data:  $\hat{\mathbf{x}}_i = G(\mathbf{z}_q^i)$ 
14:       $D_{\text{real}} = D(\mathbf{x}_i), D_{\text{fake}} = D(\hat{\mathbf{x}}_i)$ 
15:      Compute discriminator loss:

$$\mathcal{L}_{adv} = \mathbb{E}[D_{\text{fake}}] - \mathbb{E}[D_{\text{real}}] + \lambda \cdot \text{gradient penalty}$$

16:      Update  $\theta_D$  by maximizing  $\mathcal{L}_{adv}$  using  $\text{Adam}_D$ 
17:     end for
18:   Train Model:
19:     Encode  $\mathbf{x}_i$  to latent space:  $\mathbf{z}_e^i = E(\mathbf{x}_i)$ 
20:     Quantize latent vectors :  $\mathbf{z}_q^i, \mathcal{L}_{vq} = VQ(\mathbf{z}_e^i)$ 
21:     Add noise
22:     Decode to generate fake data:  $\hat{\mathbf{x}}_i = G(\mathbf{z}_q^i)$ 
23:     Compute  $\mathcal{L}_{rec}$  and  $\mathcal{L}_{perceptual}$ 
24:     Compute  $\mathcal{L}_{total}$ 
25:      $\{\theta_E, \theta_G, \theta_{VQ}\} \leftarrow \text{Adam}_G(\nabla_{\{\theta_E, \theta_G, \theta_{VQ}\}} \mathcal{L}_{total})$ 
26:   end for

```

5.2.5 Multilevel contrastive optimization

In contrast to the established contrastive mechanisms in computer vision applications, SVQ-WGAN incorporates contrastive strategies into a generative semantic communication framework, applying supervision at the pixel, latent, perceptual, and task-output levels under lossy channel simulation. We also use contrastive losses to

stabilize vector quantization, index dropout, and align task outputs (generation) under signal distortion, summarized as follows:

- **Pixel-level contrast via augmentation:** To encourage low-level semantic robustness, we apply contrastive supervision between original images and augmented views (random rotation, translation, shear, zoom, flipping, and brightness variation). These augmentations simulate spatial and photometric distortions encountered in real-world data variability. By minimizing $\|\mathbf{x} - \tilde{\mathbf{x}}\|^2$ over these augmented pairs, the model learns to preserve semantic integrity across diverse input conditions at the pixel level, where $\tilde{\mathbf{x}}$ is the augmented image.
- **Latent representation invariance:** We enforce semantic consistency by minimizing $\|z_q(\mathbf{x}) - z_q(\tilde{\mathbf{x}})\|^2$ across channel-perturbed inputs $\tilde{\mathbf{x}}$, promoting robust discrete code stability under lossy transmission. We inject AWGN into the quantized latent representations prior to index masking, simulating realistic channel distortions. During training, the model learns to align reconstructions from these degraded latents with the original input. Index masking simulates the corruption and dropout during the index transmission. This serves as an implicit contrastive mechanism between original and noisy semantic representations to minimize

$$\|z_q(\mathbf{x}) - \tilde{z}_q(\tilde{\mathbf{x}})\|^2 = \|z_q(\mathbf{x}) - \text{Mask}(z_q(\mathbf{x}) + \varepsilon)\|^2, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

, where **Mask** represents index dropout and ε reflects transmission noise.

- **Perceptual alignment:** Using deep perceptual features from a pretrained ResNet50, we define a contrastive loss in feature space: $\|\Phi(\mathbf{x}) - \Phi(\hat{\mathbf{x}})\|^2$, ensuring perceptual fidelity and task relevance beyond pixel-level matching.
- **Semantic task-level discrimination:** A Wasserstein GAN critic, trained on full reconstructions, optimizes adversarial realism under semantic distortion, in the data distribution. It enforces perceptual realism at the output level by minimizing the Wasserstein distance between the distributions of real and reconstructed images. This distance provides a more informative and task specific gradient signal than standard

divergence-based GANs, enabling stable training and encouraging reconstructions that preserve high-level semantic structure. By optimizing this distributional alignment, the model improves robustness to semantic distortion and enhances the generalizability to serve both human-perceived realism and machine-level recognition accuracy under degraded channel conditions.

As shown in table 5.1, the proposed model introduces a novel objective into contrastive learning, beyond established methodologies including SimCLR and MoCo v3, both in the objective and in the architecture.

Table 5.1: Comparison of SimCLR, MoCo v3, and SVQ-WGAN

Aspect	SimCLR [136]	MoCo v3 [137]	SVQ-WGAN
Primary Domain	Computer vision (image representation)	Computer vision (representation learning)	Semantic communication
Main Objective	Self-supervised feature learning	Self-supervised learning with memory-efficient setup	Semantic representation, compression & generation
Contrastive Learning	Contrastive loss on image pairs	Contrastive loss + momentum encoder	Multi-level contrast (latent, perceptual, pixel, task)
Data Augmentation	Two augmentations per image	Two augmentations per image	Image augmentation, spatial dropout, index masking, latent noise injection
Architecture	CNN encoder + MLP head	Vision Transformer (ViT)	CNN encoder, vector quantizer, WGAN critic, CNN decoder
Representation Type	Continuous vector embeddings	Continuous vector embeddings (ViT tokens)	Discrete latent codes
Decoder / Generator	No reconstruction	No reconstruction	Generator conditioned on discrete codebook embeddings
Evaluation	Downstream tasks (classification)	Downstream tasks (classification)	End-to-end SemCom performance (PSNR, RI, LPIPS on reconstruction) and task accuracy (classification, captioning, object detection)

5.2.6 Communication architecture

The communication pipeline is simulated using NVIDIA Sionna open-source library to evaluate the performance of the proposed system under varying channel coding rates and different SNR values. We leverage the Sionna simulator to model 5G-compliant LDPC and physical-layer effects, providing a more realistic and rigorous transmission simulation

than the simple channel models used in prior works. The channel encoder transforms the input information bits (k) into a codeword (n), while decoding is performed using a belief-propagation-based iterative algorithm with a fixed number of iterations (20). The channel is modeled for both AWGN and Rayleigh flat fading scenarios. The system configuration includes a single transmitter antenna (r_{tx}) and a single receiver antenna (r_{rx}).

A set of indices $\{j_1, j_2, \dots, j_K\}$, each of the indices representing the closest codebook vectors $\{\mathbf{c}_{j_1}, \mathbf{c}_{j_2}, \dots, \mathbf{c}_{j_K}\}$ relevant to the image, is encoded using LDPC codes as a bit stream as illustrated in figure 5.2. Each index j_k (where $k = 1, \dots, K$) is encoded as $\mathbf{u}_{c_k} = C_{\text{channel}}(j_k)$, with LDPC coding [206], where $C_{\text{channel}}(\cdot)$ denotes the channel coding function. The encoded indices $\{\mathbf{u}_{c_1}, \mathbf{u}_{c_2}, \dots, \mathbf{u}_{c_K}\}$ are then modulated using quadrature amplitude modulation (QAM), described as $\mathbf{u}_{m_k} = M(\mathbf{u}_{c_k})$ for each k , where $M(\cdot)$ is the modulation function. The set of modulated signals $\{\mathbf{u}_{m_1}, \mathbf{u}_{m_2}, \dots, \mathbf{u}_{m_K}\}$ is then transmitted over the communication channel, where it is subjected to channel noise and other impairments. At the receiver, the signal is demodulated and decoded to receive the corresponding indices $\{\hat{j}_1, \hat{j}_2, \dots, \hat{j}_K\}$. In the AWGN part of the channel, the received signal $\mathbf{y}_{\text{AWGN}} = \mathbf{z}_m + \mathbf{n}_{\text{AWGN}}$ is influenced by noise $\mathbf{n}_{\text{AWGN}}$, where each element is an independent and identically distributed Gaussian random variable with zero mean and variance σ^2 , denoted as $\mathbf{n}_{\text{AWGN}} \sim \mathcal{N}(0, \sigma^2)$.

The Rayleigh fading channel models the effects of multipath propagation, which is common in urban environments. Here, the received signal is given by $\mathbf{y}_{\text{Rayleigh}} = \mathbf{h}\mathbf{z}_m + \mathbf{n}_{\text{Rayleigh}}$, where $\mathbf{h}$ is the channel fading coefficient modeled as a complex Gaussian random variable with zero mean and unit variance ($\mathbf{h} \sim \mathcal{CN}(0, 1)$), and $\mathbf{n}_{\text{Rayleigh}}$ is a noise vector denoted as $\mathbf{n}_{\text{Rayleigh}} \sim \mathcal{N}(0, \sigma^2)$. The Rayleigh fading channel affects the amplitude of the transmitted signal, leading to variations in the received signal power. When combined with AWGN, this presents a challenging scenario for the communication system. A linear minimum mean square error (LMMSE) equalizer is used at the receiver to compensate for the effects, including fading and noise. The equalizer processes the received signal vector, $\mathbf{y}_{\text{Rayleigh}}$, the channel vector $\mathbf{h}$, and the noise variance to estimate the transmitted symbols. The effective noise variance is

computed for the subsequent demapping stage. The signal-to-noise ratio per symbol (E_s/N_0) is converted to noise variance (n_o) considering the modulation order ($\log_2 M$) and coding rate (R_c).

5.2.7 Performance metrics

SemCom based application requires novel matrices to assess system performance, because traditional metrics, including PSNR, rely on pixel-wise differences. Perpetual quality assessment is essential in modern intelligent applications such as healthcare, ITS, and XR/virtual reality (VR). MS-SSIM captures structural information and contrast sensitivity, thereby providing a more accurate representation of human-perceived quality. LPIPS [234] is more suitable to assess SemCom systems, since it leverages deep learning to evaluate perceptual similarity based on feature representations. The evaluation presented in this chapter, based on PSNR, provides an essential measure of signal fidelity, while MS-SSIM and LPIPS offer assessments that better reflect human perception. These matrices do not quantify the methodology's robustness or semantic integrity. Therefore, in this section, we present two novel metrics to evaluate semantic integrity.

Robustness index (RI)

In SemCom systems, distortions may arise from both channel impairments and imperfections in semantic representation, potentially altering the interpreted meaning even when perceptual similarity remains high. Therefore, a metric is required to jointly evaluate semantic fidelity and the system's resilience to transmission-induced distortions. Therefore, RI quantifies the ability of the proposed framework to preserve semantic information while remaining robust to channel-induced errors.

RI integrates semantic fidelity (SF) and error probability (EP), to evaluate the methodology's robustness. The similarity of the semantic content is quantified by the SF, and it is measured using the LPIPS metric. EP represents the likelihood that the reconstructed semantic information deviates significantly from the original. This probability is calculated as

$$\text{EP} = Q\left(\frac{\|\mathbf{x} - \hat{\mathbf{x}}\|}{\sigma}\right), \quad (5.14)$$

where $Q(\cdot)$ is the Q-function [235], σ is the standard deviation of the noise, and $\|\mathbf{x} - \hat{\mathbf{x}}\|$ is the Euclidean distance between the real and transmitted vectors. The RI combines SF and EP into a single metric as

$$\text{RI} = \alpha_r \cdot \text{SF} + (1 - \alpha_r) \cdot (1 - \text{EP}), \quad (5.15)$$

where α_r is a weighting factor that balances the importance of semantic fidelity and error probability.

Semantic distortion measurement

The metric is based on intersection over union (IoU) error and category error. Category error quantifies the extent to which the object categories in the original image are preserved in the generated image. This evaluation is particularly important because maintaining semantic integrity is crucial for practical applications, including robotics, autonomous vehicles, industrial automation, and augmented reality (AR), where object locations and boundaries play a key role in downstream tasks. Category error is formally defined as follows:

$$E_{\mathcal{K}} = 1 - \frac{|\mathcal{K}_{\mathbf{x}} \cap \mathcal{K}_{\hat{\mathbf{x}}}|}{|\mathcal{K}_{\mathbf{x}}|}, \quad (5.16)$$

where $\mathcal{K}_{\mathbf{x}}$ is the set of object categories detected in the ground truth image and $\mathcal{K}_{\hat{\mathbf{x}}}$ is the set of object categories detected in the generated image. The IoU [236] measures the overlap between the bounding boxes of detected objects in the original and generated images. Let $B_{\hat{\mathbf{x}}}$ represent the set of bounding boxes in the generated image $\hat{\mathbf{x}}$. For each pair of bounding boxes $b_{\mathbf{x}} \in B_{\mathbf{x}}$ and $b_{\hat{\mathbf{x}}} \in B_{\hat{\mathbf{x}}}$, the IoU is given by

$$\text{IoU}(b_{\mathbf{x}}, b_{\hat{\mathbf{x}}}) = \frac{\text{Area of Intersection}(b_{\mathbf{x}}, b_{\hat{\mathbf{x}}})}{\text{Area of Union}(b_{\mathbf{x}}, b_{\hat{\mathbf{x}}})}, \quad (5.17)$$

Chapter 5. Contrastive vector-quantized generative semantic representation for robust semantic communication

and the IoU error is given by

$$E_{\text{IoU}} = 1 - \text{mean}(\text{IoU}(B_{\mathbf{x}}, B_{\hat{\mathbf{x}}})) , \quad (5.18)$$

. Then the semantic distortion is given by

$$D_{\text{semantic}} = w_{\mathcal{K}} \cdot E_{\mathcal{K}} + w_{\text{IoU}} \cdot E_{\text{IoU}} , \quad (5.19)$$

where $w_{\mathcal{K}}$ and w_{IoU} are the weights assigned to each error term. For our experiment, we used 0.8 and 0.2, respectively, giving more importance to the category; this value can be tuned based on the application. If $|\mathcal{K}_{\mathbf{x}}| = 0$, set $E_{\mathcal{K}} = 0$ and $E_{\text{IoU}} = 0$, as there are no objects detected in the original image. However, if $|\mathcal{K}_{\mathbf{x}}| \neq 0$ and $|\mathcal{K}_{\hat{\mathbf{x}}}| = 0$, indicating that the original image contains objects but none are detected in the generated image, set both $E_{\mathcal{K}} = 1$ and $E_{\text{IoU}} = 1$, representing a complete failure in both category detection and object alignment.

5.3 Theoretical analysis

This section validates the design decisions underlying the proposed SVQ-WGAN framework, presenting theoretical grounds for the selection of main components, including quantized latent representation, index-based transmission, adversarial learning, and contrastive regularization, which work together to achieve efficient semantic representation. The validations demonstrate how the main objectives of this thesis are achieved via the developed model by achieving compression, maintaining semantic fidelity, enhancing robustness against channel noise, and maintaining generalization against different distributions.

Proposition 5.3.1. Reconstruction loss preserves task utility under semantic equivalence

Let $f : \mathbb{R}^{H \times W \times 3} \rightarrow \mathcal{Y}$ be a pretrained downstream task model (e.g., object detector), and define $f(\mathbf{x}) = (c_{\mathbf{x}}, b_{\mathbf{x}})$, where $c_{\mathbf{x}}$ is the predicted class and $b_{\mathbf{x}}$ is the predicted bounding box and f is Lipschitz continuous over a semantic manifold $\mathcal{M}_{\text{sem}} \subset \mathbb{R}^{H \times W \times 3}$,

Chapter 5. Contrastive vector-quantized generative semantic representation for robust semantic communication

$\hat{\mathbf{x}} \in \mathcal{M}_{\text{sem}}$, and if the reconstruction error satisfies $\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 \leq \varepsilon$ for some small $\varepsilon > 0$. Then the task output is preserved with a margin.

proof of proposition 5.3.1 is available in appendix A.1.1.

Proposition 5.3.2. Learned prior over quantized latents enables structured semantic decoding

Let $G : \mathbb{R}^F \rightarrow \mathbb{R}^{H \times W \times C}$ denote the generator in the VQ-based semantic communication model. G is trained to be Lipschitz continuous as per the design in section 5.2.3, i.e., there exists a constant $L > 0$ such that for all $\mathbf{z}_q^{(1)}, \mathbf{z}_q^{(2)} \in \mathbb{R}^F$,

$$\|G(\mathbf{z}_q^{(1)}) - G(\mathbf{z}_q^{(2)})\|_2 \leq L \|\mathbf{z}_q^{(1)} - \mathbf{z}_q^{(2)}\|_2.$$

proof of hypothesis 5.3.2 is available in appendix A.1.1

Proposition 5.3.3. Adversarial training with induced noise enables semantic robustness Let $\tilde{\mathbf{z}}_q = \mathbf{z}_q + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ representing discrete channel noise. Assume the discriminator D is constrained to be 1-Lipschitz. If G is trained to minimize the adversarial objective

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_x[D(\mathbf{x})] - \mathbb{E}_{\tilde{\mathbf{z}}_q}[D(G(\tilde{\mathbf{z}}_q))],$$

then $\mathbb{P}_{\text{gen}}^\varepsilon$ converges toward $\mathbb{P}_{\text{data}}$ in Wasserstein-1 distance. Consequently, samples $\hat{\mathbf{x}} = G(\tilde{\mathbf{z}}_q)$ are semantically valid with high probability.

proof of proposition 5.3.3 is available in appendix A.1.1

Proposition 5.3.4. Index dropout combined with adversarial training enables error-tolerant reconstruction

Let $s(\mathbf{x}) \in \{1, \dots, K\}^N$ denote the N quantized code indices for an input $\mathbf{x}$, and let $E : \{1, \dots, K\}^N \rightarrow \mathbb{R}^{N \times d}$ be the codebook lookup that maps indices to embeddings. Define the clean quantized embedding sequence $\mathbf{z}_q(\mathbf{x}) \triangleq E(s(\mathbf{x}))$ and a default embedding $\mathbf{z}_0 \in \mathbb{R}^d$. Let $m \in \{0, 1\}^N$ be a binary mask sampled from Bernoulli($1 - \mathbb{P}_{\text{mask}}$). Let

the corrupted latent be defined as:

$$\tilde{\mathbf{z}}_q = m \odot \mathbf{z}_q + (1 - m) \odot \mathbf{z}_0,$$

where $\mathbf{z}_0$ is a default symbol (zero vector). Let the reconstruction be $\hat{\mathbf{x}} = G(\tilde{\mathbf{z}}_q)$. If G is trained with:

$$\mathcal{L}_{\text{rec}} = \|\mathbf{x} - \hat{\mathbf{x}}\|^2, \quad \mathcal{L}_{\text{adv}} = \mathbb{E}_{\mathbf{x}}[D(\mathbf{x})] - \mathbb{E}_{\tilde{\mathbf{z}}_q}[D(G(\tilde{\mathbf{z}}_q))],$$

then for small $\mathbb{P}_{\text{mask}}$,

$$\mathbb{E}[\|G(\mathbf{z}_q) - G(\tilde{\mathbf{z}}_q)\|^2] \leq \delta(\mathbb{P}_{\text{mask}}),$$

where $\delta(\mathbb{P}_{\text{mask}})$ is the worst-case expected output perturbation caused by masking latent indices

$$\text{with } \delta \rightarrow 0 \text{ as } \mathbb{P}_{\text{mask}} \rightarrow 0,$$

.

proof of proposition 5.3.4 is available in appendix A.1.1

Proposition 5.3.5. Symbolic semantic compression bounds bandwidth usage

Let $\mathbf{z}_q(x) \in \{1, \dots, K\}^N$ be the quantized latent representation of an input x , where K is the size of the codebook and N is the number of symbols. Assume a discrete communication channel transmitting the index sequence $\mathbf{z}_q(x)$ losslessly. Then, the theoretical bandwidth B (in bits) required to transmit $\mathbf{z}_q(x)$ satisfies:

$$B = N \cdot \log_2 K,$$

and is independent of the original input size ($\dim(x)$). Furthermore, for a fixed maximum bandwidth B_{max} , the compression rate satisfies:

$$\frac{B}{\dim(x)} \ll 1,$$

provided $N \ll \dim(x)$, thus demonstrating that the learned codebook representation

achieves compression proportional to symbol count and logarithmic in vocabulary size.

proof of proposition 5.3.5 is available in appendix A.1.1

5.4 Simulation results

This section presents the implementation details, including the dataset, hyperparameter selection, and training strategy, followed by a comprehensive evaluation of the proposed system for its rate-distortion performance, noise robustness, task performance, bandwidth efficiency, and semantic distortion. The evaluation also includes the complexity and latency evaluation of the proposed system.

Training dataset

We used a diverse set of high-quality images for training, combining DIV2K [237], a high-resolution image dataset; CelebA-HQ [238], a high-resolution facial image dataset; and the entire Flickr15k dataset [239]. The preprocessing pipeline resized all images to a uniform dimension of (256, 256, 3) and normalized pixel values for consistency. To introduce pixel-level contrast, we augmented the images, and a total of 8,000 augmented samples were generated. We used transformations including random rotations (up to 20 degrees), horizontal flips, brightness variations (0.8–1.2), shifts (up to 20% of width/height), and zooms (up to 20%). The total training set was 27620, including the augmented images. The motivation behind the augmentation was to increase diversity and to generalize the model’s performance across varied scenarios. The train-test split was a 90-10 ratio.

Implementation details

To optimize the training process, we used warm-up cosine learning rate scheduling. Its initial warm-up phase is followed by a cosine decay, which provides smoother gradient updates and improved convergence. Learning rate increases linearly from a small initial value to the base learning rate over a fixed number of steps (s_{warm}) during the warm-up phase, which constitutes 15% of the total training steps, aiming to address the instability

of the gradient updates by stabilizing and ensuring effective convergence. After the warm-up phase, the learning rate follows a cosine decay to reach zero (S_{total}). The decay is designed to smoothly taper off the learning rate over the remaining steps, enhancing model performance and training efficiency. The total number of steps is calculated as

$$S_{\text{total}} = \frac{\mathcal{D}}{\text{Batch Size}} \times \text{Epochs}.$$

Batch Size is the size of each mini-batch, and Epochs is the total number of epochs. The base learning rates for the discriminator and generator were selected as $\eta_D = 0.001$ and $\eta_G = 0.004$, respectively. The Generator and Discriminator learning rate schedules are depicted in figure 5.4. We have used $\alpha = 0.25$, which weights the commitment loss in $\mathcal{L}_{\text{vq}}$.

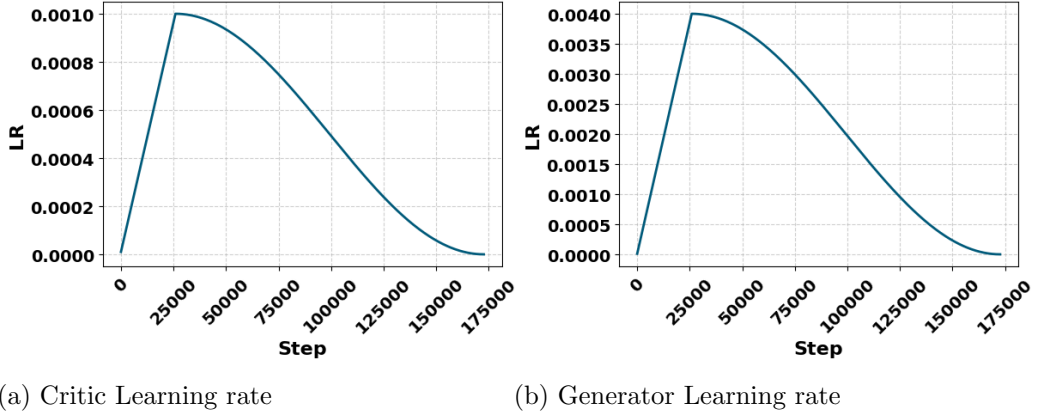


Figure 5.4: Learning rate schedule (a) critic learning rate schedule (b) generator learning rate schedule. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

To enforce contrastive objectives during the training process, we introduced AWGN noise to the quantized vector at an SNR of 10, with the noise distribution modeled as $\mathcal{N}(0, 0.1)$. The probability of masking, $\mathbb{P}_{\text{mask}}$, was set to 0.01, to simulate the potential loss of latent symbols during transmission. This allows the model to learn robust reconstructions even when part of the semantic information is missing. The training variance, denoted as τ , was computed as 0.0615, representing the spread or dispersion of values within the augmented training dataset, $\mathcal{D}$.



Figure 5.5: Spatial attention heat maps for sample images from the Kodak dataset illustrating edge-focused attention distribution. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

The semantic extraction at the encoder is done using convolutional layers with 4×4 kernels and 2×2 strides, combined with ReLU activations, to downsample the input images while preserving features. The residual connections in the encoder maintain feature continuity, improve gradient flow, and enhance stability during the training process. To standardize feature scaling, we applied batch normalization with a dropout probability of 0.05. Dropout is used for regularization and to prevent overfitting. The novel spatial attention described in section 5.2.1 emphasizes edge regions in the feature maps. Figure 5.5 shows the spatial attention heatmap applied to sample images from the Kodak dataset. Although the Sobel-based attention emphasizes edges, it may assign relatively less focus to smooth regions. However, this is mitigated through residual learning, convolutional integration, and multilevel contrastive learning, which reinforces semantic consistency across scales. As shown in figures 5.14 and 5.15, no visible artifacts are observed in smooth areas. Compared to self-attention ($\mathcal{O}(N^2)$) and multi-scale fusion networks ($\mathcal{O}(N \log N)$), the proposed method operates with near-linear complexity ($\mathcal{O}(N)$) and low memory overhead, making it efficient for real-time semantic communication.

The semantic encoding, according to the layered architecture defined in figure 4.1, is performed by the quantizer by mapping the continuous feature representations from

the encoder into discrete codebook vectors. The continuous feature vector is flattened, and one-hot encoding is applied to select the nearest embedding vectors in a learned codebook.

The lightweight generator consists of transposed convolutional layers with 4×4 kernels, strides of 2×2 , and ReLU activations to progressively upscale the latent features. The residual blocks, with an attention module, enhance reconstruction quality by preserving essential details during upsampling.

The discriminator is designed with convolutional layers with LeakyReLU activations, which enhance gradient flow during adversarial learning, and applies dropout regularization with a probability of 0.2 to prevent overfitting. The final fully connected layer produces a scalar output representing the likelihood that the input image is real or fake. We set the gradient penalty coefficient $\lambda = 10$ and the number of discriminator steps, n_D , to 5 based on empirical best practices from the literature [19]. The λ balances the contributions of adversarial loss and regularization during training. Our experiments with values $\lambda \in \{1, 5, 10, 20\}$ confirmed that $\lambda = 10$ offers the best stability-performance trade-off, avoiding both critic overfitting (low λ) and under-training (high λ).

The proposed model is validated using the traditional image codec BPG and SemCom based baseline JSCC. BPG [240] is selected as the baseline traditional codec because it delivers higher compression efficiency and perceptual quality than the older standards JPEG and JPEG2000 coding. It is a lightweight codec and widely used in SemCom literature as a baseline. The results for the Kodak dataset are replicated for the JSCC implemented in [71]. We used the pre-trained weights of the model trained at an SNR of 1 with a CPP ratio of 1/16, exhibiting a limitation of this methodology, which the model is trained for a specific channel condition, including SNR or CPP, in contrast our methodology performed well in diverse channel conditions as validated in the following section. The thesis validates the SemCom model in an AWGN and Rayleigh fading channel with an LDPC coding scheme and QPSK modulation. The model was experimented with varying SNR levels to benchmark the system with BPG encoding and JSCC. Further classification accuracy and image captioning accuracy at varying SNR levels are simulated.

5.4.1 Rate distortion analysis

This section presents an experimental validation of the proposed model for its perpetual and semantic fidelity and noise robustness. The model was configured to operate at a codebook size of 256 and BPG was configured to operate at its maximum compression level, ensuring alignment with the compression rate of the SemCom system. We set $\alpha_r = 0.25$ to balance SF and EP, placing greater emphasis on communication reliability in our novel metric RI. This value was empirically determined through preliminary experiments to achieve consistency with perceptual degradation trends, especially at low SNR. Results are presented for pixel-based fidelity by PSNR, MS-SSIM, semantic fidelity by LPIPS, and noise robustness by RI.

Experimental results on quality of transmission

The results presented in figures 5.6a and 5.7a demonstrate that SVQ-WGAN consistently achieves higher PSNR values across both AWGN and Rayleigh fading channels, particularly in low SNR conditions. It performs better at CRs 0.67 and 0.50 in contrast to BPG, and the system performs well in a fading channel, validating its performance in challenging channel conditions. The proposed system performs comparably well against the JSCC, where JSCC performs well in low SNR regions below 2 dB in AWGN channel, but at a higher transmission cost in terms of bit rate. The proposed model performs well in terms of structural similarity compared to traditional image coding, maintaining a significant difference in PSNR values, as shown in figure 5.6b, particularly at lower SNRs below 5 dB. SVQ-WGAN has comparably good MS-SSIM against the JSCC where JSCC performs well in low SNR regions below 5 dB owing to its latent transmission. The Rayleigh fading channel figure 5.7b presents a more challenging scenario due to its multipath fading characteristics. At an SNR of 4 dB and CR 0.5 and 0.67, SVQ-WGAN exhibits a much higher MS-SSIM, while BPG fails. Even at higher SNRs, BPG's performance is severely degraded under fading conditions, struggling to maintain any significant structural integrity. SVQ-WGAN, on the other hand, shows stability in preserving structural details across all coding rates, even under these harsher channel conditions.

Experimental results on semantic fidelity

This section validates the semantic preservation under noisy conditions. In the AWGN channel figure 5.6c, SVQ-WGAN consistently outperforms BPG in terms of LPIPS across all SNR levels and coding rates. Notably, at low SNRs of below 5 dB, it shows semantic preservation with low LPIPS, particularly at coding rates of CR 0.50 and CR 0.67. As the SNR increases, SVQ-WGAN's LPIPS values approach 0.3, compared to BPG's 0.55, suggesting that BPG's perceptual quality deteriorates quickly in noisy conditions due to the pixel-based reconstruction. At higher SNRs, BPG struggles to match the perceptual quality provided by SVQ-WGAN, particularly at higher coding rates. JSCC maintains better overall perceptual quality across all SNR levels in the AWGN channel, at the cost of a higher bit rate than the proposed scheme. The gap between SVQ-WGAN and BPG widens in a more challenging Rayleigh fading channel, as shown in figure 5.7c. Rayleigh fading introduces complex noise and fading effects due to multipath propagation, making it significantly more challenging to maintain semantic fidelity when pixel-based reconstruction is employed. At low SNRs (around 3 dB), SVQ-WGAN exhibits considerably lower LPIPS values compared to BPG, especially at coding rates of 0.67 and 0.5 with AWGN channel, where BPG performs poorly, with LPIPS values at settling at around 0.55 even at higher SNR rates above 7 dB. The above experimental validations underscore the proposed methodologies' advantage over traditional image coding and JSCC for maintaining quality of experience (QoE) in real-world communication systems.

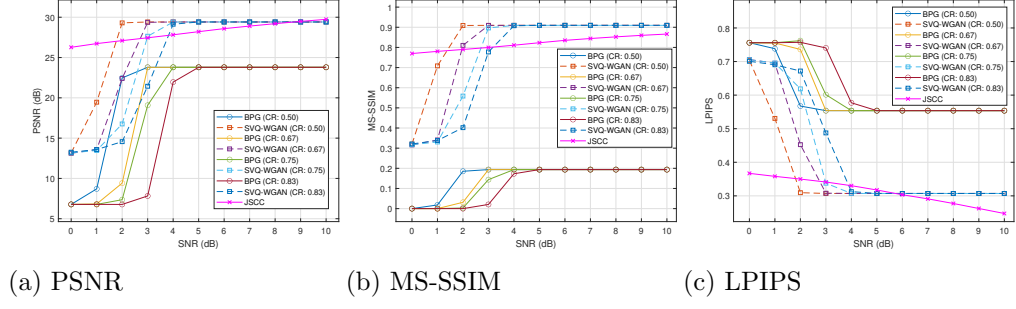


Figure 5.6: Image transmission performance on the Kodak dataset over an AWGN channel across different SNR levels using QPSK modulation and LDPC coding with multiple coding rates. Results are presented for a codebook size of 256: (a) average PSNR (higher is better), (b) average MS-SSIM (higher is better), and (c) average LPIPS (lower is better). (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

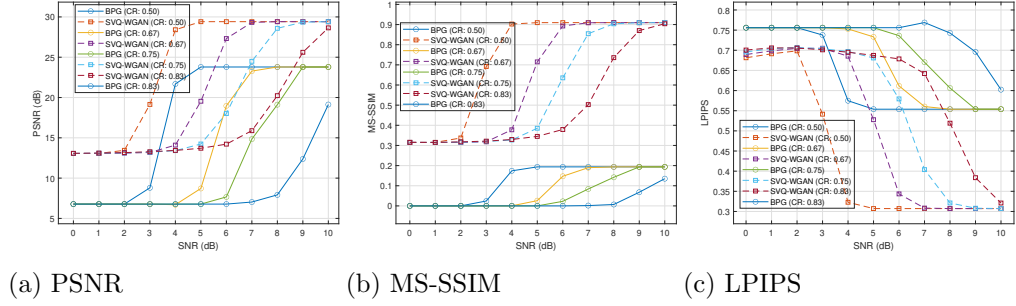


Figure 5.7: Image transmission performance over fading channel for varying SNRs for Kodak dataset with different coding rates for LDPC for QPSK: (a) average PSNR (higher the better), (b) average MS-SSIM (higher the better), and (c) average LPIPS (lower the better). (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

Experimental results on noise robustness

Figure 5.8 shows the experimental results with the novel metric RI with AWGN and Rayleigh fading channel across varying SNRs. In both AWGN and the Rayleigh fading channel, SVQ-WGAN consistently achieves a higher RI across all SNR levels and coding rates. With AWGN channel SVQ-WGAN shows a stable RI even in low SNRs, with values significantly higher than BPG, starting at SNRs as low as 2 dB, as shown in figure 5.8a. The RI behaviour over the Rayleigh fading channel, as illustrated in figure

5.8b, shows that the disparity between SVQ-WGAN and BPG becomes even more pronounced. Validated by these experimental results, we claim that the SVQ-WGAN can maintain a high robustness to channel noise and its suitability for communication systems deployed in environments with highly noisy conditions, making SVQ-WGAN a robust choice in scenarios where noise resilience is critical for maintaining the quality.

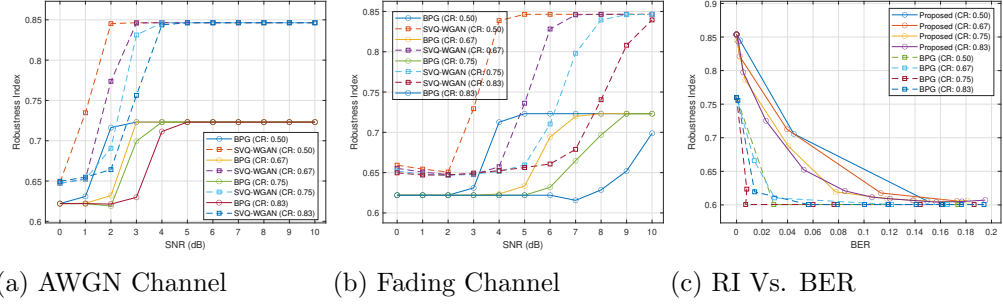


Figure 5.8: RI evaluation with different channel conditions (a) AWGN channel, (b) fading channel, (c) BER. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

Experimental results on performance versus BER

The semantic symbols are affected by the physical channel noise during transmission, and it is critical that we evaluate their effect on the performance. This section evaluates the semantic transmission efficiency with varying BERs in a Rayleigh fading channel, using LDPC coding. The pixel-level fidelity demonstrated by PSNR, illustrated in figure 5.9a, validates that the system outperforms the BPG scheme across all CRs. The proposed exhibits a gradual decline at moderate and high BERs compared to a sharp decline in BPG. The generative capabilities combined with the channel error correction enable better withstanding the effects of channel noise. The MS-SSIM and shown in figure 5.9b further validates the claim. The proposed system maintains a significant gap in MS-SSIM, while it widens as the BER increases. For instance, at higher coding rates (0.75 and 0.83), the proposed system maintains notably higher MS-SSIM values, particularly at BER values above 0.04, demonstrating the system's capacity to retain structural integrity even in highly noisy environments. The semantic fidelity presented in figure 5.9c, measured using LPIPS, further validates the claim that the SVQ-WGAN

preserves semantic information due to its generative capabilities, which stem from its contrastive training design. It maintains a lower LPIPS across all coding rates. figure 5.8c depicts the RI variation with different BERs, where our design maintains higher values than traditional image coding. At CR of 0.5 and BER of 0.05 SVQ-WGANs RI is around 0.7 whereas BPG scheme's RI drops below 0.6. This validates that combining a deep learning-based generative model with LDPC error correction enables the proposed scheme to achieve strong resilience to noise.

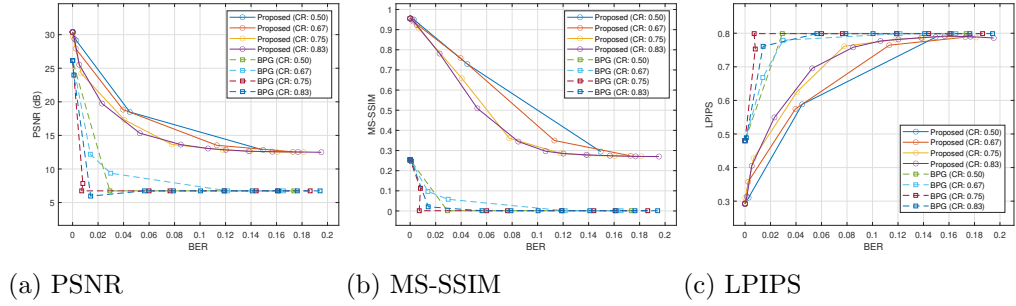


Figure 5.9: Quality versus BER for Kodak dataset over Rayleigh fading channel with LDPC coding for QPSK : (a) average PSNR (higher the better), (b) average MS-SSIM (higher the better), and (c) average LPIPS (lower the better). (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

5.4.2 Task performance

The SemCom systems are expected to be designed for M2M or M2H systems. The transmitted images can be directly processed by human or machine agents. This requirement has motivated us to evaluate the performance of our system in terms of two broad classes of semantic utility: human perception-oriented utility and machine perception-oriented utility. The former emphasizes the preservation of meaning as understood by human observers; the latter focuses on the semantic features necessary for automated reasoning and decision-making tasks. To achieve this objective, we evaluate the system against the baselines using image captioning accuracy as a proxy for human semantic understanding and classification accuracy as a representative of machine-level perception. The following section exhibits the results of the analysis.

Image captioning accuracy

Image captioning experiment is conducted on transmitted images for SVQ-WGAN and BPG, to assess the human perceived semantic preservation. The bootstrapping language-image pre-training (BLIP) image-to-text model is used to generate textual descriptions from images reconstructed after transmission through a fading channel. These generated captions were then evaluated using BERTScore F1 [241] against human-annotated ground-truth captions from the COCO dataset. BERTScore F1 captures deep semantic similarity at the language level. The results are illustrated in table 5.2.

Table 5.2: BERTScore F1 performance across varying SNR levels for BPG and the proposed model. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

SNR (dB)	BPG	SVQ-WGAN
1	0.8555	0.8368
2	0.8555	0.8278
3	0.8590	0.8644
4	0.8818	0.8826
5	0.8817	0.9088
6	0.8817	0.9118
7	0.8817	0.9118
8	0.8817	0.9118
9	0.8817	0.9118
10	0.8817	0.9118

The table 5.2 demonstrates that the proposed system gains higher BERTScore F1 values than BPG for all SNR levels above 3 dB, with higher gains in the moderate-noise regime beyond 5 dB. These improvements result from the system’s quantized latent representations, which are refined through adversarial training to preserve semantically meaningful features. In contrast, BPG prioritizes pixel-level compression without regard to semantics preservation, making it more susceptible to semantic degradation under both compression and channel impairments. Overall, SVQ-WGAN proves more effective at preserving the semantic meaning of images in language-driven evaluation tasks, as measured by human-perceived semantic fidelity.

Classification accuracy

To assess the proposed system’s utility for machine vision tasks (M2M), we evaluated classification accuracy on transmitted images using a ResNet-50 model pretrained on ImageNet and fine-tuned on the CIFAR-10 dataset. The network was first trained with its backbone frozen, then optimized end-to-end, reaching an accuracy of 96.2% on clean test samples. All images were resized to 224×224 and augmented during training to improve robustness. This evaluation serves as a proxy for machine perception. The classification task tests the receiver’s ability to reliably interpret and categorize visual information as transmitted by the indices of the quantized latents of the images as minimal semantic symbols, despite distortions introduced by the communication channel, which degrades the semantic symbol fidelity. As illustrated in figure 5.10, images transmitted through SVQ-WGAN achieved 94% accuracy at SNR levels above 4 dB, surpassing both BPG and JSCC baselines.

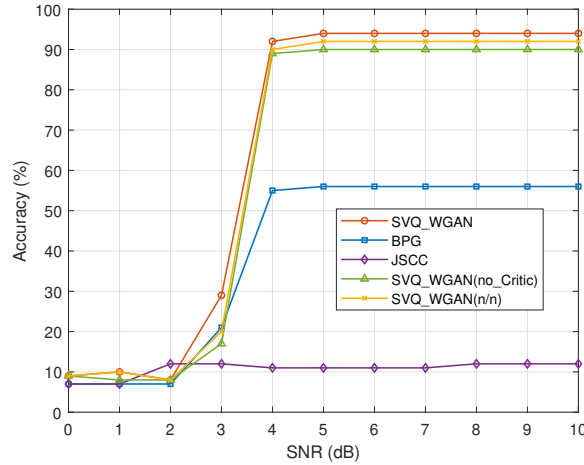


Figure 5.10: Classification accuracy on the CIFAR-10 dataset across different SNR levels. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

The inclusion of the critic in the model training process plays a key role in preserving discriminative semantic features, even under extreme noisy conditions, as reflected by the moderate drop in classification accuracy to 90% when the critic is removed from the SVQ-WGAN model. The traditional codec BPG reaches only about 56% accuracy, underscoring its limited ability to preserve semantic class-relevant visual

cues under compression and channel impairments due to its focus on pixel-based reconstruction. JSCC remains a less effective deep learning based baseline, yielding the lowest accuracy across all SNR values. This behaviour stems from its dependence on the channel characteristics used during training, which restricts its ability to generalize to mismatched noise conditions. As a result, the encoder–decoder architecture fails to maintain the discriminative features needed for reliable classification, leading to a sharp reduction in accuracy. Moreover, because JSCC prioritizes pixel fidelity over semantic preservation, its autoencoder-based design struggles to sustain task-level utility under varying channel environments, validating the importance of representation learning in SemCom.

These findings support the design of SVQ-WGAN with a strong representation learning framework that produces resilient and broadly generalizable semantic representations, supporting both human-level perception and machine-level understanding even in the presence of transmission-induced distortions.

5.4.3 Bandwidth efficiency

To assess the bandwidth efficiency of transmitting the indices of the quantized vectors, we normalize all methods by calculating the number of transmitted complex symbols per image pixel. Under a code rate of 0.5 with QPSK modulation, JSCC requires 1.0 symbols per pixel. In contrast, BPG with the same LDPC coding rate reduces this requirement to 0.562, while the proposed SVQ-WGAN further lowers it to just 0.438. This corresponds to a 56.2% decrease in transmission overhead relative to the deep learning based baseline JSCC.

The effective bitrate after LDPC redundancy is computed as $\text{bpp}_{\text{eff}} = \frac{\text{Transmitted bits}}{R_c \cdot h \cdot w}$. The bandwidth efficiency measures the number of physical-layer symbols required to transmit the minimal semantic representation. This analysis accounts for the redundancy introduced during channel encoding. For SVQ-WGAN, a 256×256 image generates 4096 quantized indices (8 bits each for $K = 256$), yielding 32,768 bits per image. At its maximum compression BPG’s average output is 5 KB (40,000 bits). Table 5.3 presents the effective bpp across codebook sizes and coding rates. At 5 dB SVQ-WGAN

consistently maintains ~ 30 dB PSNR at lower bitrates (e.g., $R_c = 0.67$), while BPG ranges between 27.0–29.5 dB. In conjunction with the transmission quality analysis, the above results demonstrate a better rate-distortion trade-off for SVQ-WGAN under identical channel conditions, achieving higher perceptual quality at reduced bandwidth.

Table 5.3: Effective bitrate comparison between the proposed compact index-based transmission paradigm and BPG with LDPC coding. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

R_c	SVQ-WGAN (128)	SVQ-WGAN (256)	SVQ-WGAN (512)	BPG
0.50	0.875	1.000	1.125	1.221
0.67	0.653	0.746	0.839	0.911
0.75	0.583	0.667	0.750	0.814
0.83	0.527	0.602	0.677	0.735

5.4.4 Model complexity

The computational complexity of the framework is mainly driven by convolutional operations and quantization. The computational complexity of the model is $\mathcal{O}(hwc^2k_l^2N_l + hwLK)$, where N_l denotes the number of layers and k_l the convolutional kernel size. The quantization step contributes only a modest overhead because it operates on a spatially reduced latent representation. In JSCC, the complexity increases due to its iterative decoding with T_{jsc} . BPG achieves the lowest computational complexity and floating-point operations per second (FLOP)s, particularly during decoding, due to its non-learned signal processing components, specifically context-adaptive binary arithmetic coding (CABAC). It delivers degraded structural quality and semantic fidelity compared with SVQ-WGAN and JSCC, especially in noisy or low-bitrate scenarios. As summarized in table 5.4, SVQ-WGAN achieves lower FLOPs than JSCC while still preserving high-fidelity reconstructions.

Table 5.4: Encoder/Decoder FLOPs ($\times 10^9$) and asymptotic complexity at 256×256 resolution. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

Metric	SVQ-WGAN	JSCC	BPG
Enc. FLOPs	~ 9.0	~ 14.5	~ 0.10
Dec. FLOPs	~ 24.0	~ 36.4	~ 0.05
Enc. Complexity	$\mathcal{O}(hwc^2k_l^2N_l + hwLK)$	$\mathcal{O}(hwc^2k_l^2N_l)$	$\mathcal{O}(hw \log(hw))$
Dec. Complexity	$\mathcal{O}(hwc^2k_l^2N_l + hwLK)$	$\mathcal{O}(T_{j_{sc}} \cdot hwc^2k_l^2N_l)$	$\mathcal{O}(hw)$

To provide a comparison of model compactness and storage requirements to validate the models’ deployability in resource-constrained environments, we report the parameter sizes of all evaluated systems. SVQ-WGAN comprises approximately 1.77 million parameters with a model size of about 6.7 MB as summarized in table 5.5. In our model, the discriminator accounts for most of the parameters, and since it is used only during training, it does not contribute to inference-time cost. In contrast to JSCC, the system contains roughly 3.73 million parameters and occupies 14.2 MB, with feature refinement performed via convolution in the decoder. Our design choice results in reduced inference cost, $2\times$ lower than deep JSCC. Results in improved efficiency, deployability, and suitability for bandwidth-constrained scenarios. For the BPG-LDPC baseline, the reported model size corresponds to codec binaries, which are typically < 1 MB, the lowest among the baselines, but with less perceptual quality.

Table 5.5: Comparison of SVQ-WGAN and deep JSCC in terms of parameters. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

Component	SVQ-WGAN	Deep JSCC
Encoder	$\sim 36K$	$\sim 1.56M$
Decoder	$\sim 40K$	$\sim 1.01M$
Vector quantizer	$\sim 4K$	None
Denoiser	None	$\sim 1.16M$
Discriminator	$\sim 1.68M$	None
Total parameters	$\sim 1.77M$	$\sim 3.73M$
Model parameter size	~ 6.7 MB	~ 14.2 MB

Inference latency was approximated based on decoder parameter counts and typical FLOP-to-time ratios on an NVIDIA T4 GPU, assuming a realistic throughput of 2.4 TFLOPs under 30% utilization for single-image inference. The SVQ-WGAN decoder

achieves significantly lower inference latency of 1.5 ms compared to the JSCC decoder 20 ms, due to its smaller model size and single-pass architecture. BPG processes the entire image using DCT, quantization, and entropy coding, achieving the lowest latency of 0.2 ms.

This analysis provided clear insights into the proposed models’ practical deployability, in resource-constrained devices and IoT platforms. The proposed model provides a trade-off between low computational overhead and substantial perceptual and semantic similarity in the transmitted images.

5.4.5 Ablation study

The design of the codebook is an important aspect in defining the language of the SemCom system. Larger codebooks provide more semantic symbols to refer to, and hence greater variability and better quality, but they increase transmission overhead. An ablation study is conducted across varying codebook sizes ($K = 128, 256, 512$) to assess the trade-off between semantic richness and transmission overhead, as shown in figure 5.11. The robustness measurement RI remained consistent across sizes, indicating that the codebook affects only semantic richness. For our experiments throughout this chapter, we selected $K = 256$ as it provides a good balance between semantic fidelity and transmission efficiency, with additional improvements observed when noise regularization was applied during training. The semantic fidelity of the model trained without adding noise to the latent vector was not optimal in terms of pixel fidelity, as indicated by the PSNR and MS-SSIM results. This confirms that injecting noise into the latent space improves pixel-level fidelity during reconstruction. The following validation provides evidence of improved semantic representation under noisy conditions through index dropping.

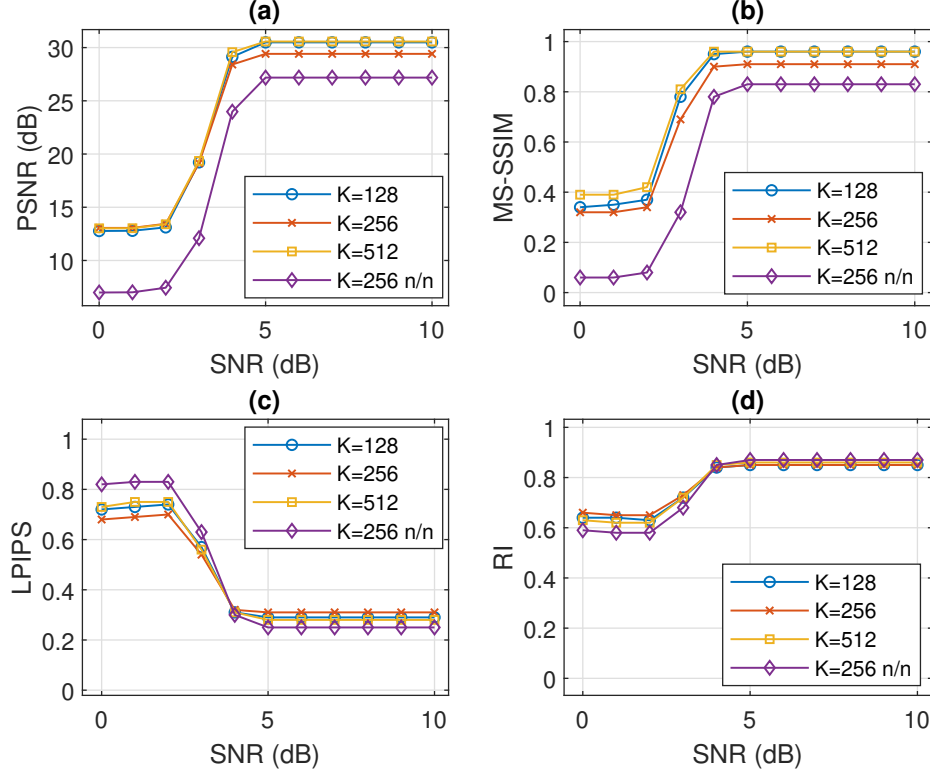


Figure 5.11: Ablation study evaluating different codebook sizes and training configurations, where n/n denotes no noise applied during training: (a) PSNR vs. SNR, (b) MS-SSIM vs. SNR, (c) LPIPS vs. SNR, and (d) RI vs. SNR. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

We used contrastive optimization to make the model robust to noise. To assess the methodology’s effectiveness, an ablation study is conducted by varying the dropout probability ($\mathbb{P}_{\text{mask}}$). By dropping the indices, it is expected to simulate index corruption and dropout during transmission. The results are shown in figure 5.12 for varying dropout rates across different SNR levels. A small dropout probability (0.01-0.1) introduces a controlled degree of stochastic degradation, encouraging the model to learn error-tolerant reconstructions. However, increasing the dropout rate to 0.5, meaning half the quantized vectors are corrupted, results in the loss of significant latent information, significantly degrading both semantic similarity and image quality, yet still maintaining a fair reconstruction of the images. Based on these validations, the experiments in the former sections were conducted at a dropout rate $\mathbb{P}_{\text{mask}} = 0.01$, providing improved

robustness to latent perturbations while preserving high reconstruction quality under realistic channel conditions. Additionally, we validated the design choice of Wasserstein distance based discriminator with adversarial training by training the model without the discriminator. As it consistently exhibits a lower quality, underscoring the role of the WGAN-based critic in enhancing perceptual realism.

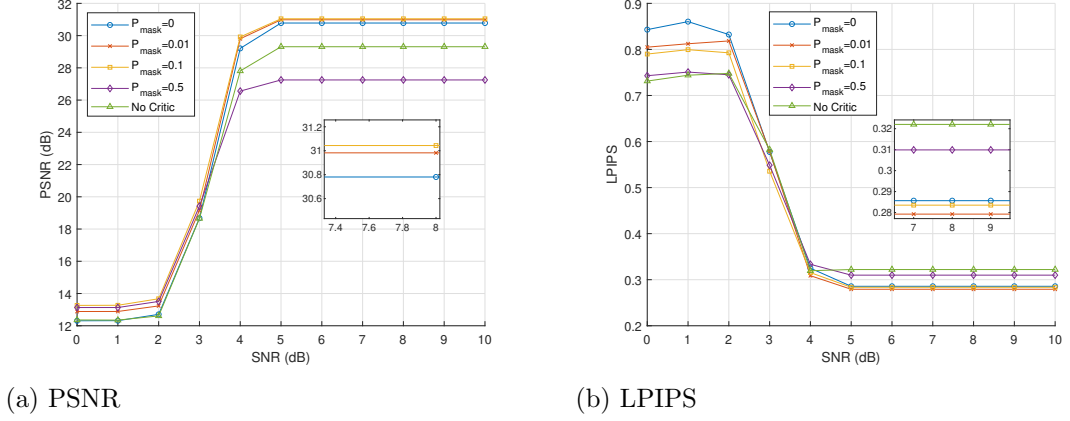


Figure 5.12: Model performance under different configurations across varying SNR levels: (a) PSNR (b) LPIPS. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

5.4.6 Generalization

An important objective of this research is to develop models that generalize to accurately interpret and reconstruct information across diverse, unseen environments, in contrast to current models that are confined to the patterns and characteristics of the training data. This is an important feature that the conceptual frameworks aim to achieve through minimal representation and semantic languages, enabling SemCom systems to operate effectively across diverse contexts, content types, and real-world scenarios. Retraining for each image resolution or dataset introduces substantial computational cost during training and increases deployment complexity. Models that perform without retaining or finetuning are highly desirable in practical scenarios such as rural telemedicine. To experimentally validate, FID [242] is calculated for a random image sample from the COCO dataset reconstructed with the trained model. Compared to other variants of GANs, including the model in [243], which is trained on the ImageNet dataset and

evaluated the FID on the COCO stuff dataset, SVQ-WGAN exhibits a lower FID in table 5.6. Further, the near-zero kernel inception distance (KID) score of 0.0073 [244] highlights the model’s capacity to align well with the original data distribution, indicating minimal divergence in the feature space.

Table 5.6: Model evaluation. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

Model	FID
SVQ-WAGN (proposed)	5.496
IC-GAN [243]	8.5
Unconditional (BigGAN) [245]	17.9
Unconditional (StyleGAN2) [246]	8.8

5.4.7 Semantic distortion

Semantic distortion was evaluated using the novel metric on randomly selected images from the COCO-Stuff [247] dataset. A pretrained object detection model, EfficientDet-D4 [248], is used to evaluate semantic distortion via category and IoU errors. The training datasets (DIV2K, CelebA-HQ, Flickr15k) align with the domain of natural images, to which the detector is trained. As illustrated in table 5.7, the images transmitted using BPG have higher distortion. In contrast, the proposed model shows lower task-oriented semantic distortion, validating its ability to preserve semantics.

Table 5.7: Task-oriented semantic distortion. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

Model	Category error	IoU error	Semantic distortion
SVQ-WGAN	0.3783	0.6671	0.4360
BPG	0.6547	0.8802	0.7009

5.4.8 Visual analysis

The scalability of SVQ-WGAN is evaluated across a range of image sizes, including high-resolution (DIV2K), facial (CelebA-HQ), and 4K images [249]. High-resolution images were divided into overlapping patches and encoded with SVQ-WGAN. Encoded output

is flattened and mapped to the codebook vector for quantization. The extracted indices of the corresponding quantized vectors are then transmitted through the LDPC transmitter as a binary stream and then over the communication channel. This process is depicted in figure 5.2. The received encoding indices are passed through a one-hot encoding layer. This is followed by a matrix multiplication with the codebook embeddings to recover the quantized latent vectors. The generator produces the reconstructed patches. The patches were blended by a Gaussian blending mask [250] into a complete image by leveraging the overlap and normalized to mitigate edge effects. This methodology ensures the transmission of high-resolution images even under challenging SNR conditions, and the model can reconstruct images of varying resolutions without retraining on different datasets.

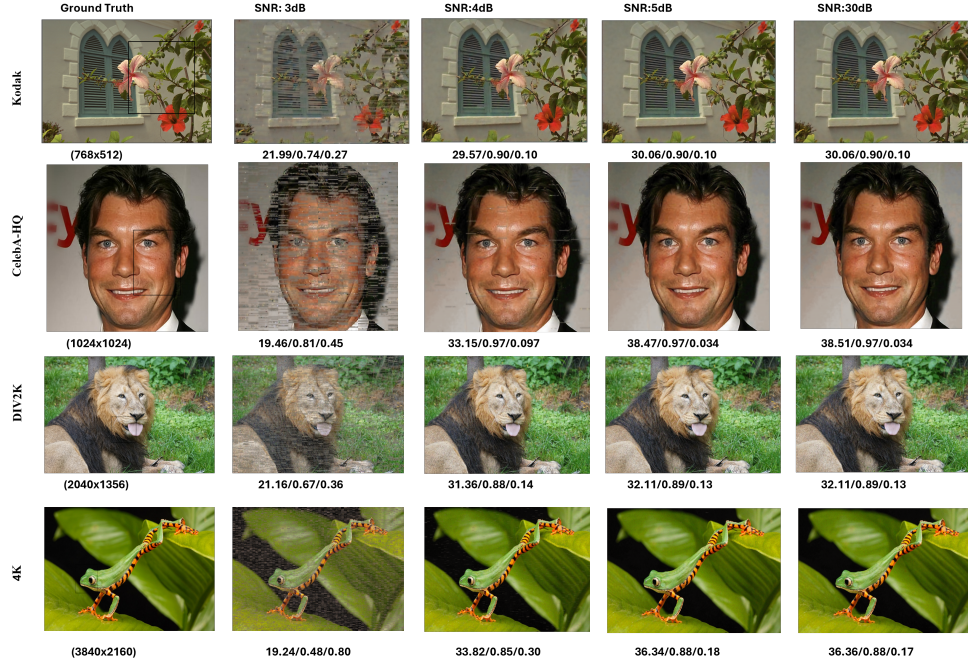


Figure 5.13: Image quality across different SNR levels for multiple datasets using QPSK modulation with LDPC coding at CR=0.5 over a fading channel, evaluated (PSNR/MS-SSIM/LPIPS). (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

Figure 5.13 presents the reconstructed images at various SNR levels for samples drawn from multiple datasets. The results highlight the method’s ability to deliver

high-quality reconstructions for human face images and similarly strong performance on high-resolution DIV2K samples, without relying on super-resolution modules or AI-based error-correction techniques at the receiver. This indicates that the proposed approach generalizes well across image resolutions without requiring task-specific training. Furthermore, figure 5.14 and figure 5.15 display enlarged patches corresponding to the highlighted regions in figure 5.13 under a no-noise setting, where the proposed method consistently exhibits superior perceptual quality.



Figure 5.14: Zoomed image patches under zero channel noise conditions for images from the Kodak and DIV2K datasets. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)



Figure 5.15: Zoomed image patches from the CelebA-HQ dataset and a 4K image under zero channel noise conditions. (Reprinted with permission from [2] under CC BY 4.0 license ©2025 IEEE)

5.5 Limitations

The proposed task-oriented distortion matrix limits its ability to generalize to a broader context because the object detection capability is restricted to the 80 classes on the

COCO dataset, which are the classes on which the object detection algorithm is trained. We opted not to include a semantic error correction module at the receiver to maintain low complexity, though this may limit extreme noise resilience below 3 dB. The model is trained on natural images, and its domain adaptability to content, including medical or satellite images, remains unexplored.

5.6 Conclusion

This chapter proposes a generative AI-based SemCom system conditioned on discrete feature vectors mapped to a codebook, which serves as the shared knowledge base and represents image features in a minimal, discrete vector form. The architecture achieves bandwidth savings by sending the codebook indices through the physical communication channel. The model demonstrated efficiency by integrating a lightweight encoder-decoder architecture, enhanced by a novel spatial attention mechanism that focuses on object edges, yielding superior performance in semantic feature extraction. The stability of the training process and perpetual quality are ensured by using a Wasserstein distance-based discriminator. Experimental results confirm that across different coding rates, the GenAI-based SemCom system outperforms BPG with LDPC channel coding in preserving perpetual quality, structural integrity, and robustness to noise. Quantitatively, SVQ-WGAN achieves an improvement of approximately 5 dB in PSNR, 0.7 in MS-SSIM, 0.45 in LPIPS, and over 0.1 in RI compared to BPG in Rayleigh fading channel with LDPC coding for QPSK. Moreover, introducing new metrics to measure robustness and task-oriented semantic distortion as a measure of semantic rate distortion underscores the model’s capability to surpass conventional bit-by-bit compression techniques. The trained model generalized to out-of-distribution data and high-resolution images without specialized training, reducing computational cost and simplifying deployment across diverse scenarios without task specific adaptation. The model demonstrates strong performance on both human-centric (captioning) and machine-centric (classification) tasks. This is enabled by a multi-level contrastive strategy spanning pixel, latent, perceptual, and semantic levels, which contributes to robustness and task consistency under varied SNR conditions. The channel-agnostic

nature of the design, which is compatible with standard coding schemes including LDPC, further reduces deployment complexity across diverse scenarios without requiring task or environment adaptation. These findings highlight the potential of codebook-based generative semantic communication systems to achieve higher efficiency, accuracy, and robustness to semantic and channel noise, marking a promising direction for future research in SemCom.

In future work, the SVQ-WGAN framework will be expanded toward a multimodal semantic communication system for video by incorporating a cross-modal feature fusion mechanism. This will include a semantic-level error correction module at the video frame level to correct the channel-introduced loss in frame transmission. This direction leverages the fusion of visual and text modalities to enhance semantic understanding and robustness during transmission. The methodology aims to align latent visual features extracted by the SVQ-WGAN encoder with pre-trained textual embeddings using a lightweight projection module. This enables joint reasoning across modalities and supports complex tasks, including cross-modal retrieval, referring expression grounding, and VQA, under channel constraints. This multimodal fusion design surpasses the unimodal captioning employed in this chapter, enabling richer semantic reasoning and greater resilience in low-bandwidth, noisy environments. In a separate effort, the domain adaptability of medical or satellite images will be explored.

Chapter 6

Semantic noise modeling and evaluation through shared knowledge base optimization

While previous chapters addressed semantic extraction and noise resilience, the design of an efficient shared knowledge base remains a key challenge in SemCom. Excessive redundancy, poor generalization, and under-utilized codebooks hinder efficient representation. This chapter presents a codebook-optimization methodology in SemCom that fulfills the pillars 1 and 2 of the conceptual framework in 1.1.1. The objective is to develop compressed yet expressive representations that improve robustness, generalization, and efficiency, and propose an effective evaluation methodology. The proposed methodology draws upon findings from recent studies on VQ-based SemCom and codebook development mechanisms, highlighting the potential for efficient, resource-aware transmission in practical communication channels.

The methodology included a soft-assignment-based quantization framework that integrates contrastive and entropy-based regularization to improve codebook diversity and semantic robustness. We further introduce an exponential moving average (EMA) codebook update mechanism to enhance stability and apply Wasserstein GAN-based adversarial training to improve distribution alignment. The main contributions of this chapter are summarized below:

Chapter 6. Semantic noise modeling and evaluation through shared knowledge base optimization

- We propose a joint regularization strategy with contrastive loss (to maximize semantic separation among embeddings) and an entropy loss (to encourage balanced codebook usage) in codebook-based SemCom. The methodology provides greater robustness and enhances semantic diversity, both of which are essential for stable task-level performance.
- The soft assignments replace hard codeword selection with a probabilistic weighting over all codebook embeddings, avoiding codebook collapse and enabling smoother latent representations.
- We introduce an EMA mechanism to update codebook embeddings to reduce instability in codeword updates. Contrastive, entropy-based, and EMA regularizations have been explored in discrete representation learning, and they have been jointly integrated into a unified SemCom framework that explicitly targets semantic robustness against both channel noise and semantic-layer perturbations.
- We incorporate a WGAN discriminator into the training pipeline to enforce semantic consistency and enhance reconstruction realism. This adversarial supervision helps suppress semantic noise by aligning the distribution of generated images with that of the original data, beyond pixel-level similarity.
- This chapter presents a comprehensive robustness evaluation framework that captures the effects of both physical-layer channel noise (via LDPC coded Rayleigh fading) [206] and semantic noise. The NVIDIA Sionna library is used to simulate a channel with QAM modulation based on real-world channel characteristics for the experiments. The proposed system demonstrates higher robustness than the baseline VQ-VAE and comparable performance against JSCC.
- This work introduces a novel FGSM based semantic robustness evaluation methodology to assess the system’s robustness to semantic noise. The method evaluates semantic noise robustness in receiver task accuracy (classification) in both the latent and pixel spaces.
- Experiments on CIFAR-10 [219] dataset demonstrate improved performance of the

Chapter 6. Semantic noise modeling and evaluation through shared knowledge base optimization

method in terms of codebook usage, perplexity, and semantic fidelity in comparison to VQ-VAE, BPG, and JSCC [71] systems. The model’s generalization to diverse datasets is evaluated on the Tiny-Imagenet dataset [220].

This chapter is based on the following article submitted for publication in IEEE Transactions on Vehicular Technology by M. Lokumarambage, T. Sivalingam, F. Dong, N. Rajatheva, A. Fernando, “Codebook-enabled semantic communication with contrastive and entropy-based optimization to reduce semantic noise.” The article is currently under review.

The structure of this chapter is as follows. Section 6.1 introduces the problem formulation, and Section 6.2 details the system architecture and presents the implementation details. Section 6.3 presents the simulation results, and Section 6.4 identifies the limitations of the methodology. Section 6.1 concludes the chapter.

6.1 Problem formulation

Semantic noise in a SemCom system occurs when latent space perturbations lead to a mismatch in the conveyed meaning or intent, despite the transmitted signal exhibiting high perceptual similarity. This mismatch becomes critical in M2M task-oriented SemCom systems. In systems using VQ with shared codebooks, where discrete embeddings represent abstract semantic units, we can expect a higher semantic distortion. Measuring the robustness and mitigation is essential. We define semantic noise as follows: Any perturbation δ such that $\mathbf{z}'_e = \mathbf{z}_e + \delta$ leading to a new reconstruction $\hat{x}' = G(Q(\mathbf{z}_e + \delta))$ causes semantic noise if,

$$f(\hat{x}) \neq f(\hat{x}'),$$

where $f(\cdot)$ is a downstream task. In this chapter, we define $f(\cdot)$ as the downstream classifier capturing semantic meaning in the form of an image class. To evaluate a system’s robustness to semantic noise, we use FGSM [251], which works by using the gradients of the neural network to create an adversarial example. We evaluate robustness to semantically disruptive noise using latent-space-based FGSM rather than

perceptually detectable distortions in the pixel space. The semantic impact of δ is evaluated by measuring the drop in classification accuracy, forming a measurable proxy for semantic robustness. VQ based systems are susceptible to semantic noise due to factors including non-linear quantization boundaries, discrete codeword jumps, codebook under-utilization, or codebook collapse. The proposed architecture is used to evaluate these issues via entropy and contrastive regularization, and EMA updates.

6.2 System model

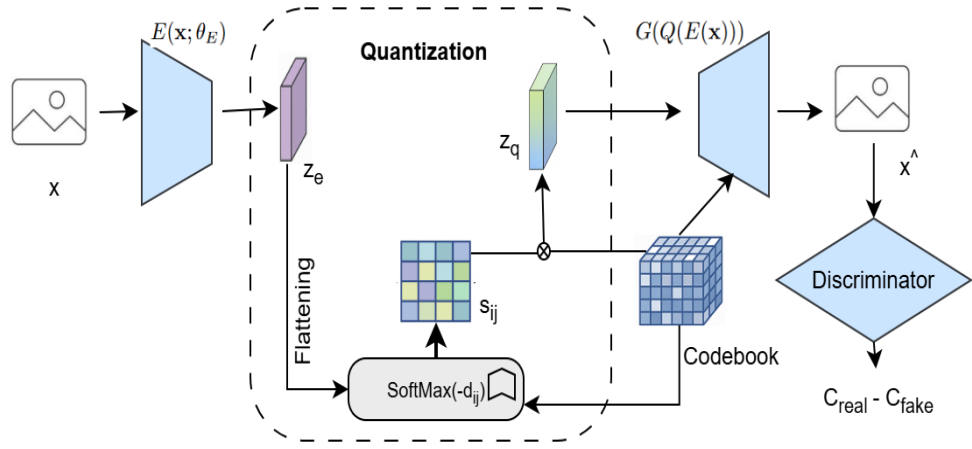


Figure 6.1: Training phase: continuous feature representations are quantized using the learned codebook and soft assignments. The complete training process is defined in algorithm 4.

This chapter proposes a novel quantization mechanism to VQ-based SemCom, comprising a semantic encoder, a VQ layer, a generator, and a discriminator by introducing codebook optimization. This novel methodology encourages an efficient representation of the input data distribution. It optimizes codebook embeddings, leading to high-quality, task-relevant generations by avoiding semantic noise introduced by the semantic extraction and semantic coding processes, which corresponds to Level B of the improved theoretical model of SemCom described in the previous chapter. This chapter addresses the codebook optimization problem by introducing a novel mechanism to validate the hypothesis that optimizing the semantic knowledge base reduces semantic noise, and by

Chapter 6. Semantic noise modeling and evaluation through shared knowledge base optimization

providing a comprehensive experimentation framework to evaluate the semantic noise associated with it, complementing the works introduced in the previous chapters.

6.2.1 The quantization process

The encoder E maps the input image $\mathbf{x} \in \mathbb{R}^{h \times w \times c}$, (where h , w , and c denote the height, width, and number of channels) into a latent representation $\mathbf{z}_e \in \mathbb{R}^{a \times b \times d}$, where a , b , and d denote the dimensionality of the latent space. Here, $\mathbf{z}_e$ is defined as

$$\mathbf{z}_e = E(\mathbf{x}; \theta_E), \quad (6.1)$$

where θ_E are the learnable parameters of the encoder. The encoder comprises a series of convolutional layers with strided downsampling to reduce spatial dimensions, thereby providing a continuous latent representation that effectively captures key semantics. VQ then maps this continuous representation into a discrete representation. In contrast to conventional quantization, soft assignments use a weighted combination of embeddings to assign the nearest embedding vector, enabling differentiability. Soft assignments are probabilities that represent how much each embedding vector in the codebook is “assigned” to a given input vector. The distance between $\mathbf{z}_{e,i}$ and corresponding embedding $\mathbf{e}_j$ is calculated using the squared Euclidean distance as

$$d_{ij} = \|\mathbf{z}_{e,i} - \mathbf{e}_j\|^2. \quad (6.2)$$

We map the negative distances $-d_{ij}$ using a softmax function to obtain the soft assignment probability (s_{ij}) as

$$s_{ij} = \frac{\exp(-d_{ij})}{\sum_{k=1}^K \exp(-d_{ik})}, \quad (6.3)$$

where K is number of total embeddings and s_{ij} is the probability of assigning $\mathbf{z}_{e,i}$ to embedding $\mathbf{e}_j$. Using negative distances gives larger values to closer embeddings. The quantized vector $\mathbf{z}_q$ is the weighted sum of the embeddings using the soft assignment

probability defined as

$$\mathbf{z}_q = \sum_{j=1}^K s_{ij} \mathbf{e}_j. \quad (6.4)$$

Soft assignments encourage a smooth mapping between embedding vectors instead of a direct mapping to the nearest embedding. The quantization process is shown in algorithm 2.

Algorithm 2 Quantization Process (Forward pass)

- 1: **Input:** Encoder outputs $\mathbf{z}_e$, codebook $\{\mathbf{e}_j\}_{j=1}^K$
 - 2: **Output:** Quantized vectors $\mathbf{z}_q$
 - 3: **for** each input $z_{e,i}$ **do**
 - 4: Compute distances: $d_{ij} = \|\mathbf{z}_{e,i} - \mathbf{e}_j\|^2$
 - 5: Compute soft assignment: $s_{ij} = \text{softmax}(-d_{ij})$
 - 6: Quantize using weighted sum: $\mathbf{z}_{q,i} = \sum_{j=1}^K s_{ij} \mathbf{e}_j$
 - 7: **end for**
-

We maintain a representative codebook by updating a EMA based codebook vector, which tracks both the frequency of codebook usage and the contribution of each embedding vector in the input space. For each step, we maintain an EMA cluster size M_j , which is the count of how often the embedding is used, and E_j is the sum of all encoder outputs for a particular codebook vector $\mathbf{e}_j$. At each training step t , the cluster size is updated as

$$M_j^{(t)} = \gamma M_j^{(t-1)} + (1 - \gamma) \sum_{i=1}^N v_{ij}, \quad (6.5)$$

where $v_{ij} \in \{0, 1\}$ is one-hot encoding of the soft assignment showing whether the i -th encoder output vector $\mathbf{z}_{e,i} \in \mathbb{R}^d$ is assigned to the j -th embedding vector $\mathbf{e}_j$, $M_j \in \mathbb{R}$ is the cluster size representing the effective number of assignments to codebook embedding $\mathbf{e}_j$ and γ is the EMA decay factor. The aggregated embedding sum is updated as

$$\mathbf{E}_j^{(t)} = \gamma \mathbf{E}_j^{(t-1)} + (1 - \gamma) \sum_{i=1}^N v_{ij} \mathbf{z}_{e,i}, \quad (6.6)$$

where $\mathbf{E}_j \in \mathbb{R}^d$ is the aggregated sum of all encoder output vectors assigned to $\mathbf{e}_j$. The updated embedding $\mathbf{e}_j^{(t)}$ is normalized by $M_j^{(t)}$ to obtain the new embedding for the step given by

$$\mathbf{e}_j^{(t)} = \frac{\mathbf{E}_j^{(t)}}{M_j^{(t)} + \epsilon}, \quad (6.7)$$

where ϵ is used to avoid division by zero. The codebook embedding update is shown in algorithm 3. Semantic noise arises in SemCom when perturbations in the latent representation cause shifts in the conveyed meaning, even if the reconstructed signal remains perceptually similar. An EMA-based approach incrementally adjusts each codebook vector based on both recent assignments and historical usage, thereby smoothing updates and mitigating sudden fluctuations compared to direct gradient-based or hard-assignment methods. This incremental update, combined with previous assignments, maintains the semantic integrity of each codebook vector over time by mitigating the risk of semantic shifts. By acting as a temporal filter, EMA stabilizes codebook evolution and maintains distinct, semantically coherent embeddings, thereby enhancing resilience against semantic drift and improving end-to-end meaning preservation in the presence of semantic noise due to quantization in the SemCom pipeline.

Algorithm 3 Codebook embedding update (EMA)

- 1: **Input:** Flattened inputs $\mathbf{z}_e$, one-hot assignments v_{ij} , decay γ , previous cluster sizes $M_j^{(t-1)}$, previous embedding sums $\mathbf{E}_j^{(t-1)}$
 - 2: **Output:** Updated embeddings $\{\mathbf{e}_j^{(t)}\}$
 - 3: **for** $j \leftarrow 1$ to K **do**
 - 4: Update cluster size: $M_j^{(t)} = \gamma M_j^{(t-1)} + (1 - \gamma) \sum_i v_{ij}$
 - 5: Update embedding sum: $\mathbf{E}_j^{(t)} = \gamma \mathbf{E}_j^{(t-1)} + (1 - \gamma) \sum_i v_{ij} \mathbf{z}_{e,i}$
 - 6: Normalize embedding: $\mathbf{e}_j^{(t)} = \frac{\mathbf{E}_j^{(t)}}{M_j^{(t)} + \epsilon}$
 - 7: **end for**
-

The generator (G) uses transposed convolution layers to upsample the spatial dimension to generate the image from $\mathbf{z}_q$ as

$$\hat{\mathbf{x}} = G(\mathbf{z}_q; \theta_G), \quad (6.8)$$

where $\hat{\mathbf{x}}$ is the generated image and θ_G are the generator parameters.

6.2.2 Discriminator

We used a Wasserstein distance-based discriminator to evaluate the realism of the generated images, similar to the implementation in the previous chapter. By penalizing large distributional shifts between reconstructed and real data, Wasserstein loss enforces that the semantic decoder remains within the underlying structure of the input, even when subject to compression or channel noise. This distance metric captures not only pixel-level similarity but also the semantic geometry of the data manifold, supporting conceptually robust reconstructions. The discriminator consists of three convolutional blocks with channel sizes 64, 128, and 256, respectively; each block includes a 3×3 convolution (stride 2, padding “same”), a LeakyReLU activation with $\alpha = 0.2$. Layer normalization is applied only in the second and third convolutional blocks to stabilize deeper-layer activations and improve gradient flow, while avoiding normalization in the first block helps preserve local feature variability, which is important for early-stage representation learning. Dropout is applied after each convolutional block (with a rate of 0.3) to prevent overfitting of the discriminator and maintain training stability by regularizing feature learning. The final dense layer maps the flattened feature representation to a single scalar output, serving as the critic score in the Wasserstein distance formulation, without applying an activation function. We applied a gradient penalty to enforce smoothness in the discriminator.

6.2.3 Loss function

The model is optimized during training using a series of loss terms based on pixel-wise differences and adversarial training. Reconstruction loss measures the pixel-wise difference between the input and the generated image, expressed as

$$\mathcal{L}_{\text{recon}} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2. \quad (6.9)$$

We use the perpetual loss term to align the generated images with human visual perception. The calculation uses the ResNet-50 [233] network to extract features from

ground-truth and generated images. Perpetual loss is calculated as,

$$\mathcal{L}_{\text{perc}} = \|\Phi(\mathbf{x}) - \Phi(\hat{\mathbf{x}})\|^2, \quad (6.10)$$

where $\Phi(\cdot) : \mathbb{R}^{h_p \times w_p \times c_p} \rightarrow \mathbb{R}^{d_p}$ denotes the feature extraction function implemented by a ResNet-50 network pretrained on ImageNet.

Contrastive loss optimizes the codebook by promoting diverse and balanced utilization of embeddings and serves as a regularization to prevent embeddings from becoming similar. This is achieved by maximizing the pairwise squared Euclidean distances between all embeddings. The loss penalizes the embeddings that are close to each other by negating the pairwise distances, which gives high values for small distances in the contrastive loss term. This encourages the model to spread out the embeddings. Contrastive loss is given by

$$\mathcal{L}_{\text{contrastive}} = \mathbb{E} \left[\max \left(0, 1 - \frac{\mathbf{z}_e \cdot \mathbf{e}_j}{\|\mathbf{z}_e\| \|\mathbf{e}_j\|} \right) \right]. \quad (6.11)$$

Contrastive loss is essential when two embeddings representing semantically distinct concepts are too close in the latent space, channel noise, or minor feature perturbations may cause their assignments to swap, resulting in semantic errors at the receiver. By maximizing distances between the embeddings, contrastive loss reduces such assignments, ensuring that each embedding captures a distinct semantic concept and is robust to perturbations while channel transfer. Entropy-based regularization plays an important role in improving codebook utilization and preventing the model from becoming overly dependent on a small subset of embeddings. The objective is to maximize entropy over the distribution of selected embeddings. This improves codebook utilization and diversity in quantization and is given by

$$\mathcal{L}_{\text{entropy}} = -\mathbb{E}_i \left[\sum_{j=1}^K s_{ij} \log(s_{ij}) \right], \quad (6.12)$$

where $\mathbb{E}_i[\cdot]$ is the expectation operator that averages over all input vectors in the batch. Entropy loss complements contrastive loss by encouraging uniform usage of the

Chapter 6. Semantic noise modeling and evaluation through shared knowledge base optimization

entire codebook. Without this regularization, the SemCom system may overuse a small subset of embeddings, leading to codebook collapse, reduced representational capacity of the codebook embeddings, and forcing multiple semantics to share the same code, thereby increasing semantic ambiguity and semantic noise. Maximizing entropy over the embedding selection distribution ensures that all codebook entries remain active and representative of different semantics in the latent space. In addition to these novel loss terms, we have also used commitment loss and codebook loss terms. Commitment loss ensures that the encoder produces encoder output $\mathbf{z}_e$ that is well represented by the quantized vectors. Commitment loss is defined as

$$\mathcal{L}_{\text{commit}} = \|\mathbf{z}_e - \text{sg}(\mathbf{z}_q)\|_2^2, \quad (6.13)$$

where $\text{sg}(\cdot)$ is the stop-gradient operator. Codebook loss ensures embeddings in the codebook stay close to the encoder output:

$$\mathcal{L}_{\text{codebook}} = \|\text{sg}(\mathbf{z}_e) - \mathbf{z}_q\|_2^2. \quad (6.14)$$

In combination, the loss terms optimize model training such that the codebook loss aligns embeddings with encoder outputs, the commitment loss ensures encoder outputs commit to a specific embedding, the contrastive loss spreads embeddings apart to maintain semantic distinctiveness, and the entropy loss ensures balanced usage to avoid representational bottlenecks. Together, they produce a well-separated, fully utilized codebook that reduces semantic noise by making latent assignments more stable, distinct, and meaning-preserving under noisy channel conditions. The total loss is a weighted sum given by

$$\begin{aligned} \mathcal{L}_{\text{total}} = & \mathcal{L}_{\text{recon}} + \eta_{\text{adv}}\mathcal{L}_{\text{adv}} + \eta_{\text{perc}}\mathcal{L}_{\text{perc}} + \eta_{\text{commit}}\mathcal{L}_{\text{commit}} \\ & + \eta_{\text{code}}\mathcal{L}_{\text{codebook}} + \eta_{\text{cont}}\mathcal{L}_{\text{contrastive}} + \eta_{\text{ent}}\mathcal{L}_{\text{entropy}}, \end{aligned} \quad (6.15)$$

where $\eta_{\text{commit}}, \eta_{\text{cont}}, \eta_{\text{ent}}, \eta_{\text{code}}$ are hyperparameters. The train algorithm is defined in algorithm 4.

6.2.4 Semantic inference pipeline

The encoder extracts latent features from input images, and the vector quantizer compresses these features into quantized latent representations. These representations are serialized into bitstreams and transmitted through a simulated physical-layer. All simulations use a Sionna-based flat Rayleigh-fading channel with AWGN. The received vector at symbol time t is modeled as $\mathbf{y}_t = \mathbf{H}_t \mathbf{x}_t + \mathbf{n}_t$, where $\mathbf{H}_t \in \mathbb{C}^{N_r \times N_t}$ has zero-mean, unit-variance. We evaluate with $N_t = N_r = 1$ to provide a controlled, widely used baseline that isolates semantic-layer gains, where N_t is the number of transmitter antennas and N_r is the number of receiver antennas. The noise $\mathbf{n}_t \sim \mathcal{CN}(\mathbf{0}, N_0 \mathbf{I}_{N_r})$, with N_0 computed from the target E_s/N_0 accounting for modulation order M_C and LDPC code rate $R_c = k_c/n_c$. We employ 4-QAM, soft-output demapping, and linear minimum mean square error (MMSE) equalization. For channel coding, we use $k_c = 512$, $R_c \in \{1/2, 2/3, 3/4\}$. The shared codebook between the transmitter and receiver ensures that transmitted vectors retain their semantic meaning.

6.2.5 Interpretation of codebook clustering

We reinterpret the DBSCAN clustering [252] of codebook embeddings as an experiment that evaluates the model’s ability to learn conceptually meaningful representations. In a robust SemCom system, codebook vectors should align with high-level, generative factors including object category or scene semantics, rather than low-level pixel correlations. To evaluate this, we apply DBSCAN to the t-distributed stochastic neighbor embedding (t-SNE)-reduced embeddings of our codebook and examine whether the resulting clusters align with CIFAR-10 class labels. The quantized latent representation $\mathbf{z}_q$ of an input image $\mathbf{x}$ is mapped to a set of codebook indices $K = \{k_{ij} \mid i \in \{1, \dots, h\}, j \in \{1, \dots, w\}\}$, where k_{ij} corresponds to the codebook embedding $\mathbf{e}_{k_{ij}}$. For a dataset of N samples, the indices k are aggregated for each class c_{ls} as $\mathcal{S}_{c_{ls}} = \bigcup_{n:y_n=c_{ls}} \mathbf{k}_n$. The frequency of a codebook index k for class c_{ls} is computed as $n_c(k) = \sum_{k' \in \mathcal{S}_c} \mathbb{1}(k' = k)$. The embeddings $\mathcal{E} = \{\mathbf{e}_k \mid k = 1, \dots, K\}$ are reduced to a two-dimensional space $\mathbf{u}_k = \text{t-SNE}(\mathbf{e}_k)$ for clustering using DBSCAN, producing clusters $\mathcal{C}_\ell = \{k \mid \mathbf{u}_k \in \text{cluster } \ell\}$. The class contribution to each cluster is quantified as $m_{\ell,c} = \sum_{k \in \mathcal{C}_\ell} n_c(k)$, and the top-5 classes

Algorithm 4 Training procedure of the proposed model with WGAN and EMA-based codebook

```

1: Input: Training data  $\mathbf{x}$ , encoder  $E$ , decoder  $G$ , vector quantizer  $Q$ , discriminator
    $D$ , noise layer  $n$ , perceptual model  $\Phi$ , training variance  $\sigma^2$ 
2: Output: Trained generator (encoder–decoder) and discriminator
3: for each training step do
4:   Step 1: Train the discriminator
5:   for  $i \leftarrow 1$  to discriminator_steps do
6:     Sample mini-batch of real images  $\mathbf{x}$ 
7:     Compute quantized reconstructions:  $\hat{\mathbf{x}} = G(Q(E(\mathbf{x})))$ 
8:     Compute critic scores:  $D(\mathbf{x})$  and  $D(\hat{\mathbf{x}})$ 
9:     Compute discriminator loss:  $\mathcal{L}_{adv} = \mathbb{E}[D(\hat{\mathbf{x}})] - \mathbb{E}[D(\mathbf{x})]$ 
10:    Compute gradient penalty  $\mathcal{L}_{GP}$  on interpolated samples
11:    Update  $D$  using loss  $\mathcal{L}_{adv} + \lambda\mathcal{L}_{GP}$ 
12:   end for
13:   Step 2: Train the generator (encoder–decoder with VQ)
14:   Compute encoder output:  $\mathbf{z}_e = E(\mathbf{x})$ 
15:   Apply vector quantization:  $\mathbf{z}_q = Q(\mathbf{z}_e)$ 
16:   Reconstruct image:  $\hat{\mathbf{x}} = G(\mathbf{z}_q)$ 
17:   Compute reconstruction loss:  $\mathcal{L}_{recon} = \|\mathbf{x} - \hat{\mathbf{x}}\|^2 / \sigma^2$ 
18:   Compute perceptual loss:  $\mathcal{L}_{perc} = \|\Phi(\mathbf{x}) - \Phi(\hat{\mathbf{x}})\|^2$ 
19:   Compute VQ-related loss:
        $\mathcal{L}_{VQ} = \eta_{code}\mathcal{L}_{codebook} + \eta_{commit}\mathcal{L}_{commit} + \eta_{cont}\mathcal{L}_{contrastive} + \eta_{ent}\mathcal{L}_{entropy}$ 
20:   Compute total generator loss:
        $\mathcal{L}_{total} = \mathcal{L}_{recon} + \eta_{perc}\mathcal{L}_{perc} + \mathcal{L}_{VQ}$ 
21:   Update  $E$ ,  $G$ , and  $Q$  to minimize  $\mathcal{L}_{total}$ 
22: end for

```

for each cluster are determined. Clusters are visualized in the t-SNE space with centers $\boldsymbol{\mu}_\ell = \frac{1}{|\mathcal{C}_\ell|} \sum_{k \in \mathcal{C}_\ell} \mathbf{u}_k$ and radii $r_\ell = \frac{1}{|\mathcal{C}_\ell|} \sum_{k \in \mathcal{C}_\ell} \|\mathbf{u}_k - \boldsymbol{\mu}_\ell\|$, annotated with their top-5 classes and contributions. This alignment indicates that the learned codebook vectors abstract over visual variability to reflect latent causal concepts (e.g., “car”, “truck”, “bird”), even though no class supervision is used during training. This clustering behavior supports the hypothesis that our combination of contrastive and entropy-based optimization not only improves codebook utilization but also encourages alignment with semantically and causally coherent representations.

FGSM based semantic noise evaluation

Practical sources of distortion, including symbol index flips due to channel fading or quantization mismatch, can be modeled as latent perturbations that may alter the conveyed intent even when the reconstructed signal remains visually close to the original. Evaluating robustness with FGSM in latent space, therefore, provides a rigorous way to test semantic stability. Similar experiments with pixel space perturbations have been explored in [90] to assess semantic robustness through masked VQ-VAE codebooks. The authors of [80] investigate robustness via a vector-to-index transformer approach, and latent adversarial autoencoder studies [253] show how perturbations in latent variables expose weaknesses in learned representations. We use FGSM in pixel and latent spaces to evaluate the robustness of the proposed method. Pixel-space FGSM is used to manipulate the input image semantically. This perturbation directly manipulates the input pixels within a bounded ϵ , enabling assessment of the model’s robustness to visually imperceptible yet semantically misleading changes. We get the adversarial image x_{adv} , as

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}_c(f(x), y_l)), \quad (6.16)$$

where ϵ is the pixel perturbation magnitude, and $\mathcal{L}_c$ is the categorical cross-entropy loss with respect to class label y_l .

Latent-space (semantic) FGSM is used in the proposed test setup to apply adversarial perturbations directly in the latent space rather than the pixel domain. Let $E(\cdot)$ denote the encoder, $Q(\cdot)$ the vector quantizer, and $G(\cdot)$ the generator that reconstructs the image from quantized latents. The gradient of the classification loss $\mathcal{L}_c(f(G(Q(z_e))), y_l)$ with respect to z_e is computed, and a perturbation is applied according to

$$z_e^{\text{adv}} = \mathbf{z}_e + \epsilon_z \cdot \text{sign}(\nabla_{\mathbf{z}_e} \mathcal{L}). \quad (6.17)$$

The perturbed latent representation is then quantized. The generated image under latent perturbation is

$$\hat{x}' = G(Q(\mathbf{z}_e^{\text{adv}})). \quad (6.18)$$

Chapter 6. Semantic noise modeling and evaluation through shared knowledge base optimization

Latent-space FGSM attack serves as a proxy for semantic noise, directly disturbing the codebook-based representation and enabling a controlled evaluation of the system’s semantic robustness. The results are illustrated in section 6.3.3.

6.3 Simulation results

We conducted tests on the model, training it on the CIFAR-10 dataset with codebook sizes of 64, 128, and 256. We evaluated codebook usage and perplexity for different codebook sizes and measured transmission accuracy for a codebook size of 64. Additionally, we assessed the model’s generalization on diverse datasets, including Tiny-Imagenet samples.

Implementation details

Data augmentation was applied to expand and diversify the training dataset, improving model generalization. The transformations included random rotations (up to 15°), horizontal and vertical shifts (up to 10% of the image dimensions), shearing, zooming, and horizontal flipping. 10,000 augmented images were generated from randomly selected samples in the training set. This approach enhanced the model’s robustness by exposing it to a broader range of variations, reducing the risk of overfitting. We set the hyperparameters $\eta_{\text{commit}} = 0.25$, $\eta_{\text{cont}} = 0.1$, $\eta_{\text{ent}} = 0.1$. We employed a warm-up cosine learning rate schedule with a base learning rate of 0.001, a warm-up learning rate of 1×10^{-5} and 15% of the total steps allocated to the warm-up phase. The schedule ensures stable initial updates and smooth convergence through cosine decay over the remaining training steps.

6.3.1 Codebook analysis

In this experiment, we intend to evaluate the diversity, utilization, and representation efficiency of the learned semantic representations. Figure 6.2 illustrates the effectiveness of the proposed codebook optimization strategy compared to baseline VQ-VAE, based on three quantitative indicators: codebook usage, perplexity, and codebook embedding

entropy.

Codebook usage is defined as the number of codebook vectors that are actively used during inference, and we measured this for the CIFAR-10 test dataset. Higher usage indicates that the learned latent representation distributes semantic information more uniformly across the codebook rather than collapsing to a small subset of vectors for decoding. The results in figure 6.2a show that the proposed model achieves significantly higher utilization compared to the baseline VQ-VAE. The proposed method achieves 96.88% usage for a codebook size of 64, while the VQ-VAE baseline uses only 60.94% of its embeddings. Even with a higher codebook size, the proposed methodology gains higher codebook usage as a percentage, embedding size 128 with 92.97% usage and embedding size 256 with 70.31%. It is evident that with large embedding sizes, a higher percentage of codebooks are not utilized during inference, providing an indication for optimal codebook size in design. The results indicate that the proposed training strategy effectively prevents codebook collapse and promotes richer semantic representation.

Figure 6.2b further analyzes the learned codebook structure using perplexity and codebook embedding entropy. Perplexity [254] reflects the effective number of embeddings used by the encoder. Formally, perplexity is defined as the exponentiated entropy of the codebook selection distribution:

$$\text{Perplexity} = \exp \left(- \sum_{i=1}^K p_i \log(p_i + \epsilon) \right),$$

where p_i is the normalized frequency of the i -th codebook embedding and ϵ is a small constant to avoid instability when $p_i = 0$. Higher Perplexity means that the semantic space is well distributed, capturing diverse semantic concepts, and the learned language is expressive. Similarly, embedding entropy quantifies the uncertainty or distribution uniformity of codeword selection and provides a measure of representation diversity.

$$H_{\text{emb}} = - \sum_{i=1}^K p_i \log p_i$$

For example, if $p_i = 1$ and the rest are zero, the entropy becomes zero.

Chapter 6. Semantic noise modeling and evaluation through shared knowledge base optimization

The values in table 6.1 demonstrate that both perplexity and entropy increase as the embedding size grows, indicating improved representation capacity. The proposed method maintains high perplexity values for embedding sizes of 64, 128, and 256, validating balanced codeword utilization. The corresponding entropy values follow a similar trend, confirming that the learned codebook captures diverse semantic patterns in the data. Overall, these results indicate that the proposed semantic codebook optimization approach produces more expressive and balanced semantic embeddings, which result in improved reconstruction quality, higher semantic fidelity, and greater robustness to semantic noise in the communication system, as validated in the following sections.

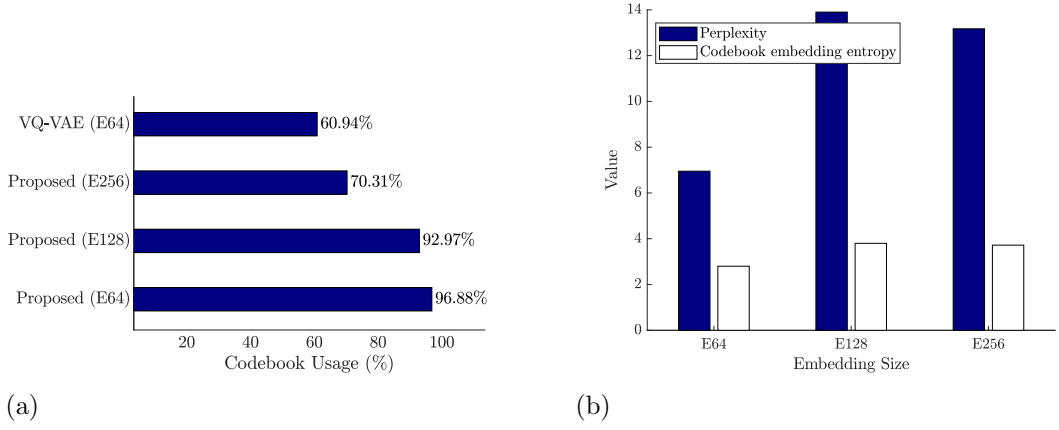


Figure 6.2: (a) Codebook usage for different codebook sizes for the proposed model and VQ-VAE. (b) Perplexity and codebook embedding entropy for different codebook sizes.

Table 6.1: Perplexity and codebook embedding entropy values for different embedding sizes.

Embedding Size	Perplexity	Embedding Entropy
E64	6.95	2.80
E128	13.90	3.80
E256	13.17	3.72

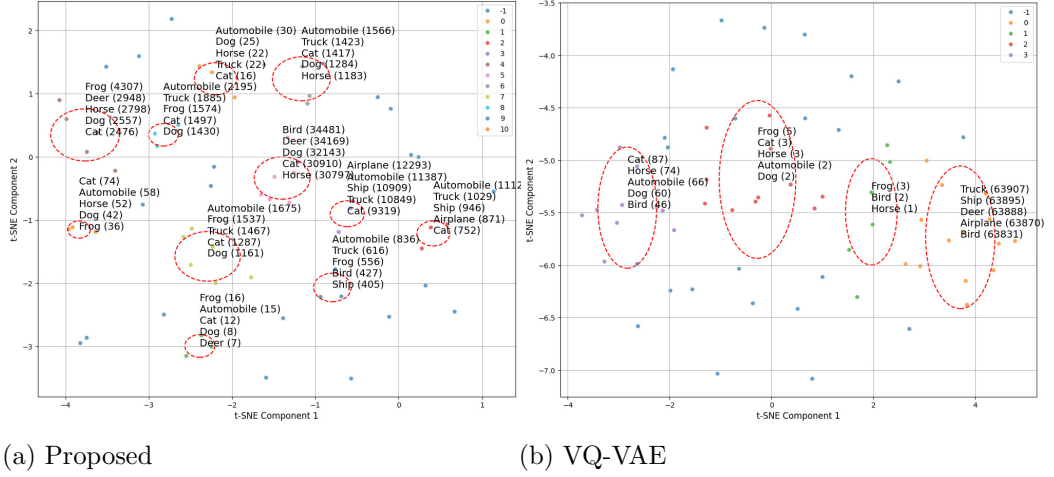


Figure 6.3: DBSCAN clustering of codebook embeddings for CIFAR-10 class labels and clustering circles.

6.3.2 Clustering

The clustering patterns in figure 6.3 reveal a critical insight into how our model reduces semantic noise. In the proposed system, embeddings form tight clusters aligned with CIFAR-10 semantic classes, indicating that the codebook captures concept-level abstractions. These abstractions are resilient to semantic changes in the input (e.g., background variation), ensuring that the intended meaning, the object class, is preserved throughout encoding and decoding. In contrast, the VQ-VAE baseline exhibits dispersed, overlapping clusters, in which embeddings fail to separate causes from correlations. Such entanglement can lead to semantic drift, where minor perturbations in the input result in significant changes to the reconstructed message’s meaning. Thus, clustering is not just a representational improvement; it directly contributes to semantic noise suppression by maintaining message intent throughout the communication process.

The comparison of codebook vector clustering between the proposed model in figure 6.3a and VQ-VAE in figure 6.3b, as visualized through t-SNE plots, reveals significant differences in grouping of semantic information through CIFAR-10 class labels. The VQ-GAN demonstrates tighter, more compact clusters with higher class homogeneity, indicating that its codebook vectors are more semantically meaningful and robust against semantic noise. This compactness is advantageous for SemCom systems, as it

Chapter 6. Semantic noise modeling and evaluation through shared knowledge base optimization

facilitates efficient transmission with a reduced risk of semantic ambiguity. Conversely, the VQ-VAE exhibits more dispersed clusters and greater class overlap, suggesting a less precise encoding of semantic information. While the VQ-VAE’s broader representations may capture variability, this dispersion introduces potential vulnerabilities to semantic noise, as overlapping regions may lead to misinterpretation or loss of semantic clarity during communication. A higher number of vectors in VQ-VAE fall into noise in the clustering algorithm than in the proposed model, representing underutilization. Overall, the proposed method is better suited for semantic noise reduction due to its well-separated clusters, which enhance the robustness and accuracy of transmitted semantic representations.

6.3.3 Semantic noise robustness

For the downstream classification task, we used a ResNet-50 backbone pre-trained on ImageNet and fine-tuned on CIFAR-10. The images are first resized to 224×224 and normalized, with standard augmentation (random flips, rotations, and zooms) applied to improve generalization. The ResNet-50 base is initially frozen, followed by fine-tuning of the full network at a low learning rate to adapt to the CIFAR-10 domain. A global average pooling layer and batch normalization are applied on top of the convolutional features, followed by a ReLU-activated dense layer with dropout regularization, and a final 10-way softmax output. Training uses Adam with categorical cross-entropy and early stopping, providing a strong baseline classifier for robustness evaluation.

For pixel-space FGSM, perturbations were applied directly to resized 224×224 images, and for latent-space FGSM, perturbations were introduced at the encoder output $\mathbf{z}_e$. The perturbed latents were then quantized and decoded, and the classifier evaluated the reconstructed images. For pixel FGSM, we used $\epsilon \in [0, 0.02]$, corresponding to 0–5/255 on the normalized $[0, 1]$ scale, a standard range in adversarial literature [255] that ensures distortions remain imperceptible to barely perceptible. For latent FGSM, we adopted the same range $\epsilon_z \in [0, 0.02]$, which is small relative to the encoder’s activation scale but sufficient to induce codeword changes in the quantizer. Accuracies were computed with a batch size of 128, and results are reported as the average top-1

classification accuracy over the entire CIFAR-10 test set, as shown in figure 6.4.

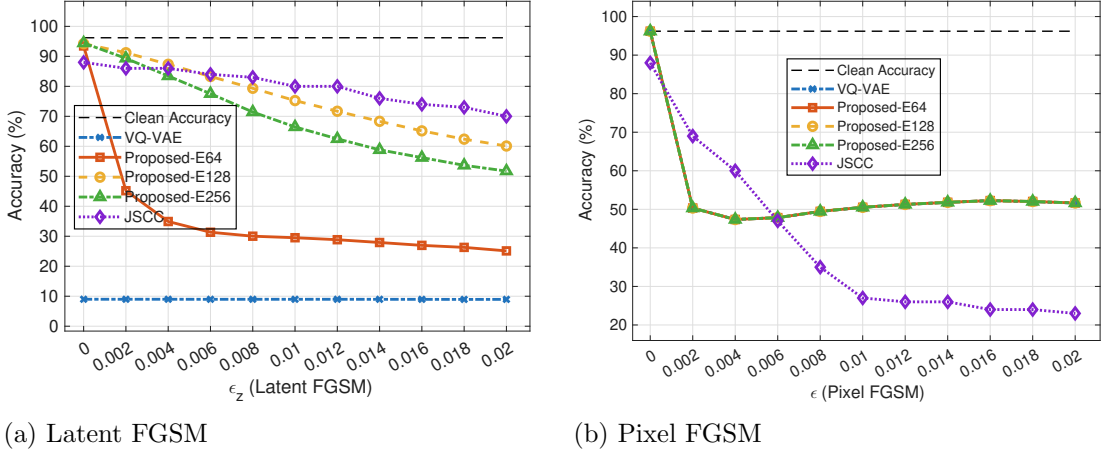


Figure 6.4: Classification accuracy under semantic-level (latent) and pixel-level perturbations for varying codebook sizes 64,128, and 256 of the proposed system.

In the latent FGSM setting shown in figure 6.4a, classification accuracy falls noticeably for all systems as ϵ increases. Compared with the VQ-VAE baseline, the proposed models demonstrate greater resilience, particularly when larger codebook sizes (128 and 256) are used, since these configurations provide richer codebook usage and more reliable latent representations. The JSCC scheme achieves the highest accuracy with a higher perturbation range. This outcome is valid because JSCC employs an analog transmission strategy that does not rely on discrete codebook indices and is therefore less vulnerable to small perturbations in the latent domain. In contrast, quantized embeddings in VQ-based systems are more susceptible to semantic errors once perturbations shift encoder outputs across codeword boundaries. In the pixel FGSM case, shown in figure 6.4b, the proposed system and the VQ-VAE exhibit similar behavior: an initial drop in accuracy, followed by stabilization around 50–60%. JSCC accuracy falls much lower compared to the discrete latent models because the perturbation is applied in the pixel space, and the generative nature of the proposed model can mitigate the effect of pixel-level perturbations.

6.3.4 Quality of transmission

The proposed model exhibits superior PSNR performance, as shown in table 6.2, especially at higher SNR values, compared to traditional methods including BPG and deep learning-based VQ-VAE. It shows comparable performance with JSCC at SNRs above 5 dB, where JSCC is robust for channel noise at low SNR regions. The efficient use of the codebook results in better-quality generations than VQ-VAE. The generative capability of the proposed model allows for efficient recovery of lost details during transmission by leveraging learned feature priors to generate high-quality images.

Table 6.2: PSNR (dB) (higher is better) versus SNR on CIFAR-10 over fading channel for different code rates (R_c).

SNR	VQ-VAE			Proposed			BPG			JSCC
	$R_c=0.50$	$R_c=0.67$	$R_c=0.75$	$R_c=0.50$	$R_c=0.67$	$R_c=0.75$	$R_c=0.50$	$R_c=0.67$	$R_c=0.75$	
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	23.50
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	25.19
3	0.00	0.00	0.00	0.00	0.00	0.00	11.52	0.00	0.00	26.55
4	0.00	0.00	0.00	0.00	0.00	0.00	15.21	7.41	0.00	27.47
5	21.76	0.00	0.00	35.38	0.00	0.00	19.38	13.48	5.77	28.26
6	22.32	0.00	0.00	36.77	0.00	0.00	19.38	18.92	11.69	28.91
7	22.32	18.91	0.00	36.77	33.81	0.00	19.38	19.38	16.61	29.49
8	22.32	22.32	0.00	36.77	36.77	4.14	19.38	19.38	19.38	30.03
9	22.32	22.32	21.19	36.77	36.77	32.64	19.38	19.38	19.38	30.55
10	22.32	22.32	22.32	36.77	36.77	36.66	19.38	19.38	19.38	31.05

The proposed model shows better structural integrity than other methods, as shown in table 6.3. Generative architecture captures global and local features, enabling better reconstructions when channel conditions improve. This is advantageous over traditional methods like BPG, which fail to maintain structural integrity at lower SNRs. This efficient usage of discrete codebook vectors has enabled the model to achieve a higher MS-SSIM than VQ-VAE.

Table 6.3: MS-SSIM (higher is better) versus SNR on CIFAR-10 over fading channel for different code rates (R_c).

SNR	VQ-VAE			Proposed			BPG			JSCC
	$R_c=0.50$	$R_c=0.67$	$R_c=0.75$	$R_c=0.50$	$R_c=0.67$	$R_c=0.75$	$R_c=0.50$	$R_c=0.67$	$R_c=0.75$	
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.828
2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.865
3	0.000	0.000	0.000	0.030	0.000	0.000	0.099	0.000	0.000	0.890
4	0.660	0.000	0.000	0.850	0.000	0.000	0.354	0.026	0.000	0.907
5	0.752	0.000	0.000	0.980	0.060	0.000	0.361	0.220	0.000	0.919
6	0.752	0.558	0.000	0.990	0.700	0.000	0.361	0.344	0.218	0.929
7	0.752	0.744	0.249	0.990	0.980	0.320	0.361	0.361	0.297	0.937
8	0.752	0.752	0.681	0.990	0.990	0.860	0.361	0.361	0.361	0.943
9	0.752	0.752	0.752	0.990	0.990	0.970	0.361	0.361	0.361	0.948
10	0.752	0.752	0.752	0.990	0.990	0.990	0.361	0.361	0.361	0.953

Despite the channel noise introduced, adversarial training and efficient codebook usage have achieved better human-perceived image fidelity in the transmitted images. Although JSCC achieves better performance below SNR of 5 dB, the proposed methodology achieves LPIPS value of 0.01 at SNR 6 dB, compared to 0.06 of JSCC as shown in table 6.4. Our method employs discrete quantization with LDPC coding, resulting in a “cliff effect”; a sharp drop in performance at low SNR. JSCC, being analog, gracefully degrades in the presence of noise. However, above 5 dB SNR, our model outperforms JSCC in PSNR and LPIPS due to better codebook utilization and more effective adversarial decoding.

Table 6.4: LPIPS (lower is better) versus SNR on CIFAR-10 over fading channel for different code rates (R_c).

SNR	VQ-VAE			Proposed			BPG			JSCC
	$R_c=0.50$	$R_c=0.67$	$R_c=0.75$	$R_c=0.50$	$R_c=0.67$	$R_c=0.75$	$R_c=0.50$	$R_c=0.67$	$R_c=0.75$	
1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.202
2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.167
3	1.000	1.000	1.000	1.000	1.000	1.000	0.596	1.000	1.000	0.139
4	0.314	1.000	1.000	0.120	1.000	1.000	0.462	0.632	1.000	0.119
5	0.282	1.000	1.000	0.010	1.000	1.000	0.460	0.524	1.000	0.103
6	0.281	0.371	1.000	0.010	0.260	1.000	0.460	0.472	0.558	0.091
7	0.281	0.286	0.413	0.010	0.030	0.390	0.460	0.460	0.496	0.081
8	0.281	0.281	0.339	0.010	0.010	0.140	0.460	0.460	0.460	0.072
9	0.281	0.281	0.283	0.010	0.010	0.010	0.460	0.460	0.460	0.065
10	0.281	0.281	0.281	0.010	0.010	0.010	0.460	0.460	0.460	0.059

6.3.5 Generalization and scalability to diverse datasets

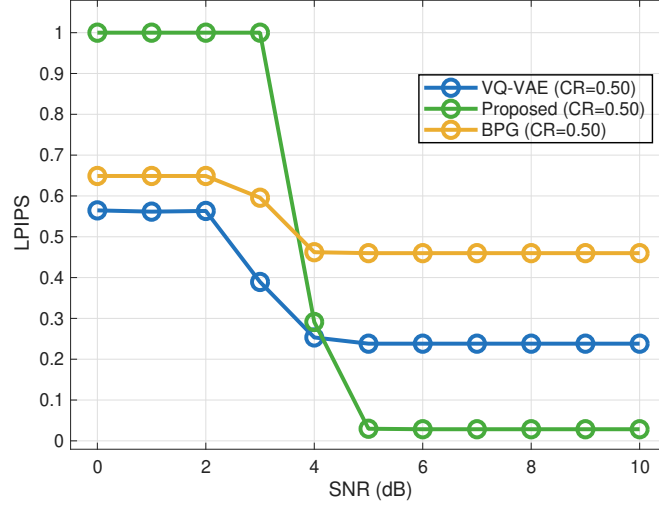


Figure 6.5: LPIPS performance on Tiny-ImageNet for image transmission over a fading channel at $R_c = 0.5$. The comparison includes BPG, VQ-VAE, and the proposed method, demonstrating generalization under a more diverse data distribution.

We evaluate the proposed model on Tiny-ImageNet (200 classes) [220] and inspect the samples at a channel code rate of 0.5. On representative ImageNet samples, the LPIPS vs. SNR curve in figure 6.5 exhibits the same trend observed on CIFAR-10, characterized by a “cliff effect” at very low SNR (≤ 4 dB). The model shows an advantage over VQ-VAE and BPG once the channel improves (≥ 5 dB). Qualitatively, ImageNet reconstructions in figure 6.6 preserve object identity and scene layout for SNRs above 4–5 dB, indicating that the learned codebook remains semantically stable on diverse content. In terms of computational complexity, the quantization step scales linearly with the number of latent tokens and the codebook size, $\mathcal{O}(abKd)$. The codebook adds only Kd parameters, enabling diverse datasets to be handled by controlling latent resolution ab and tuning K or d . These results suggest the method generalizes to larger and more varied distributions with predictable compute growth and consistent robustness trends.

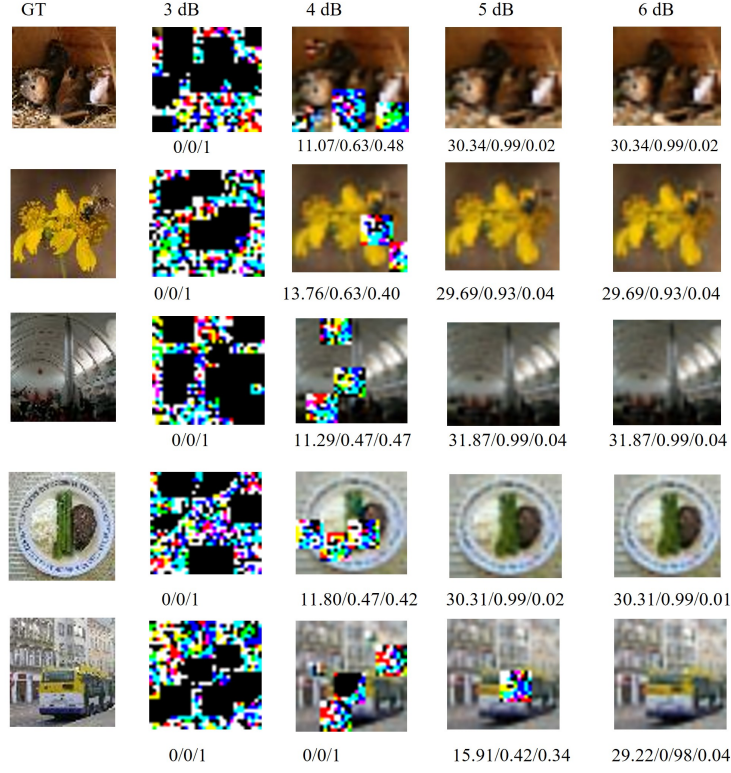


Figure 6.6: Visual quality of transmitted representative Imagenet samples with fading channel at code rate of 0.5. The quality metrics illustrated are (PSNR/MS-SSIM/LPIPS) at SNR = 3–6 dB, GT – ground truth image.

6.3.6 Computational cost and scalability

The complexity of the proposed model matches that of a VQ-VAE with the same (K, d) ; the use of soft assignments adds only a small constant factor relative to hard nearest-neighbor selection. In CNN-based JSCC without quantization, the encoder/decoder complexity grows approximately as $\mathcal{O}(ab\bar{c}^2L)$ with latent resolution ab , feature width $\bar{c}$, and depth L . The additional VQ computation is a single batched general matrix–matrix multiply that is typically negligible. In the implementation of the JSCC, the feature dimensions are larger and include more convolutional blocks, resulting in greater complexity than the proposed model. In contrast, transform codecs including BPG, employing DCT transformation, quantization, and entropy coding, incur the lowest complexity of $\mathcal{O}(n^2 \log n)$, where n^2 represents the spatial dimensions. The proposed

Chapter 6. Semantic noise modeling and evaluation through shared knowledge base optimization

design exhibits predictable, GPU-friendly scaling with higher-quality transmission. Compute scales linearly with ab and K , while memory scales with Kd ; scalability to larger datasets is achieved by controlling the latent grid and increasing K or d .

6.4 Limitations and future work

While this chapter introduces a conceptual framework to highlight the importance of optimizing the shared knowledge base for efficient SemCom design, the main limitation of the empirical validation is the use of a low-resolution dataset. Additionally, since this model is developed to explore the effect of codebook design on the overall performance of the SemCom system, as the central component that defines the semantic language, the experiments have not considered the codebook’s dynamic adaptability to achieve optimal system performance. Future work will focus on adaptive codebook design that adjusts size and structure online to data and channel conditions. Furthermore, improvements will focus on mitigating the low-SNR cliff through adaptive rate/power allocation and unequal protection for codewords, and tighter source–channel co-design that optimizes semantic fidelity directly for downstream tasks for application-specific SemCom systems. Extending the framework to richer modalities, including video, 3D video, and high-resolution datasets while preserving temporal complexity is another promising direction.

6.5 Conclusion

This chapter addressed the problem of modeling the semantic noise and designing optimal shared knowledge bases in SemCom, a relatively underexplored topic in the current literature. We introduce a novel soft-assignment based VQ. The loss terms, contrastive loss and entropy loss, jointly regularize quantization and the codebook, stabilized via EMA codebook updates, and trained with a WGAN-based critic. The communication pipeline is a practical physical-layer implementation using Sionna with LDPC coding, 4-QAM, and a flat Rayleigh fading channel. We modeled semantic noise relative to SemCom. A novel methodology is introduced that evaluates the model’s

Chapter 6. Semantic noise modeling and evaluation through shared knowledge base optimization

robustness to both pixel- and latent-space FGSM perturbations and assesses it on a downstream classification task. The latent-space FGSM evaluation highlighted improved semantic robustness over VQ-VAE, especially with larger codebooks, confirming the value of diversity-promoting regularization and stabilized codebook updates. The system achieved balanced codebook utilization of up to 96.88% for a codebook size of 64, with higher perplexity and improved perceptual quality on CIFAR-10, demonstrating generalization to Tiny-ImageNet. Clustering analyses further validated that learned codebook vectors align with the semantic classes of the data distribution, supporting the claim that the extracted semantic tokens can preserve representations in the presence of noise. Across fading channels, the proposed method outperformed VQ-VAE, JSCC, and BPG in terms of PSNR, MS-SSIM, and LPIPS, notably when the SNR exceeded 5 dB.

Chapter 7

Conclusion

This thesis presents a novel body of work that contributes to the development of communication architectures for wireless image transmission by conveying meaning rather than raw pixel data, aligning with emerging theories in SemCom. The research investigated whether generative AI-driven SemCom can enable robust, scalable, and meaning-centric communication architectures based on existing physical layer classical channel coding to transmit the extracted minimally sufficient semantic symbols. The thesis is built upon the central hypothesis that learned, discrete, and generative semantic representations can effectively convey the meaning. In contrast to continuous latents extracted by channel-coupled encoders, representation learning-based probabilistic deep learning techniques in SemCom can achieve superior robustness, efficiency, and generalization across diverse tasks and data distributions. The methodology proves the hypothesis through a sequence of progressively structured investigations and experiments, coupled with designed communication architectures based on evolving SemCom theories to validate it, revealing limitations and open challenges. The designed communication architectures are primarily grounded in the conceptual architecture motivated by the evolving SemCom theories.

First, the experimental results demonstrate that minimal semantic representations replace pixel-based image transmission to achieve bandwidth savings while maintaining sufficient semantic fidelity, validating hypothesis H1. The study in chapter 3 revealed that compact symbolic representation based on pixel boundaries, namely semantic

segmentation maps, can achieve compression gains of approximately $20\times$, maintaining object-level meaning under noisy channels where traditional codecs fails, specifically JPEG. This supports the hypothesis that semantic abstraction, in its simple symbolic form, coupled with generative decoders, can fundamentally support the semantic rate-distortion tradeoff in SemCom systems. However, the experiment also revealed that such symbolic representations are sensitive to boundary distortions caused by channel noise, motivating the need for more expressive and resilient semantic languages that can be shared between the transmitter and the receiver.

Based on the limitations identified in the initial SemCom design in chapter 3, the remainder of the experiments and communication architecture designs explored image transmission in wireless communication channels with VQ based implementation. As the first implementation in this research, a lightweight vector-quantized semantic communication model is introduced for edge scenarios, with an architecture that supports knowledge base and model weight sharing and inference to validate hypotheses H1, H2, and H3. It demonstrated computational efficiency suitable for resource-constrained platforms. The model performs image reconstruction and classification tasks at the receiver with two-step training optimization using quantized latent vectors, achieving 85.52% classification accuracy. The system demonstrated a 25% reduction in transmitted symbols, with LDPC at channel code rate 0.5 compared to JSCC, while achieving lower computational complexity and shorter inference latency of 1.2–1.6 ms for the proposed model vs. 10–12 ms for JSCC. This implementation validates the VQ-based index transmission strategy to achieve minimal representation and bandwidth efficiency, bridging theory and research with deployable semantic communication pipelines that align with the conceptual framework in 1.1.1, which guides the design decisions of this thesis and validates the hypothesis.

Third, we validate the hypotheses H1, H2, H3, and H4 in chapter 5, which is the main contribution of this thesis. It validates that vector-quantized latent representations, when combined with generative decoding and trained with contrastive regularization, can provide robustness to noise. Further, the system is stable during training and yields a generalizable semantic manifold compared to continuous latent representations

with encoders trained jointly with channel conditions and traditional image coding methods. The system reduces bandwidth utilization by transmitting indices of the relevant codebook vectors, resulting in a 56.2% reduction in symbol transmission load and low model complexity compared to JSCC. SVQ-WGAN demonstrated a PSNR gain of approximately 7 dB above 5 dB SNR levels, surpassing BPG across both channel conditions and comparable performance over JSCC. The design demonstrates LPIPS reductions close to 0.3 compared to 5.5 with BPG and MS-SSIM of approximately 0.8. In machine-perceived task performance, the model achieves approximately 94% in classification accuracy above 4 dB SNR and human-perceived task performance approximately 91% in image captioning accuracy above 4 dB SNR. Further, the model shows a strong robustness to noise with approximately 0.8 in RI. SVQ-WGAN demonstrates strong generalization with an FID of 5.496. These results validate the hypothesis that discrete semantic manifolds function as a shared semantic language, improving robustness to channel noise and enabling task-agnostic usage.

Further, the thesis studies the effect of semantic-based shared code books on efficient representation to mitigate semantic noise, which will arise in the semantic coding process, which is the quantization of the continuous latent representations to map to the codebook vectors. The effects were partially mitigated, but not fully resolved, through soft assignment and entropy regularization. Additionally, the experiments in chapters 4-6 demonstrate that decoupling semantic representation learning from physical-layer reliability is both feasible and advantageous. By transmitting the codebook indices that are protected by LDPC codes, the proposed models based on VQ achieve superior perpetual fidelity against JSCC, avoiding channel-specific retraining. These results strongly support the hypothesis that semantic robustness should be addressed at the representation level, while physical-layer errors are more effectively handled by mature channel coding schemes. However, the performance of the VQ based methods degrades in low SNR regions below approximately 3 dB, suggesting lightweight semantic error correction mechanisms at the decoder as a promising future direction.

One of the research objectives of this thesis is to evaluate SemCom beyond pixel-based matrices. The thesis presents novel evaluation methodologies that capture

perceptual quality, task relevance, robustness to noise, codebook effectiveness, and predictive sufficiency across image and video transmission. Task-oriented semantic retention is evaluated through classification and image captioning accuracy to resemble machine-perceived and human-perceived task fidelity. We introduce a novel metric to evaluate the channel noise robustness utilizing EP and SF. The semantic distortion measurement quantifies the boundary-based semantic fidelity in image transmission in SemCom by quantifying IoU distortion and category error, where the SVQ-WGAN shows superior performance over both metrics. For discrete semantic representations, codebook utilization, perplexity, and clustering consistency are introduced to analyze shared knowledge structure and semantic learning stability. An explicit evaluation is done using pixel-based and latent-based FGSM to quantify the semantic noise robustness of SemCom architecture.

These limitations delineate the scope within which the proposed frameworks are effective, and they can direct further research. Based on the validated hypotheses and insights gained from these limitations, the following section discusses promising directions for future research.

7.1 Future work

Extending upon the contributions and limitations identified relative to the research presented in the thesis, several promising research directions are identified towards advancing the SemCom for real-world deployment and complying with the theoretical maturity. Adaptive semantic codebooks are a promising novel direction that can evolve dynamically to changing data distributions, task objectives, and channel conditions. The codebook in this thesis is a static knowledge base learned offline and remains fixed during deployment. A practical problem is that the task requirements, channel conditions, and data contexts and domains are constantly changing in the real world. Meta-learning, as a principled solution, can enable VQ codebooks to rapidly adapt with feedback, without retraining the models. This direction can advance SemCom towards more autonomous and context-aware operation.

The proposed SemCom frameworks are based on unimodal visual representation,

mainly images and video. Often, most real-world applications involve modalities including audio, language, and sensor signals. A promising future direction is to develop multimodal semantic representation with multimodal feature fusion, where meaning emerges by combining different modalities. The representation alignment can be enhanced by using techniques including cross-modal attention, hierarchical vector quantization, or contrastively aligned multimodal codebooks. Open challenges remain in developing mechanisms that dynamically weight the modalities' importance based on CSI, task relevance, temporal dynamics, and uncertainty, presenting an important step towards developing multimodal SemCom systems.

This research has intentionally avoided the semantic level error correction mechanism at the decoder to keep the design lightweight. A key future direction is to develop semantic-level error-correction mechanisms that operate in the latent space rather than at the pixel level, similar to the median-filtering-based error-correction module used in chapter 3. Semantic error correction can be implemented using lightweight neural modules, including denoising autoencoders, gated residual correction blocks, or prototype-based latent refinement networks, which selectively repair corrupted semantic indices or embeddings. By using feedback mechanisms and CSI, these modules can be adaptive and jointly trained end-to-end. The parameters, including masking probabilities, correction weights, and confidence weights, can be adjusted based on the SNR conditions, where semantic error correction can be omitted at high SNR levels to minimize latency. Such CSI aware semantic error correction mechanisms can preserve semantic integrity, while maintaining the decoder's lower complexity.

Similarly, in VQ based SemCom CSI information can be incorporated at various levels in the pipeline. At the quantization step, CSI information can be used to dynamically select the codebook size or dimension, enabling the system to adapt to channel conditions by changing the semantic representation. Further, at the transmission phase, CSI can be used to implement unequal error protection by assigning stronger channel coding to semantically important indices. At the receiver, CSI-aware confidence scores can be assigned to received quantized vectors or indices to give more attention to or implement corrective mechanisms under degraded channel conditions. Most

importantly, these modular CSI integrations preserve the design of classical channel protection at the wireless channel and avoid the instability and retraining overhead associated with JSCC approaches. Future work will extend the proposed image-based semantic communication framework to video transmission, where temporal dependencies between frames introduce additional semantic dynamics and compression opportunities. Finally, standardizing semantic communication performance matrices is identified as an important direction to evaluate all SemCom systems against common baselines. In current implementations, SemCom systems are evaluated on datasets developed for different purposes relative to various tasks, specifically the COCO dataset. Developing benchmark datasets is an open direction for research, as are task-agnostic semantic similarity measures and predictive sufficiency metrics, which would allow fair comparison across semantic communication systems and accelerate scientific progress in the field.

Bibliography

- [1] M. U. Lokumarambage, V. S. S. Gowrisetty, H. Rezaei, T. Sivalingam, N. Rajatheva, and A. Fernando, “Wireless end-to-end image transmission system using semantic communications,” *IEEE Access*, vol. 11, pp. 37 149–37 163, 2023.
- [2] M. Lokumarambage, T. Sivalingam, F. Dong, N. Rajatheva, and A. Fernando, “Generative AI-based vector quantized end-to-end semantic communication system for wireless image transmission,” *IEEE Transactions on Machine Learning in Communications and Networking*, vol. 3, pp. 1050–1074, 2025.
- [3] C. Shannon, “A mathematical theory of communication,” *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [4] D. Wheeler and B. B. Natarajan, “Engineering semantic communication: A survey,” *IEEE Access*, vol. 11, pp. 13 965 – 13 995, 2023.
- [5] W. Yang, H. Du, Z. Liew, W. Y. B. Lim, Z. Xiong, D. T. Niyato, X. Chi, X. S. Shen, and C. Miao, “Semantic communications for future internet: Fundamentals, applications, and challenges,” *IEEE Communications Surveys and Tutorials*, vol. 25, no. 1, pp. 213 – 250, 2023.
- [6] C. Liang, H. Du, Y. Sun, D. Niyato, J. Kang, D. Zhao, and M. A. Imran, “Generative AI-driven semantic communication networks: Architecture, technologies, and applications,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 11, no. 1, pp. 27–47, 2025.
- [7] S. Ali, W. Saad, N. Rajatheva, K. Chang, D. Steinbach, B. Sliwa, C. Wietfeld, K. Mei, H. Shiri, H.-J. Zepernick, T. M. C. Chu, I. Ahmad, J. Huusko, J. Suutala,

Bibliography

- S. Bhadauria, V. Bhatia, R. Mitra, S. Amuru, R. Abbas, B. Shao, M. Capobianco, G. Yu, M. Claes, T. Karvonen, M. Chen, M. Girnyk, and H. Malik, “6G white paper on machine learning in wireless communication networks,” 2020. [Online]. Available: <https://arxiv.org/abs/2004.13875>
- [8] Y. Liu, X. Wang, Z. Ning, M. Zhou, L. Guo, and B. Jedari, “A survey on semantic communications: Technologies, solutions, applications and challenges,” *Digital Communications and Networks*, vol. 10, no. 3, pp. 528–545, 2024.
- [9] C. Chaccour, W. Saad, M. Debbah, Z. Han, and H. V. Poor, “Less data, more knowledge: Building next generation semantic communication networks,” *IEEE Communications Surveys & Tutorials*, pp. 1–1, 2024.
- [10] International Telecommunication Union Radiocommunication Sector (ITU-R), “IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond,” International Telecommunication Union, ITU-R Report M.2370-0, July 2015, accessed: 2022-06-02. [Online]. Available: https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2370-2015-PDF-E.pdf
- [11] V. Karthikeyan and K. Subbulakshmi, “Chapter 4 - millimeter wave and terahertz communication,” in *Human-Centric Integration of 6G-Enabled Technologies for Modern Society*, A. K. Tyagi and S. Tiwari, Eds. Academic Press, 2025, pp. 57–69. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B9780443274343000040>
- [12] A. Shafie, N. Yang, C. Han, J. M. Jornet, M. Juntti, and T. Kürner, “Terahertz communications for 6G and beyond wireless networks: Challenges, key advancements, and opportunities,” *IEEE Network*, vol. 37, no. 3, pp. 162–169, 2023.
- [13] Y. Karaçora, C. Chaccour, A. Sezgin, and W. Saad, “Event-based beam tracking with dynamic beamwidth adaptation in terahertz (THz) communications,” *IEEE Transactions on Communications*, vol. 71, no. 10, pp. 6195–6210, 2023.

Bibliography

- [14] Z. Qin, X. Tao, J. Lu, W. Tong, and G. Y. Li, “Semantic communications: Principles and challenges,” 2022. [Online]. Available: <https://arxiv.org/abs/2201.01389>
- [15] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, “Neural discrete representation learning,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6309–6318.
- [16] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Shen, and C. Miao, “Semantic communications for future internet: Fundamentals, applications, and challenges,” *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 213–250, 2023.
- [17] Z. Zhang, M. Li, and J. Yu, “On the convergence and mode collapse of GAN,” in *SIGGRAPH Asia 2018 Technical Briefs*, ser. SA ’18. New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: <https://doi.org/10.1145/3283254.3283282>
- [18] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde, “Optimal mass transport: Signal processing and machine-learning applications,” *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 43–59, 2017.
- [19] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, “Improved training of wasserstein GANs,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 5769–5779.
- [20] T. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. USA: Prentice Hall PTR, 2001.
- [21] N. Haverkamp and M. Hoeltje, “Semantic theories, linguistic essences, and knowledge of meaning,” *Synthese*, vol. 199, no. 5, pp. 14 459–14 490, 2021.

Bibliography

- [22] R. A. Howard, “Information value theory,” *IEEE Transactions on Systems Science and Cybernetics*, vol. 2, no. 1, pp. 22–26, 1966.
- [23] Y. Zhong, “A theory of semantic information,” *China Communications*, vol. 14, no. 1, pp. 1–17, 2017.
- [24] J. G. Proakis and M. Salehi, *Digital Communications*, 5th ed. New York: McGraw–Hill, 2008.
- [25] Q. Cao, Y. Shao, and S. Ge, “On the zero-error capacity of semantic channels with input and output memories,” *IEEE Wireless Communications Letters*, vol. 14, no. 3, pp. 896–900, 2025.
- [26] M. K. Farshbafan, W. Saad, and M. Debbah, “Common language for goal-oriented semantic communications: A curriculum learning framework,” in *ICC 2022 - IEEE International Conference on Communications*, 2022, pp. 1710–1715.
- [27] Y. Shao, Q. Cao, and D. Gündüz, “A theory of semantic communication,” *IEEE Transactions on Mobile Computing*, vol. 23, no. 12, pp. 12 211–12 228, 2024.
- [28] J. Feist, *Significance in Language: A Theory of Semantics*, 1st ed. Routledge, 2022. [Online]. Available: <https://doi.org/10.4324/9781003259381>
- [29] Q. Hu, G. Zhang, Z. Qin, Y. Cai, G. Yu, and G. Y. Li, “Robust Semantic Communications with Masked VQ-VAE Enabled Codebook,” *IEEE Transactions on Wireless Communications*, vol. 22, no. 12, pp. 8707–8722, 2023.
- [30] Z. Xu, B. Sun, X. Luo, and X. Liu, “VQ-VAE with Improved Codebook Learning,” in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 1280–1284.
- [31] Y. Zhou, Y. Sun, G. Chen, X. Xu, H. Chen, B. Huang, S. Cui, and P. Zhang, “MOC-RVQ: Multilevel Codebook-Assisted Digital Generative Semantic Communication,” in *GLOBECOM 2024 - 2024 IEEE Global Communications Conference*, 2024, pp. 2078–2083.

Bibliography

- [32] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, “The roadmap to 6G: AI empowered wireless networks,” *IEEE Communications Magazine*, vol. 57, no. 8, pp. 84–90, 2019.
- [33] S. Iyer, R. Khanai, D. Torse, R. J. Pandya, K. M. Rabie, K. Pai, W. U. Khan, and Z. Fadlullah, “A survey on semantic communications for intelligent wireless networks,” *Wireless Personal Communications*, vol. 129, no. 1, pp. 569–611, 2023.
- [34] Q. Lan, D. Wen, Z. Zhang, Q. Zeng, X. Chen, P. Popovski, and K. Huang, “What is semantic communication? a view on conveying meaning in the era of machine intelligence,” *Journal of Communications and Information Networks*, vol. 6, no. 4, pp. 336 – 371, 2021.
- [35] X. Luo, H. H. Chen, and Q. Guo, “Semantic communications: Overview, open issues, and future research directions,” *IEEE Wireless Communications*, vol. 29, no. 1, pp. 210 – 219, 2022.
- [36] K. Niu, J. Dai, S. Yao, S. Wang, Z. Si, X. Qin, and P. Zhang, “A paradigm shift toward semantic communications,” *IEEE Communications Magazine*, vol. 60, no. 11, pp. 113 – 119, 2022.
- [37] Y. Wang, C. Mingzhe, T. Luo, W. M. Saad, D. T. Niyato, H. V. Vincent Poor, and S. R. Cui, “Performance optimization for semantic communications: An attention-based reinforcement learning approach,” *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 9, pp. 2598 – 2613, 2022.
- [38] S. Jiang, Y. Liu, Y. Zhang, P. Luo, K. Cao, J. Xiong, H. Zhao, and J. Wei, “Reliable semantic communication system enabled by knowledge graph,” *Entropy*, vol. 24, no. 6, 2022.
- [39] Q. Zhou, R. Li, Z. Zhao, Y. Xiao, and H. Zhang, “Adaptive bit rate control in semantic communication with incremental knowledge-based HARQ,” *IEEE Open Journal of the Communications Society*, vol. 3, pp. 1076 – 1089, 2022.

Bibliography

- [40] P. Yi, Y. Cao, X. Kang, and Y. C. Liang, “Deep learning-empowered semantic communication systems with a shared knowledge base,” *IEEE Transactions on Wireless Communications*, vol. 23, no. 6, pp. 6174 – 6187, 2024.
- [41] F. Zhou, Y. Li, M. Xu, L. Yuan, Q. Wu, R. Q. Hu, and N. Al-Dhahir, “Cognitive semantic communication systems driven by knowledge graph: Principle, implementation, and performance evaluation,” *IEEE Transactions on Communications*, vol. 72, no. 1, pp. 193 – 208, 2024.
- [42] B. Wang, R. Li, J. Zhu, Z. Zhao, and H. Zhang, “Knowledge enhanced semantic communication receiver,” *IEEE Communications Letters*, vol. 27, no. 7, pp. 1794 – 1798, 2023.
- [43] A. Li, X. Liu, G. Wang, and P. Zhang, “Domain knowledge driven semantic communication for image transmission over wireless channels,” *IEEE Wireless Communications Letters*, vol. 12, no. 1, pp. 55 – 59, 2023.
- [44] C. Liang, D. Li, Z. Lin, and H. Cao, “Selection-based image generation for semantic communication systems,” *IEEE Communications Letters*, vol. 28, no. 1, pp. 34 – 38, 2024.
- [45] F. Jiang, Y. Peng, L. Dong, K. Wang, K. Yang, C. Pan, and X. You, “Large AI model-based semantic communications,” *IEEE Wireless Communications*, vol. 31, no. 3, pp. 68 – 75, 2024.
- [46] F. Jiang, L. Dong, Y. Peng, K. Wang, K. Yang, C. Pan, and X. You, “Large AI model empowered multimodal semantic communications,” *IEEE Communications Magazine*, vol. 63, no. 1, pp. 76 – 82, 2025.
- [47] Z. Yang, C. Mingzhe, G. Li, Y. Yang, and Z. Zhang, “Secure semantic communications: Fundamentals and challenges,” *IEEE Network*, vol. 38, no. 6, pp. 513 – 520, 2024.

Bibliography

- [48] H. Peng, Z. Zhang, Y. Liu, Z. Su, T. H. Luan, and N. Cheng, “Semantic communication in non-terrestrial networks: A future-ready paradigm,” *IEEE Network*, vol. 38, no. 4, pp. 119 – 127, 2024.
- [49] G. Xin, F. Pingyi, and K. B. Letaief, “Semantic communication: A survey of its theoretical development,” *Entropy*, vol. 26, no. 2, 2024.
- [50] H. Zhou, Y. Deng, X. Liu, N. Pappas, and A. Nallanathan, “Goal-oriented semantic communications for 6G networks,” *IEEE Internet of Things Magazine*, vol. 7, no. 5, pp. 104 – 110, 2024.
- [51] T. M. Getu, G. Kaddoum, and M. Bennis, “A survey on goal-oriented semantic communication: Techniques, challenges, and future directions,” *IEEE Access*, vol. 12, pp. 51 223 – 51 274, 2024.
- [52] E. C. Calvanese Strinati and S. Barbarossa, “6G networks: Beyond shannon towards semantic and goal-oriented communications,” *Computer Networks*, vol. 190, 2021.
- [53] S. Iyer, R. Khanai, D. A. Torse, R. J. Pandya, K. M. Rabie, K. Pai, W. U. Khan, and Z. M. Fadlullah, “A survey on semantic communications for intelligent wireless networks,” *Wireless Personal Communications*, vol. 129, no. 1, pp. 569 – 611, 2023.
- [54] P. A. Stavrou and M. Kountouris, “The role of fidelity in goal-oriented semantic communication: A rate distortion approach,” *IEEE Transactions on Communications*, vol. 71, no. 7, pp. 3918 – 3931, 2023.
- [55] W. Xu, Z. Yang, D. W. K. Ng, M. Levorato, Y. C. Eldar, and M. Debbah, “Edge learning for 5G networks with distributed signal processing: Semantic communication, edge computing, and wireless sensing,” *IEEE Journal on Selected Topics in Signal Processing*, vol. 17, no. 1, pp. 9 – 39, 2023.
- [56] W. Wu, Y. Yang, Y. Deng, and A. Hamid Aghvami, “Goal-oriented semantic communications for robotic waypoint transmission: The value and age of

Bibliography

- information approach,” *IEEE Transactions on Wireless Communications*, vol. 23, no. 12, pp. 18 903–18 915, 2024.
- [57] A. Kosta, N. Pappas, A. Ephremides, and V. Angelakis, “Age and value of information: Non-linear age case,” in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 326–330.
- [58] S. Chen, E. Spyrakos-Papastavridis, Y. Jin, and Y. Deng, “Goal-oriented semantic communication for robotic arm reconstruction in digital twin,” in *GLOBECOM 2024 - 2024 IEEE Global Communications Conference*, 2024, pp. 4830–4835.
- [59] Y. Xu, N. Li, and Y. Deng, “Goal-oriented semantic communication for video transmission via optical flow-based autoencoder,” in *ICC 2025 - IEEE International Conference on Communications*, 2025, pp. 1972–1977.
- [60] N. Li and Y. Deng, “Goal-oriented semantic communication for wireless image transmission via stable diffusion,” in *ICC 2025 - IEEE International Conference on Communications*, 2025, pp. 3748–3753.
- [61] Z. Wang, Y. Deng, and A. Hamid Aghvami, “Goal-oriented semantic communications for avatar-centric augmented reality,” *IEEE Transactions on Communications*, vol. 72, no. 12, pp. 7982–7995, 2024.
- [62] J. Park, Y. Oh, S. Kim, and Y. Jeon, “Joint source-channel coding for channel-adaptive digital semantic communications,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 11, no. 1, pp. 75 – 89, 2025.
- [63] Z. Lyu, G. Zhu, J. Xu, B. Ai, and S. R. Cui, “Semantic communications for image recovery and classification via deep joint source and channel coding,” *IEEE Transactions on Wireless Communications*, vol. 23, no. 8, pp. 8388 – 8404, 2024.
- [64] G. Zhang, Q. Hu, Z. Qin, Y. Cai, G. Yu, and X. Tao, “A unified multi-task semantic communication system for multimodal data,” *IEEE Transactions on Communications*, vol. 72, no. 7, pp. 4101 – 4116, 2024.

Bibliography

- [65] C. Liu, C. Guo, Y. Yang, and N. Jiang, “Adaptable semantic compression and resource allocation for task-oriented communications,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 10, no. 3, pp. 769 – 782, 2024.
- [66] H. Zhang, H. Wang, Y. Li, K. Long, and A. Nallanathan, “DRL-driven dynamic resource allocation for task-oriented semantic communication,” *IEEE Transactions on Communications*, vol. 71, no. 7, pp. 3992 – 4004, 2023.
- [67] J. Kang, H. Du, Z. Li, Z. Xiong, S. Ma, D. T. Niyato, and Y. Li, “Personalized saliency in task-oriented semantic communications: Image transmission and performance analysis,” *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 186 – 201, 2023.
- [68] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, “Task-oriented multi-user semantic communications,” *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 9, pp. 2584 – 2597, 2022.
- [69] H. Xie, Z. Qin, and G. Y. L. Li, “Task-oriented multi-user semantic communications for VQA,” *IEEE Wireless Communications Letters*, vol. 11, no. 3, pp. 553 – 557, 2022.
- [70] E. Bourtsoulatzé, D. Burth Kurka, and D. Gündüz, “Deep joint source-channel coding for wireless image transmission,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567–579, 2019.
- [71] C. Lee, X. Hu, and H. Kim, “Deep joint source-channel coding with iterative source error correction,” in *The 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*. The 26th International Conference on Artificial Intelligence and Statistics, 2023.
- [72] H. Zhang and M. Tao, “SNR-EQ-JSCC: Joint source-channel coding with SNR-based embedding and query,” *IEEE Wireless Communications Letters*, vol. 14, no. 3, pp. 881 – 885, 2025.

Bibliography

- [73] P. Ren, J. Wang, Z. Wang, X. Hou, X. Zhang, and C. Jiang, “M-JSCC: An asymmetric semantic communication architecture for 6G intelligent networks,” in *2025 IEEE 101st Vehicular Technology Conference*, 2025, pp. 1–6.
- [74] A. D. Raha, A. Adhikary, M. Gain, Y. Park, W. Saad, and C. S. Hong, “DD-JSCC: Dynamic deep joint source-channel coding for semantic communications,” in *ICC 2025 - IEEE International Conference on Communications*, 2025, pp. 3754–3759.
- [75] H. Li, J. Wang, X. Li, and K. Zhu, “Joint semantic feature selection and resource allocation optimization via DIF-JSCC in IoVs,” in *IEEE INFOCOM 2025 - IEEE Conference on Computer Communications Workshops*, 2025, pp. 1–6.
- [76] S. Yang, B. Shen, X. Huang, and Q. Chen, “Prenet aided DF-JSCC: An optimized approach for wireless semantic image transmission over dynamic radio channels,” *IEEE Transactions on Vehicular Technology*, 2025.
- [77] J. Huang, K. Yuan, C. Huang, and K. Huang, “D²-JSCC: Digital deep joint source-channel coding for semantic communications,” *IEEE Journal on Selected Areas in Communications*, vol. 43, no. 4, pp. 1246 – 1261, 2025.
- [78] M. Nemati and J. Choi, “All-in-one: VQ-VAE for end-to-end joint source-channel coding,” 2022, preprint. [Online]. Available: <https://www.techrxiv.org/doi/full/10.36227/techrxiv.19294622.v1>
- [79] Q. Fu, H. Xie, Z. Qin, G. Slabaugh, and X. Tao, “Vector quantized semantic communication system,” *IEEE Wireless Communications Letters*, vol. 12, no. 6, pp. 982–986, 2023.
- [80] Y. Zhang, W. Liu, P. Li, X. Chen, and H. V. Poor, “Transformer-based semantic communication system for robust image transmission,” in *2024 IEEE International Conference on Communications (ICC)*, 2024, pp. 1–6.
- [81] W. Li, W. Zhang, X. Chen, W. Liu, and H. V. Poor, “Semantic communication with large language models for robust message understanding,” *IEEE Transactions on Wireless Communications*, vol. 23, no. 10, pp. 10 342–10 355, 2024.

Bibliography

- [82] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Commun. ACM*, vol. 63, no. 11, 2014.
- [83] J. Wu, C. Wu, Y. Lin, T. Yoshinaga, L. Zhong, X. Chen, and Y. Ji, “Semantic segmentation-based semantic communication system for image transmission,” *Digital Communications and Networks*, vol. 10, no. 3, pp. 519–527, 2024.
- [84] E. Erdemir, T.-Y. Tung, P. L. Dragotti, and D. Gündüz, “Generative joint source-channel coding for semantic image transmission,” *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 8, pp. 2645–2657, 2023.
- [85] D. Ye, X. Wang, and X. Chen, “Lightweight generative joint source-channel coding for semantic image transmission with compressed conditional GANs,” in *2023 IEEE/CIC International Conference on Communications in China (ICCC Workshops)*, 2023, pp. 1–6.
- [86] X. Chen, D. Feng, Q. He, Y. Sun, and X.-G. Xia, *Content-Aware Robust Semantic Transmission of Images over Wireless Channels with GANs*. Wiley, 2024, ch. 7. [Online]. Available: <https://doi.org/10.1002/9781394223336.ch7>
- [87] Q. He, H. Yuan, D. Feng, B. Che, Z. Chen, and X.-G. Xia, “Robust semantic transmission of images with generative adversarial networks,” in *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, 2022, pp. 3953–3958.
- [88] M. Letafati, S. Ali, and M. Latva-Aho, “Conditional denoising diffusion probabilistic models for data reconstruction enhancement in wireless communications,” *IEEE Transactions on Machine Learning in Communications and Networking*, vol. 3, pp. 133–146, 2025.
- [89] Y. Miao, J. Yan, Y. Wang, Z. Li, and D. Hu, “A semantic communication system based on vector quantization and generative model,” in *2024 6th International Conference on Communications, Information System and Computer Engineering (CISCE)*, 2024, pp. 46–50.

Bibliography

- [90] Q. Hu, G. Zhang, Z. Qin, Y. Cai, G. Yu, and G. Y. Li, “Robust semantic communications with masked VQ-VAE enabled codebook,” *IEEE Transactions on Wireless Communications*, vol. 22, no. 12, pp. 8707–8722, 2023.
- [91] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
- [92] W. L. Hamilton, R. Ying, and J. Leskovec, “Representation learning on graphs: Methods and applications,” 2018. [Online]. Available: <https://arxiv.org/abs/1709.05584>
- [93] G. Hinton, “How to represent part-whole hierarchies in a neural network,” Feb. 2021, arXiv:2102.12627 [cs]. [Online]. Available: <http://arxiv.org/abs/2102.12627>
- [94] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” 2022. [Online]. Available: <https://arxiv.org/abs/1312.6114>
- [95] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, “beta-VAE: Learning basic visual concepts with a constrained variational framework,” in *International Conference on Learning Representations (ICLR)*, 2017.
- [96] M. Mehralian and B. Karasfi, “RDCGAN: Unsupervised representation learning with regularized deep convolutional generative adversarial networks,” in *2018 9th Conference on Artificial Intelligence and Robotics and 2nd Asia-Pacific International Symposium*, 2018, pp. 31–38.
- [97] M. Sana and E. C. Strinati, “Learning semantics: An opportunity for effective 6G communications,” in *2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC)*, 2022, pp. 631–636.
- [98] C. K. Thomas, C. Chaccour, W. Saad, M. Debbah, and C. S. Hong, “Causal reasoning: Charting a revolutionary course for next-generation AI-native wireless networks,” *IEEE Vehicular Technology Magazine*, vol. 19, no. 1, pp. 16–31, 2024.

Bibliography

- [99] C. K. Thomas, C. Chaccour, W. Saad, and M. Debbah, “Causal reasoning: Foundations of semantic communication systems,” in *Foundations of Semantic Communication Networks*, W. Saad, C. Chaccour, C. K. Thomas, and M. Debbah, Eds. Wiley-IEEE Press, 2025, ch. 11, pp. 217–240.
- [100] S. H. Wanasekara, V. D. Nguyen, K. S. Wong, M. D. Nguyen, S. Chatzinotas, and O. A. Dobre, “SC-GIR: Goal-oriented semantic communication via invariant representation learning for image transmission,” *IEEE Transactions on Mobile Computing*, 2025.
- [101] W. Hua, L. Xiong, S. Liu, L. Chen, X. M. Hong, J. F. C. Mota, and X. Cheng, “Classification-driven discrete neural representation learning for semantic communications,” *IEEE Internet of Things Journal*, vol. 11, no. 9, pp. 16 061 – 16 073, 2024.
- [102] D. Chen and W. Hua, “Hierarchical VAE based semantic communications for POMDP tasks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 5540–5544.
- [103] N. Hello, P. Di Lorenzo, and E. C. Strinati, “Semantic communication enhanced by knowledge graph representation learning,” in *2024 IEEE 25th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2024, pp. 876–880.
- [104] G. Chen, G. Nan, Z. Jiang, H. Du, R. Shi, Q. Cui, and X. Tao, “Lightweight and robust wireless semantic communications,” *IEEE Communications Letters*, vol. 28, no. 11, pp. 2633 – 2637, 2024.
- [105] J. Ding, Y. Zhang, Y. Shang, Y. Zhang, Z. Zong, J. Feng, Y. Yuan, H. Su, N. Li, N. Sukiennik, F. Xu, and Y. Li, “Understanding world or predicting future? a comprehensive survey of world models,” *ACM Comput. Surv.*, vol. 58, no. 3, Sep. 2025.

Bibliography

- [106] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.
- [107] A. Wang, H. Chen, Z. Lin, J. Han, and G. Ding, “RepViT: Revisiting mobile CNN from ViT perspective,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 15 909–15 920.
- [108] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” 2021. [Online]. Available: <https://arxiv.org/abs/2010.11929>
- [109] D. Ha and J. Schmidhuber, “World models,” *arXiv preprint arXiv:1803.10122*, 2018. [Online]. Available: <https://arxiv.org/abs/1803.10122>
- [110] —, “Recurrent world models facilitate policy evolution,” in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018.
- [111] Y. Seo, J. Kim, S. James, K. Lee, J. Shin, and P. Abbeel, “Multi-view masked world models for visual robotic manipulation,” in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML’23. JMLR.org, 2023.
- [112] X. Long, Q. Zhao, K. Zhang, Z. Zhang, D. Wang, Y. Liu, Z. Shu, Y. Lu, S. Wang, X. Wei, W. Li, W. Yin, Y. Yao, J. Pan, Q. Shen, R. Yang, X. Cao, and Q. Dai, “A survey: Learning embodied intelligence from physical simulators and world models,” 2025. [Online]. Available: <https://arxiv.org/abs/2507.00917>
- [113] D. Wheeler and B. Natarajan, “Conceptual learning and causal reasoning for semantic communication,” *IEEE Transactions on Cognitive Communications and Networking*, pp. 1–1, 2025.

Bibliography

- [114] A. M. Girgis, A. Valcarce, and M. Bennis, “Time-series JEPA for predictive remote control under capacity-limited networks,” 2025. [Online]. Available: <https://arxiv.org/abs/2406.04853>
- [115] H. H. Kim, S. Yu, S. Yuan, and C. Tomasi, “Cross-attention transformer for video interpolation,” in *Proceedings of the Asian Conference on Computer Vision (ACCV) Workshops*, 2022, pp. 320–337.
- [116] Y. Cai, “Robust and adaptive semantic noise for complex secure communication networks,” *Physical Communication*, vol. 72, 2025.
- [117] L. Hu, Y. Li, H. Zhang, L. Yuan, F. Zhou, and Q. Wu, “Robust semantic communication driven by knowledge graph,” in *2022 9th International Conference on Internet of Things: Systems, Management and Security (IOTSMS)*, 2022, pp. 1–5.
- [118] X. Lai, X. Li, S. Yang, and B. Shen, “Robust text semantic communication system assisted by knowledge graph,” *Beijing Youdian Xueyuan Xuebao/Journal of Beijing University of Posts And Telecommunications*, vol. 48, no. 4, pp. 97 – 116, 2025.
- [119] X. Ba, X. Zhang, S. Li, J. Yuan, and J. Hu, “Multimodal semantic communication system based on graph neural networks,” *Intelligence and Robotics*, vol. 5, no. 3, pp. 805 – 826, 2025.
- [120] H. Wei, W. Ni, W. Xu, W. Jiang, D. T. Niyato, and P. Zhang, “Semantic communications for multi-view generation,” *IEEE Communications Letters*, vol. 28, no. 6, pp. 1308 – 1312, 2024.
- [121] T. M. Getu, W. M. Saad, G. Kaddoum, and M. Bennis, “Performance limits of a deep learning-enabled text semantic communication under interference,” *IEEE Transactions on Wireless Communications*, vol. 23, no. 8, pp. 10 213 – 10 228, 2024.

Bibliography

- [122] S. Choi, J. Oh, T. S. Do, S. Hong, J. Kim, and S. Cho, “A survey on noise detection and correction in semantic communication,” in *2025 Sixteenth International Conference on Ubiquitous and Future Networks (ICUFN)*, 2025, pp. 23–25.
- [123] P. Samarathunga, T. Fernando, V. S. S. Gowrisetty, T. Atulugama, and A. Fernando, “Autoencoder based image quality metric for modelling semantic noise in semantic communications,” *Electronics Letters*, vol. 60, no. 4, 2024.
- [124] H. Niu, L. Wang, Z. Lu, K. Du, and X. Wen, “Deep learning enabled video semantic transmission against multi-dimensional noise,” in *2023 IEEE Globecom Workshops*, 2023, pp. 1267–1272.
- [125] X. Li, X. Zhang, Q. Cui, and X. Tao, “Robust semantic communication systems based on image transmission,” in *2023 22nd International Symposium on Communications and Information Technologies (ISCIT)*, 2023, pp. 419–423.
- [126] W. Hu, C. Guo, Z. Zhu, Y. Yang, and C. Feng, “Task-oriented semantic communication with large language model enabled knowledge base,” in *ICC 2025 - IEEE International Conference on Communications*, 2025, pp. 6279–6284.
- [127] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.
- [128] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4217–4228, 2021.
- [129] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network ,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jul. 2017, pp. 105–114.

Bibliography

- [130] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, “ESRGAN: Enhanced super-resolution generative adversarial networks,” in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019, pp. 63–79.
- [131] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976.
- [132] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.
- [133] H. Shi, L. Wang, G. Ding, F. Yang, and X. Li, “Data augmentation with improved generative adversarial networks,” in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 73–78.
- [134] Y. Jiménez-Gaona, D. Carrión-Figueroa, V. Lakshminarayanan, and M. José Rodríguez-Álvarez, “GAN-based data augmentation to improve breast ultrasound and mammography mass classification,” *Biomedical Signal Processing and Control*, vol. 94, p. 106255, 2024.
- [135] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” 2019. [Online]. Available: <https://arxiv.org/abs/1807.03748>
- [136] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proceedings of the 37th International Conference on Machine Learning*, ser. ICML’20. JMLR.org, 2020, pp. 1597–1607.
- [137] X. Chen, S. Xie, and K. He, “An empirical study of training self-supervised vision transformers,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9620–9629.

Bibliography

- [138] C. Chaccour and W. M. Saad, “Disentangling learnable and memorizable data via contrastive learning for semantic communications,” *Conference Record of the Asilomar Conference on Signals, Systems and Computers*, vol. 2022-October, pp. 1175 – 1179, 2022.
- [139] S. Tang, Q. Yang, L. Fan, X. Lei, Y. Deng, and A. Nallanathan, “Contrastive learning based semantic communication for wireless image transmission,” in *2023 IEEE 98th Vehicular Technology Conference*, 2023, pp. 1–6.
- [140] S. Tang, Q. Yang, L. Fan, X. F. Lei, A. Nallanathan, and G. K. Karagiannidis, “Contrastive learning based semantic communications,” *IEEE Transactions on Communications*, pp. 1 – 1, 2024.
- [141] Y. Wang, W. Ni, W. Yi, X. Xu, P. Zhang, and A. Nallanathan, “Federated contrastive learning for personalized semantic communication,” *IEEE Communications Letters*, vol. 28, no. 8, pp. 1875 – 1879, 2024.
- [142] X. Niu, D. Li, J. Zhao, L. Tan, and R. Bai, “ESCS: An expandable semantic communication system for multimodal data based on contrastive learning,” *IEEE Communications Letters*, vol. 29, no. 2, pp. 368 – 372, 2025.
- [143] T. Zhang, F. Wu, Z. Tian, C. Zhang, and S. Yu, “MOUSSE: A multimodality-oriented unified semantic communication system by contrastive learning,” in *2025 IEEE Wireless Communications and Networking Conference (WCNC)*, 2025, pp. 1–6.
- [144] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” 2022. [Online]. Available: <https://arxiv.org/abs/1312.6114>
- [145] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, “Neural discrete representation learning,” 2018. [Online]. Available: <https://arxiv.org/abs/1711.00937>
- [146] A. Razavi, A. van den Oord, and O. Vinyals, *Generating diverse high-fidelity images with VQ-VAE-2*. Red Hook, NY, USA: Curran Associates Inc., 2019.

Bibliography

- [147] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, “Speech resynthesis from discrete disentangled self-supervised representations,” 2021. [Online]. Available: <https://arxiv.org/abs/2104.00355>
- [148] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, “Soundstream: An end-to-end neural audio codec,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2022.
- [149] P. Esser, R. Rombach, and B. Ommer, “Taming transformers for high-resolution image synthesis,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 873–12 883.
- [150] Z. Jin, Y. Li, T. Song, W.-K. Jia, and X. Song, “SVQ-VAE: Federated-learning-based semantic-aware communication for vehicular networks,” *IEEE Internet of Things Journal*, vol. 12, no. 23, pp. 51 289–51 304, 2025.
- [151] M. Nemati, J. Park, and J. Choi, “VQ-VAE empowered wireless communication for joint source-channel coding and beyond,” in *GLOBECOM 2023 - 2023 IEEE Global Communications Conference*, 2023, pp. 3155–3160.
- [152] Y. Cai, “Robust and adaptive semantic noise for complex secure communication networks,” *Physical Communication*, vol. 72, p. 102763, 2025.
- [153] Y.-C. Chao, Y. Chen, W. Wang, A. Wijesinghe, S. Wanninayaka, S. Zhang, and Z. Ding, “Task-driven semantic quantization and imitation learning for goal-oriented communications,” in *ICC 2025 - IEEE International Conference on Communications*, 2025, pp. 1–6.
- [154] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.
- [155] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby,

Bibliography

- “An image is worth 16×16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations (ICLR)*, 2021.
- [156] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, 2020, p. 213–229.
- [157] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, “ViViT: A video vision transformer,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 6836–6846.
- [158] M. Assran, A. Bardes, D. Fan, Q. Garrido, R. Howes, Mojtaba, Komeili, M. Muckley, A. Rizvi, C. Roberts, K. Sinha, A. Zholus, S. Arnaud, A. Gejji, A. Martin, F. R. Hogan, D. Dugas, P. Bojanowski, V. Khalidov, P. Labatut, F. Massa, M. Szafraniec, K. Krishnakumar, Y. Li, X. Ma, S. Chandar, F. Meier, Y. LeCun, M. Rabbat, and N. Ballas, “V-JEPA 2: Self-supervised video models enable understanding, prediction and planning,” 2025.
- [159] K. Drozdov, R. Shwartz-Ziv, and Y. LeCun, “Video representation learning with joint-embedding predictive architectures,” 2024. [Online]. Available: <https://arxiv.org/abs/2412.10925>
- [160] H. Seo, J. Park, M. Bennis, and M. Debbah, “Semantics-native communication via contextual reasoning,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 9, no. 3, pp. 604–617, 2023.
- [161] T. M. Getu, G. Kaddoum, and M. Bennis, “A survey on goal-oriented semantic communication: Techniques, challenges, and future directions,” *IEEE Access*, vol. 12, pp. 51 223–51 274, 2024.
- [162] G. Yan, K. Liu, C. Liu, and J. Zhang, “Edge intelligence for internet of vehicles: A survey,” *IEEE Transactions on Consumer Electronics*, vol. 70, no. 2, pp. 4858–4877, 2024.

Bibliography

- [163] Z. Wan, S. Liu, Z. Xu, W. Ni, Z. Chen, and F. Wang, “Semantic communication method based on compression ratio optimization for vision tasks in the artificial intelligence of things,” *IEEE Transactions on Consumer Electronics*, vol. 70, no. 2, pp. 4934–4944, 2024.
- [164] Z. Qin, J. Ying, D. Yang, H. Wang, and X. Tao, “Computing networks enabled semantic communications,” *IEEE Network*, vol. 38, no. 2, pp. 122–131, 2024.
- [165] Y. Wu, S. Tang, L. Zhang, L. Fan, X. Lei, and X. Chen, “Resilient machine learning-based semantic-aware MEC networks for sustainable next-G consumer electronics,” *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 2188–2199, 2024.
- [166] G. Zhang, Q. Hu, Z. Qin, Y. Cai, G. Yu, and X. Tao, “A unified multi-task semantic communication system for multimodal data,” *IEEE Transactions on Communications*, vol. 72, no. 7, pp. 4101–4116, 2024.
- [167] Z. Hu, C. You, T. Liu, D. Wen, Y. Hu, Y. Cui, Y. Gong, and K. Huang, “Semantic communication meets edge intelligence: Semantic-relay-aided text transmissions,” *IEEE Internet of Things Journal*, vol. 11, no. 24, pp. 39 409–39 423, 2024.
- [168] X. Xu, B. Xu, S. Han, C. Dong, H. Xiong, R. Meng, and P. Zhang, “Task-oriented and semantic-aware heterogeneous networks for artificial intelligence of things: Performance analysis and optimization,” *IEEE Internet of Things Journal*, vol. 11, no. 1, pp. 228–242, 2024.
- [169] W. F. Lo, N. Mital, H. Wu, and D. Gündüz, “Collaborative semantic communication for edge inference,” *IEEE Wireless Communications Letters*, vol. 12, no. 7, pp. 1125–1129, 2023.
- [170] J. Im, N. Kwon, T. Park, J. Woo, J. Lee, and Y. Kim, “Attention-aware semantic communications for collaborative inference,” *IEEE Internet of Things Journal*, vol. 11, no. 22, pp. 37 008–37 020, 2024.

Bibliography

- [171] H. Lin, T. Nagarajan, N. Ballas, M. Assran, M. Komeili, M. Bansal, and K. Sinha, “VEDIT: Latent prediction architecture for procedural video representation learning,” 2024. [Online]. Available: <https://arxiv.org/abs/2410.03478>
- [172] H. Lin, A. Zala, J. Cho, and M. Bansal, “Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning,” in *Proceedings of the Conference on Language, Vision, and Multimodal (COLM)*, 2024.
- [173] G. Shrivastava and A. Shrivastava, “Video prediction by modeling videos as continuous multi-dimensional processes,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, p. 7236–7245.
- [174] M. A. Yilmaz, A. Bilican, and A. M. Tekalp, “FG-DFPN: Flow guided deformable frame prediction network,” *arXiv preprint arXiv:2503.11343*, 2025. [Online]. Available: <https://arXiv.org/abs/2503.11343>
- [175] W. Yuan, Y. He, W. Shen, Y. Dong, X. Gu, Z. Dong, L. Bo, and Q. Huang, “MogenTS: motion generation based on spatial-temporal joint modeling,” in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, ser. NIPS ’24. Red Hook, NY, USA: Curran Associates Inc., 2025.
- [176] C. V. Aravinda, T. Al-Shehari, N. A. Alsadhan, S. Shetty, G. Padmajadevi, and K. R. U. K. Reddy, “A novel hybrid architecture for video frame prediction: combining convolutional LSTM and 3D CNN,” *J. Real-Time Image Process.*, vol. 22, no. 1, Jan. 2025. [Online]. Available: <https://doi.org/10.1007/s11554-025-01626-w>
- [177] S. Ge, T. Hayes, H. Yang, X. Yin, G. Pang, D. Jacobs, J.-B. Huang, and D. Parikh, “Long video generation with time-agnostic vqgan and time-sensitive transformer,” in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 102–118.

Bibliography

- [178] D. Gunduz, Y. Xiao, K. Yang, Z. Yang, and H. V. Poor, “Beyond transmitting bits: Context, semantics, and task-oriented communications,” *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 5–36, 2022.
- [179] T.-Y. Tung and D. Gündüz, “DeepWiVe: Deep-learning-aided wireless video transmission,” *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 9, pp. 2570–2583, 2022.
- [180] S. Wang, J. Dai, Z. Liang, K. Niu, Z. Si, C. Dong, X. Qin, and P. Zhang, “Wireless deep video semantic transmission,” *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 214–229, 2023.
- [181] H. Li, H. Tong, S. Wang, N. Yang, Z. Yang, and C. Yin, “Video semantic communication with major object extraction and contextual video encoding,” in *2024 IEEE Wireless Communications and Networking Conference (WCNC)*, 2024, pp. 1–6.
- [182] M. Karaa, R. Bahria, H. Ghazzai, and L. Sboui, “A semantic compression framework for video surveillance applications,” in *2024 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, 2024, pp. 133–134.
- [183] H. Niu, L. Wang, Z. Lu, K. Du, and X. Wen, “Deep learning enabled video semantic transmission against multi-dimensional noise,” in *2023 IEEE Globecom Workshops (GC Wkshps)*, 2023, pp. 1267–1272.
- [184] G. Iglesias, E. Talavera, and A. Díaz-Álvarez, “A survey on GANs for computer vision: Recent research, analysis and taxonomy,” *Computer Science Review*, vol. 48, p. 100553, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1574013723000205>
- [185] Z. Xing, Q. Feng, H. Chen, Q. Dai, H. Hu, H. Xu, Z. Wu, and Y.-G. Jiang, “A survey on video diffusion models,” *ACM Comput. Surv.*, vol. 57, no. 2, Nov. 2024. [Online]. Available: <https://doi.org/10.1145/3696415>

Bibliography

- [186] N. Li, Y. Deng, and D. Niyato, “Goal-oriented semantic communication for wireless video transmission via generative AI,” *arXiv preprint arXiv:2502.20927*, 2025. [Online]. Available: <https://Arxiv.org/abs/2502.20927>
- [187] H. Yin, L. Qiao, Y. Ma, S. Sun, K. Li, Z. Gao, and D. Niyato, “Generative video semantic communication via multimodal semantic fusion with large model,” *arXiv preprint arXiv:2502.13838*, 2025.
- [188] Y. Peng, L. Xiang, K. Yang, K. Wang, and M. Debbah, “Semantic communications with computer vision sensing for edge video transmission,” 2025. [Online]. Available: <https://arxiv.org/abs/2503.07252>
- [189] J. Guo, W. Chen, Y. Sun, J. Xu, and B. Ai, “VideoQA-SC: Adaptive semantic communication for video question answering,” *IEEE Journal on Selected Areas in Communications*, vol. 43, no. 7, pp. 2462–2477, 2025.
- [190] Y. Wang, H. Han, Y. Feng, J. Zheng, and B. Zhang, “Semantic communication empowered 6G networks: Techniques, applications, and challenges,” *IEEE Access*, vol. 13, pp. 28 293 – 28 314, 2025.
- [191] Y. Yang, J. Zhou, Z. Yang, and M. R. Shikh-Bahaei, “Fluid antenna-enabled near-field integrated sensing, computing, and semantic communication for emerging applications,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 11, no. 5, pp. 3062 – 3078, 2025.
- [192] O. Günlü, “Randomized distributed function computation with semantic communications: Applications to privacy,” in *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2024, pp. 1–6.
- [193] J. Chen, J. Wang, C. Jiang, and J. Wang, “Age of incorrect information in semantic communications for NOMA aided XR applications,” *IEEE Journal on Selected Topics in Signal Processing*, vol. 17, no. 5, pp. 1093 – 1105, 2023.
- [194] W. Yu and J. Zhao, “Semantic communications, semantic edge computing, and semantic caching with applications to the metaverse and 6G mobile networks,”

Bibliography

- in *2023 IEEE 43rd International Conference on Distributed Computing Systems (ICDCS)*, 2023, pp. 983–984.
- [195] A. Bhattacharyya, A. Sau, and S. K. Mahato, “Semantic communication with networked AI in 6G: Application scenarios, insights from practical experiments, standardization aspects,” *International Conference on Communication Systems and Networks, COMSNETS*, no. 2025, pp. 1264 – 1268, 2025.
- [196] A. Cavagna, N. Li, A. Iosifidis, and Q. Zhang, “Semantic communication enabling robust edge intelligence for time-critical IoT applications,” in *2023 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2023, pp. 1617–1622.
- [197] M. Li, S. Li, K. Jin, Y. Li, and J. Tang, “Semantic communication-aided computation offloading for latency-aware applications in edge computing networks,” in *2025 IEEE/CIC International Conference on Communications in China (ICCC)*, 2025, pp. 1–6.
- [198] A. Athama, K. Wang, X. Chen, and Y. Li, “Semantic communications for healthcare applications: Opportunities and challenges,” in *2024 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT)*, 2024, pp. 383–388.
- [199] Y. Xiao, X. Fan, Z. Xie, and Y. Lu, “Offline knowledge base and attention-driven semantic communication for image-based applications in ITS scenarios,” *Big Data and Cognitive Computing*, vol. 9, no. 9, 2025.
- [200] H. Wang, Z. Li, X. Zhao, Z. Chang, and F. Hu, “A universal speech semantic communication framework for multi-task applications based on unsupervised models,” *IEEE Internet of Things Journal*, 2025.
- [201] I. A. Ivanova and G. Rzadkowski, “Mechanisms of semantic communication of information and dynamics of complex systems. application to financial markets,” *Chaos, Solitons and Fractals*, vol. 200, 2025.

Bibliography

- [202] S. Sharma and H. Verma, “A vision on extensions of canonical correlation analysis for data interpretation applications in compatible semantic communication,” *International Journal of Sensors, Wireless Communications and Control*, vol. 11, no. 6, pp. 696 – 699, 2021.
- [203] H. Rezaei, N. Rajatheva, and M. Latva-Aho, “High-throughput rate-flexible combinational decoders for multi-kernel polar codes,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 70, no. 11, pp. 4492–4504, 2023.
- [204] —, “Low-latency multi-kernel polar decoders,” *IEEE Access*, vol. 10, pp. 119 460–119 474, 2022.
- [205] H. Rezaei, E. Abbasi, N. Rajatheva, and M. Latva-Aho, “Unrolled, pipelined, and stage-folded architectures for encoding of multi-kernel polar codes,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 32, no. 11, pp. 2107–2120, 2024.
- [206] R. G. Gallager, “Low-density parity-check codes,” *IRE Transactions on Information Theory*, vol. 8, no. 1, pp. 21–28, 1962.
- [207] N. Corporation, *Sionna 1.1.0 Documentation*, 2025, accessed: 2025-03-11. [Online]. Available: <https://nvlabs.github.io/sionna/api/channel.html>
- [208] J. Hoydis, S. Cammerer, F. A. Aoudia, A. Vem, N. Binder, G. Marcus, and A. Keller, “Sionna: An open-source library for next-generation physical layer research,” 2023. [Online]. Available: <https://arxiv.org/abs/2203.11854>
- [209] J.-H. Lee and J. Kim, “Sionna: Introduction to embedded open-source semantic communication platforms,” in *2023 International Conference on Information Networking (ICOIN)*, 2023, pp. 775–777.
- [210] NVIDIA Corporation, “Forward error correction (fec)- sionna phy api,” <https://nvlabs.github.io/sionna/phy/api/fec.html>, 2025, accessed: 2023-03-10.

Bibliography

- [211] H. Caesar, J. Uijlings, and V. Ferrari, “COCO-Stuff: Thing and Stuff Classes in Context,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1209–1218.
- [212] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, “Semantic image synthesis with spatially-adaptive normalization,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2337–2346.
- [213] E. Arıkan, “Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels,” *IEEE Transactions on Information Theory*, vol. 55, no. 7, pp. 3051–3073, 2009.
- [214] F. Ercan, C. Condo, and W. J. Gross, “Reduced-Memory High-Throughput Fast-SSC Polar Code Decoder Architecture,” *IEEE Workshop on Signal Processing Systems (SiPS)*, pp. 1–6, 2017.
- [215] R. Senthamizh Selvi, B. Aarthi Varshini, and S. Deekshiga, “A comparative analysis of image denoising filters for salt and pepper noise,” in *2023 9th International Conference on Smart Structures and Systems (ICSSS)*, 2023, pp. 1–10.
- [216] The MathWorks, Inc., *Two-Dimensional Median Filtering - medfilt2*, MathWorks, Natick, Massachusetts, USA, 2022, accessed: October 30, 2022. [Online]. Available: <https://www.mathworks.com/help/images/ref/medfilt2.html>
- [217] S. Panchanathan, N. Gamaz, and A. Jain, “JPEG based scalable image compression,” *Computer Communications*, vol. 19, no. 12, pp. 1001–1013, 1996.
- [218] G. Shi, D. Gao, X. Song, J. Chai, M. Yang, X. Xie, L. Li, and X. Li, “A new communication paradigm: From bit accuracy to semantic fidelity,” *arXiv preprint arXiv:2101.12649*, 2021.
- [219] A. Krizhevsky and G. Hinton, “CIFAR-10 (canadian institute for advanced research),” 2009, accessed: 2024-11-09. [Online]. Available: <https://www.cs.toronto.edu/~kriz/cifar.html>

Bibliography

- [220] F.-F. Li, A. Karpathy, and J. Johnson, “Tiny ImageNet Visual Recognition Challenge,” <https://tiny-imagenet.herokuapp.com/>, 2015, accessed: 2023-05-04.
- [221] Y. Shao, Q. Cao, and D. Gunduz, “A theory of semantic communication,” 2025. [Online]. Available: <https://arxiv.org/abs/2212.01485>
- [222] F. Zhao, Y. Sun, R. Cheng, and M. A. Imran, “Background knowledge aware semantic coding model selection,” in *2022 IEEE 22nd International Conference on Communication Technology (ICCT)*, 2022, pp. 84–89.
- [223] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, “Gradient surgery for multi-task learning,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS ’20. Red Hook, NY, USA: Curran Associates Inc., 2020.
- [224] Z. Qin, F. Gao, B. Lin, X. Tao, G. Liu, and C. Pan, “A generalized semantic communication system: From sources to channels,” *IEEE Wireless Communications*, vol. 30, no. 3, pp. 18–26, 2023.
- [225] J. Ballé, D. C. Minnen, S. Singh, S. J. Hwang, and N. Johnston, “Variational image compression with a scale hyperprior,” *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, vol. abs/1802.01436, 2018. [Online]. Available: <https://api.semanticscholar.org/CorpusID:3611540>
- [226] M. M. Fouda, Z. M. Fadlullah, M. I. Ibrahem, and N. Kato, “Privacy-preserving data-driven learning models for emerging communication networks: A comprehensive survey,” *IEEE Communications Surveys & Tutorials*, pp. 1–1, 2024.
- [227] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein GAN,” in *Proceedings of the 34th International Conference on Machine Learning*. PMLR, 2017, pp. 214–223.
- [228] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting of*

Bibliography

- the Association for Computational Linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.
- [229] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2008.
- [230] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, “Improved training of wasserstein GANs,” in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.
- [231] Y. Ni and P. Koniusz, “ $\odot$ chain: Enhancing generalization in data-efficient gans via lipschitz continuity constrained normalization,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 6763–6774.
- [232] Z. Zhou, Y. Song, L. Yu, H. Wang, J. Liang, W. Zhang, Z. Zhang, and Y. Yu, “Understanding the effectiveness of lipschitz-continuity in generative adversarial nets,” 2018. [Online]. Available: <https://arxiv.org/abs/1807.00751>
- [233] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.
- [234] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.
- [235] W. Lindsey, “Design of block-coded communication systems,” *IEEE Transactions on Communication Technology*, vol. 15, no. 4, pp. 525–534, 1967.
- [236] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (VOC) challenge,” in *International journal of computer vision*, vol. 88, no. 2. Springer, 2010, pp. 303–338.
- [237] E. Agustsson and R. Timofte, “NTIRE 2017 challenge on single image super-resolution: Dataset and study,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017.

Bibliography

- [238] L. simon, “CelebA-HQ dataset,” <https://www.kaggle.com/datasets/lamsimon/celebahq>, 2021, accessed: 2024-01-04.
- [239] M. Hodosh, P. Young, and J. Hockenmaier, “Flickr8k: A dataset for image and sentence based image description,” 2013, aaccessed Oct. 3, 2023. [Online]. Available: https://github.com/jbrownlee/Datasets/releases/tag/Flickr8k_Dataset
- [240] F. Bellard, “BPG image format,” <https://bellard.org/bpg/>, 2014, accessed: 2024-06-26.
- [241] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTscore: Evaluating text generation with BERT,” in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://arxiv.org/abs/1904.09675>
- [242] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local nash equilibrium,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6629–6640.
- [243] A. Casanova, M. Careil, J. Verbeek, M. Drozdal, and A. Romero Soriano, “Instance-conditioned GAN,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 27 517–27 529, 2021.
- [244] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, “Demystifying MMD GANs,” *arXiv preprint arXiv:1801.01401*, 2018.
- [245] A. Brock, J. Donahue, and K. Simonyan, “Large scale GAN training for high fidelity natural image synthesis,” in *International Conference on Learning Representations*, 2019. [Online]. Available: <https://openreview.net/forum?id=B1xsqj09Fm>
- [246] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, “Training generative adversarial networks with limited data,” in *Proc. NeurIPS*, 2020.

Bibliography

- [247] H. Caesar, J. Uijlings, and V. Ferrari, “COCO-stuff: The official homepage of the COCO-stuff dataset,” <https://github.com/nightrome/cocostuff?tab=readme-ov-file#downloads>, 2025, accessed: 2023-09-03.
- [248] M. Tan, R. Pang, and Q. V. Le, “EfficientDet: Scalable and efficient object detection,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 778–10 787.
- [249] E. Makov, “Images 4K dataset,” <https://www.kaggle.com/datasets/evgeniumakov/images4k>, 2020, accessed: 2025-07-06.
- [250] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd ed. Pearson Education, 2002.
- [251] TensorFlow Authors, “Adversarial example using fast gradient sign method (FGSM),” https://www.tensorflow.org/tutorials/generative/adversarial_fgsm, 2023, accessed: 2025-08-25.
- [252] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996, p. 226–231.
- [253] S. Pidhorskyi, D. A. Adjeroh, and G. Doretto, “Adversarial latent autoencoders,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 14 092–14 101.
- [254] A. Roy, A. Vaswani, A. Neelakantan, and N. Parmar, “Theory and experiments on vector quantized autoencoders,” 2018. [Online]. Available: <https://arxiv.org/abs/1805.11063>
- [255] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” 2015. [Online]. Available: <https://arxiv.org/abs/1412.6572>

Appendix A

Appendix A

This appendix presents the mathematical proofs of the propositions in Chapter 6 and training curves of SVQ-WGAN

A.1 Mathematical proofs

We present mathematical proofs to support the theoretical claims underlying the model SVQ-WGAN.

A.1.1 Proof of proposition 5.3.1

Reconstruction loss preserves task utility under semantic equivalence

Proof. Since f is assumed to enforce Lipschitz condition on $\mathcal{M}_{\text{sem}}$, there exists a constant $L_f > 0$ such that for all $\mathbf{x}, \hat{\mathbf{x}} \in \mathcal{M}_{\text{sem}}$,

$$d_{\mathcal{Y}}(f(\hat{\mathbf{x}}), f(\mathbf{x})) \leq L_f \cdot \|\hat{\mathbf{x}} - \mathbf{x}\|_2$$

where $d_{\mathcal{Y}}$ denotes a task-output distance metric (e.g., class label mismatch or IoU divergence). If $\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 \leq \varepsilon$, then

$$d_{\mathcal{Y}}(f(\hat{\mathbf{x}}), f(\mathbf{x})) \leq L_f \cdot \sqrt{\varepsilon}$$

For sufficiently small ε , the output perturbation is below any classification or localization

Appendix A. Appendix A

threshold, resulting in identical task outputs: If f is piecewise constant (in classification tasks), or if the task-specific output metric satisfies $d_Y(f(\hat{\mathbf{x}}), f(\mathbf{x})) < \delta$ for some task-dependent threshold δ , then small input perturbations (noise) $\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 \leq \varepsilon$ satisfying $L_f \cdot \sqrt{\varepsilon} < \delta$ ensure $f(\hat{\mathbf{x}}) = f(\mathbf{x})$.

Justification. In this design, the semantic manifold $\mathcal{M}_{\text{sem}}$ is represented by the training distribution, which combines DIV2K, CelebA-HQ, and Flickr15k datasets, augmented for contrastive learning. The learned manifold captures the structural and perceptual regularities common across natural images, supporting the assessment of task utility on a different dataset (MS-COCO). The model is trained to represent the noisy or incomplete inputs back into this manifold, which preserves meaningful semantic content even under compression or transmission degradation. Thereby providing robustness and generalization, without explicit task-based training. The low FID scores reported in Section 5.4.6, empirically supports that $\hat{\mathbf{x}} \in \mathcal{M}_{\text{sem}}$. Thus, minimizing reconstruction loss preserves task-relevant semantics without requiring explicit task supervision.

Empirical validation The image classification results on the CIFAR-10 dataset in Fig. 5.10 demonstrate that SVQ-WGAN maintains high classification accuracy (94%) at SNR levels above 4 dB, closely approaching the clean baseline. Additionally, Table 5.2 shows higher BERTScore F1 values for SVQ-WGAN in image captioning tasks on the COCO dataset at moderate and high SNRs.

□

A.1.2 Proof of Proposition 5.3.2

Learned prior over quantized latents enables structured semantic decoding

Proof. Let $\mathcal{T}$ be a stochastic perturbation process applied to the quantized latent code $\mathbf{z}_q$, such that $\tilde{\mathbf{z}}_q = \mathcal{T}(\mathbf{z}_q) = \mathbf{z}_q + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ or represents index masking. Then,

$$\|G(\tilde{\mathbf{z}}_q) - G(\mathbf{z}_q)\|_2 \leq L_G \|\tilde{\mathbf{z}}_q - \mathbf{z}_q\|_2 = L_G \|\varepsilon\|_2.$$

Appendix A. Appendix A

Taking expectation over the noise distribution, the generator’s output under perturbation satisfies:

$$\mathbb{E}_{\mathcal{T}} [\|G(\tilde{\mathbf{z}}_q) - \mathbf{x}\|^2] \leq L_G^2 \mathbb{E}_{\mathcal{T}} [\|\varepsilon\|^2] + \|G(\mathbf{z}_q) - \mathbf{x}\|^2,$$

The proof indicates theoretically that the designed generator learns semantically consistent outputs even with corrupted latents. The generator is constructed from Lipschitz-continuous layers with gradient penalty imposed with stable training, hence $G(\mathbf{z}_q)$ exhibits stable decoding behavior across the learned prior distribution. $\square$

Empirical validation As illustrated in the ablation study in Fig.5.11, different codebook sizes result in consistent performance across PSNR, MS-SSIM, and LPIPS, even under low SNR and fading channels. The generalization to the COCO with low FID (5.496) and KID (0.0073) scores, as reported in Section 5.4.6, validates that the generator decodes quantized latent vectors in a structured and semantically meaningful manner.

A.1.3 Proof of proposition 5.3.3

Adversarial training with induced noise enables semantic robustness

Proof. Let $\mathbb{P}_{data}$ be the real image distribution and $\mathbb{P}_{gen}^\varepsilon$ the distribution of reconstructed images under perturbed latents $\tilde{\mathbf{z}}_q$. The adversarial objective approximates the Wasserstein distance:

$$W(\mathbb{P}_{data}, \mathbb{P}_{gen}^\varepsilon) = \sup_{\|D\|_L \leq 1} [\mathbb{E}_x[D(\mathbf{x})] - \mathbb{E}_{\tilde{\mathbf{z}}_q}[D(G(\tilde{\mathbf{z}}_q))]].$$

Minimizing $\mathcal{L}_{adv}$ with respect to G encourages $\mathbb{P}_{gen}^\varepsilon \rightarrow \mathbb{P}_{data}$, implying $G(\tilde{\mathbf{z}}_q) \in \mathcal{M}_{sem}$, even when $\tilde{\mathbf{z}}_q \neq \mathbf{z}_q$. Moreover, if G is Lipschitz continuous with constant L_G , then:

$$\|G(\tilde{\mathbf{z}}_q) - G(\mathbf{z}_q)\|_2 \leq L_G \cdot \|\varepsilon\|_2.$$

Appendix A. Appendix A

Hence, for small $\|\varepsilon\|$, the output $G(\tilde{\mathbf{z}}_q)$ remains close to the clean reconstruction $G(\mathbf{z}_q)$, preserving semantic content. $\square$

Empirical validation As illustrated in the ablation study Fig.5.11, the result without noise regularization shows a degraded quality in the transmitted images at all SNR levels. The ablation study on classification without the adversarial training results in degraded classification accuracy, shown in Fig. 5.10

A.1.4 Proof of proposition 5.3.4

Index dropout combined with adversarial training enables error-tolerant reconstruction

Proof. For each position $i \in \{1, \dots, N\}$, let $S_{\text{drop}} = \{i : m_i = 0\}$ denote the dropped indices. The difference between the corrupted and clean latent vectors is bounded by:

$$\mathbb{E}[\|\tilde{\mathbf{z}}_q - \mathbf{z}_q\|^2] \leq N \cdot \mathbb{P}_{\text{mask}} \cdot \max_i \|\mathbf{z}_q[i] - \mathbf{z}_0\|^2,$$

where N denotes the total number of latent vectors in the quantized representation, and $\mathbf{z}_q[i] \in \mathbb{R}^d$ is the i -th latent embedding. The index $i \in \{1, \dots, N\}$ refers to positions within the latent sequence. The term $\max_i \|\mathbf{z}_q[i] - \mathbf{z}_0\|^2$ captures the worst-case replacement error introduced by the default vector $\mathbf{z}_0$ under index dropout. Assuming G is Lipschitz with constant L_G , we have:

$$\mathbb{E}[\|G(\mathbf{z}_q) - G(\tilde{\mathbf{z}}_q)\|^2] \leq L_G^2 \cdot \mathbb{E}[\|\tilde{\mathbf{z}}_q - \mathbf{z}_q\|^2].$$

Taking expectation and combining with the bound above gives the stated $\delta(\mathbb{P}_{\text{mask}})$ and the limit $\delta(\mathbb{P}_{\text{mask}}) \rightarrow 0$ as $\mathbb{P}_{\text{mask}} \rightarrow 0$. The reconstruction loss ensures $G(\tilde{\mathbf{z}}_q) \approx \mathbf{x}$, while the adversarial loss ensures $G(\tilde{\mathbf{z}}_q) \in \mathcal{M}_{\text{sem}}$. Thus, reconstructions are semantically valid under index masking. $\square$

Empirical validation This is empirically validated through the ablation study of different variants of the model in Fig. 5.12, where the SVQ-WGAN with $P_{mask}=0$, shows a degraded quality in PSNR and LPIPS.

A.1.5 Proof of proposition 5.3.5

Symbolic semantic compression bounds bandwidth usage

Proof. Each codebook vector $\mathbf{z}_q[i] \in \{1, \dots, K\}$ is an index into a codebook of size K , and requires $\log_2 K$ bits to encode. Since the full latent sequence $\mathbf{z}_q(x)$ contains N codebook vectors, the total transmission bandwidth is:

$$B = N \cdot \log_2 K.$$

Let $x \in \mathbb{R}^{H \times W \times C}$ be the original input, with total size $D = HWC$. Then, the compression ratio is:

$$\frac{B}{D} = \frac{N \cdot \log_2 K}{HWC}.$$

In practice, $N = H'W'$, where $H', W' \ll H, W$, due to spatial downsampling in the encoder. As such, this ratio indicates substantial compression.

Furthermore, B can be explicitly controlled by choosing N (via encoder stride) and K (via codebook size). Hence, symbolic quantized representations allow deterministic control over bandwidth usage, unlike continuous latent codes used in JSCC systems.

Assume each symbol is transmitted over a channel using LDPC coding with coding rate $R_c \in (0, 1]$.

Then, the total number of transmitted bits is:

$$B_{tx} = \frac{N \cdot \log_2 K}{R_c},$$

and the effective compression ratio becomes:

$$\frac{B_{tx}}{HWC} = \frac{N \cdot \log_2 K}{R_c \cdot HWC}.$$

Thus, the use of LDPC increases the transmission bandwidth by a factor of $1/R_c$, while providing robustness against bit errors, making the model suitable for lossy channels. The encoder design enables explicit trade-off control between bandwidth efficiency and channel resilience.

Empirical validation The bandwidth efficiency of the symbolic latent representation is empirically verified in Section 5.4.3. SVQ-WGAN transmits fewer bits per pixel (0.438 bpp vs. 0.562 for BPG and 1.0 for JSCC) while achieving comparable or better perceptual quality as illustrated in Fig.5.6,5.7 $\square$

A.2 Training stability

Training and validation loss curves over 150 epochs for four key components of the SVQ-WGAN framework are presented in Fig. A.1, illustrating training stability and convergence. The reconstruction loss decreases sharply during the initial 40 epochs and gradually stabilizes for both training and validation sets. This indicates that the generator learns to reconstruct images more accurately over time. The validation loss initially fluctuates more, but after 70 epochs, it stabilizes closely with the training loss. The close tracking of the validation curve with the training curve shows effective generalization and no overfitting. The VQ loss consists of codebook and commitment terms, and converges smoothly after an initial transient phase of approximately 50 epochs. The consistent close convergence between training and validation losses reflects successful codebook learning and encoder commitment to the quantized representations. The discriminator loss initially fluctuates, which is typical in adversarial training. However, it rapidly stabilizes and remains steady beyond 50 epochs. The curve validates the effectiveness of the Wasserstein loss with gradient penalty ($\lambda = 10$), which prevents gradient explosion or vanishing. The stable discriminator behavior contributes to more consistent updates for the generator, avoiding mode collapse and enabling high-fidelity reconstructions. The total loss, comprising reconstruction, VQ, adversarial, and perceptual losses, shows a steady trend and stabilizes after 40 epochs.

The consistent convergence and alignment of the loss curves across all components

Appendix A. Appendix A

demonstrate that the SVQ-WGAN model is stably trained. This stability is a result of key architectural and training design choices, including vector quantization with a well-regularized codebook, Wasserstein GAN with a gradient penalty, noise injection, and index dropout for robustness, as well as cosine learning rate scheduling. These mechanisms collectively ensure smooth optimization dynamics and effective learning under noisy channel conditions.

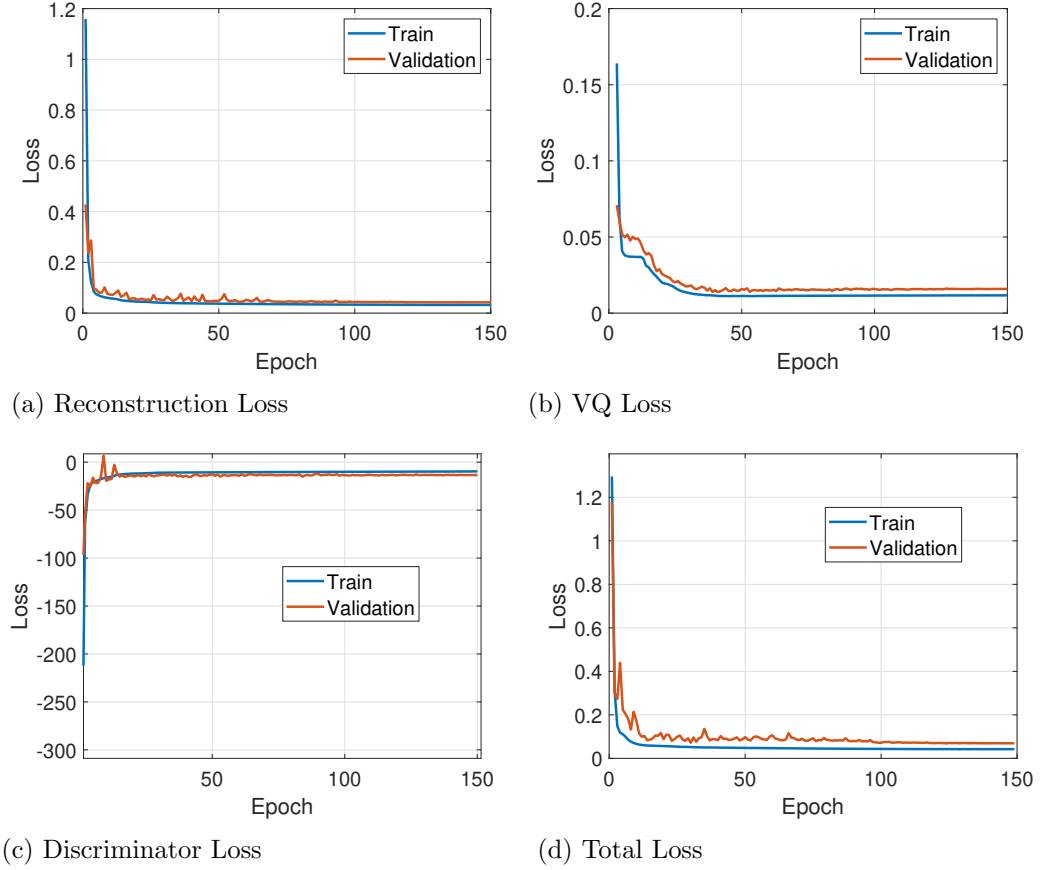


Figure A.1: losses (a) Reconstruction Loss (b) VQ Loss. (c) Discriminator Loss (d) Total Loss.