

# Semantic-Aware Attention-Driven JSCC for Efficient Video Transmission over Wireless Channels

G.Coppola, M.Narimani Kenari, G.Sciddurlo, N.Cordeschi, L.A.Grieco, G.Boggia  
Department of Electrical and Information Engineering, Politecnico di Bari, 70125 Bari, Italy  
Consorzio Nazionale Interuniversitario per le Telecomunicazioni, 43124 Parma, Italy  
Email: g.coppola2@studenti.poliba.it, m.narimanikenari@phd.poliba.it, {name.surname}@poliba.it

**Abstract**—Separation-based video coding often fails in low-latency, noisy wireless scenarios due to channel sensitivity and lack of task awareness. Semantic communication, leveraging end-to-end task-oriented architectures such as Deep Joint Source-Channel Coding (DJSCC), offers a robust alternative. This paper introduces a lightweight video transmission framework that integrates a Semantic Attention (*SemAtt*) module with a Signal to Noise Ratio (SNR) modulation (*SNRMod*) for key-frame encoding and decoding. *SemAtt* uses semantic segmentation maps as compact priors to enhance feature discriminability with minimal computational overhead, while semantic priors also aid non-key frame reconstruction via lightweight generative models. Experiments on high-resolution urban scenes show improved robustness under low SNR, with a 5.2% reduction in LPIPS and a 6.3% increase in mIoU over conventional coding schemes, achieved with only a 9.9% increase in GFLOPs and 1.1% in trainable parameters. The framework is thus suitable for real-time, bandwidth-constrained wireless applications.

**Keywords**—Semantic communication, Joint Source-Channel Coding, deep learning, semantic segmentation.

## I. INTRODUCTION

The rapid growth of multimedia platforms and high-bandwidth wireless access has made video the dominant form of Internet traffic, accounting for nearly 80% of global traffic and expected to grow further [1]. This has driven research on efficient and robust wireless video transmission, traditionally based on source-channel separation. In these architectures, source coding (e.g., Advanced Video Coding (AVC)) compresses video into bitstreams, while channel coding protects them from wireless noise impairments. Despite their success, such designs are sensitive to channel mismatches: when actual conditions deviate from the assumed code rate, decoding errors can increase sharply, leading to complete reconstruction failures, known as the *cliff effect*. Furthermore, conventional schemes ignore semantic relevance and downstream task requirements, resulting in inefficient resource utilization [2], which is critical for emerging applications like wireless virtual and augmented reality requiring ultra-low latency with constrained devices [2].

Semantic communications have recently emerged as an end-to-end, task-oriented paradigm [3], transmitting only task-relevant information and potentially overcoming the limitations of conventional architectures. Deep learning techniques have proven effective for extracting and leveraging semantic

features in multimedia transmission [4], [5]. In particular, Deep Joint source-Channel Coding (DJSCC) methods enhance robustness to channel variations, especially under low SNRs by providing graceful performance degradation while reducing latency and improving bandwidth efficiency [4]. Recent video-focused DJSCC frameworks often adopt key and non-key frame structures [2], transmit latent features with or without explicit semantic priors [2], [6]–[8], and aim to exploit task-relevant information. While recent works have successfully integrated semantic information for video reconstruction, many existing frameworks rely on computationally intensive modules or prioritize the generative reconstruction of non-key frames. There remains a need for lightweight architectures that can robustly transmit key frames under variable channel conditions by fully exploiting compact semantic priors without incurring significant computational overhead.

Motivated by these limitations, this work makes the following contributions:

- We propose a Deep Learning (DL)-enabled DJSCC framework for key-frame transmission that integrates semantic attention (*SemAtt*) modules leveraging segmentation maps as priors to guide encoding and enhance feature discriminability. We focus on key frames since semantic errors at this stage can propagate and degrade the reconstruction of subsequent frames.
- We further introduce a novel Segmentation Adjacency Matrix (SGM)-based mechanism combined with SNR-aware modulation, improving semantic interpretability and reconstruction performance across varying channel conditions with negligible additional model complexity.
- Experimental results on benchmark datasets demonstrate high reconstruction fidelity under low SNR. Compared to conventional separation-based schemes prone to the digital cliff effect, our approach achieves a 5.2% reduction in Learned Perceptual Image Patch Similarity (LPIPS) and a 6.3% improvement in Mean Intersection over Union (mIoU).

The remainder of this paper is organized as follows. Section II describes the proposed system and semantic attention modules. Section III reports performance evaluation, and Section IV concludes the paper.

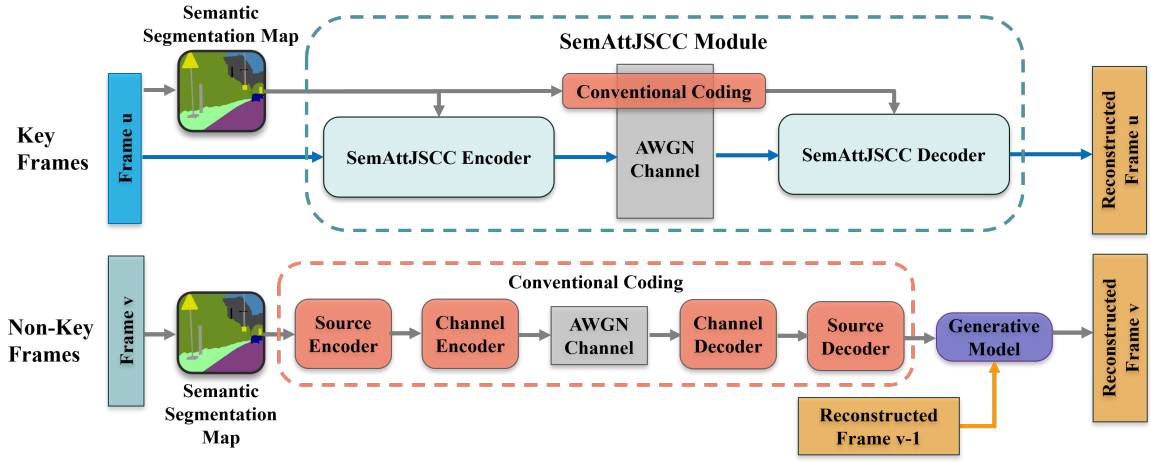


Fig. 1: Semantic-aided Joint Source-Channel Coding (JSCC) framework overview.

## II. SEMANTIC-AIDED JSCC FRAMEWORK FOR VIDEO TRANSMISSION

This section introduces a semantic-aided JSCC framework for key-frame video transmission. Inspired by the object-attribute-relation paradigm in [7], the proposed architecture incorporates a *Semantic Attention (SemAtt)* mechanism that embeds semantic priors directly into the encoding and decoding stages.

### A. Proposed Framework Architecture

Figure 1 illustrates the proposed framework. We consider wireless video transmission where sequences are divided into Group of Pictures (GoP), following [2]. The upper pipeline shows key-frame processing, with key frames defined as the first frame of each GoP. Semantic features are extracted from each key frame using pretrained HRNetV2 and Optical Character Recognition (OCR) models [9]–[11] to generate high-resolution segmentation maps. These maps, along with the RGB frame, are fed into the *SemAttJSCC* module, an end-to-end trainable system for joint source and channel coding. The module comprises a *SemAttJSCC Encoder* at the transmitter and a *SemAttJSCC Decoder* at the receiver. The encoder jointly processes the RGB image and its semantic map to generate complex-valued symbols transmitted over an Additive White Gaussian Noise (AWGN) channel. Given their compact representation, semantic maps are also transmitted via a parallel conventional source-channel coding pipeline, using 4-QAM modulation and a 1/3-rate Low-Density Parity Check (LDPC) code [7]. At the receiver, the decoded semantic information and key-frame symbols are jointly used by the decoder to reconstruct the RGB frame, with a symmetric architecture enabling end-to-end optimization.

The lower branch of the Figure 1, instead, corresponds to the non-key-frame processing pipeline within each GoP. To save bandwidth, only their semantic maps are transmitted via the separate coding scheme. At the receiver, the decoded map and the previously reconstructed RGB frame are fed into a

vid2vid-based generative model [12] to synthesize the current frame. The *vid2vid* model is pretrained on the target dataset and deployed in inference mode at the receiver. As *vid2vid* operates exclusively on non-key frames, it does not impact the complexity analysis of key-frame transmission.

### B. The SemAttJSCC Module

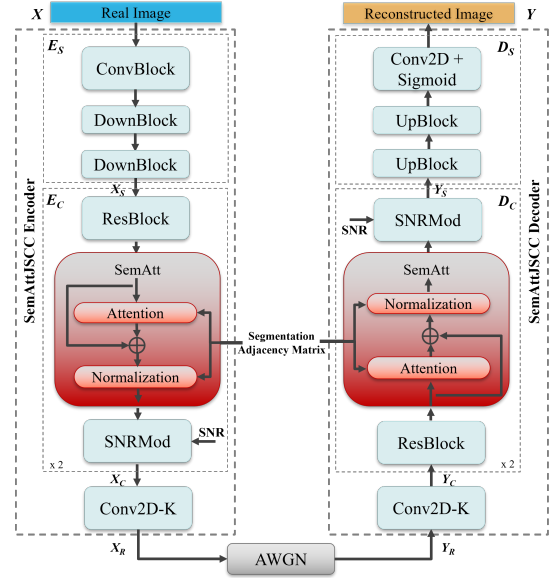


Fig. 2: SemAttJSCC module architecture.

The proposed *SemAttJSCC* module, designed for key-frame transmission, extends the DynamicJSCC framework [13]. As shown in Figure 2, it comprises a jointly trained encoder-decoder pair: the *SemAttJSCC Encoder* and *Decoder*, each logically divided into source and channel blocks.

The source encoder  $E_S$  receives an RGB image  $X \in \mathbb{R}^{3 \times H \times W}$  and extracts high-level features, reducing spatial resolution by a factor  $G$  and expanding channel dimension to  $C$ , resulting in  $X_S \in \mathbb{R}^{C \times H/G \times W/G}$ . We adopt  $G = 4$

and  $C = 256$  as in [13]. The channel encoder  $E_C$  applies shape-preserving transformations, including a residual block (*ResBlock*) and an SNR-aware modulation block (*SNRMod*), to incorporate channel state information. A key novelty is the insertion of a *Semantic Attention (SemAtt)* block prior to *SNRMod* (highlighted in Figure 2), enabling semantic-aware feature modulation. Inspired by [14], [15], the *SemAtt* block exploits semantic segmentation priors via the SGM, guiding both attention and normalization stages.

The *SemAtt* block has three stages:

- **Attention layer:** processes latent features together with the SGM from a downsampled segmentation map.
- **Residual connection:** adds the attention output to the input features for stable gradient propagation [16].
- **Semantic-aware normalization:** adjusts activations based on region-specific statistics.

The SGM  $A \in \{0, 1\}^{L \times L}$  is constructed from  $M$  semantic regions  $S_m$ , for  $m = 1, \dots, M$ , where  $L = H \times W$ :

$$a_{ij} = \begin{cases} 1, & \text{if pixels } i, j \in S_m \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

To reduce memory and computation, a compressed representation  $A_r \in \{0, 1\}^{L \times M}$  is used, preserving semantic relationships with only  $L$  non-zero entries.

The feature tensor is reshaped into  $F \in \mathbb{R}^{C \times L}$ , and a deviation matrix  $D$  is computed as

$$d_{ij} = f_{ij} - \frac{\sum_{k=1}^L f_{ik} a_{kj}}{\sum_{k=1}^L a_{kj}}, \quad (2)$$

capturing channel-wise deviations from region-wise means. For each semantic region  $S_m$ , the attention map  $\Psi^{S_m} \in \mathbb{R}^{C \times C}$  is defined by [15]:

$$\psi_{ij}^{S_m} = \frac{e^{\frac{1}{|S_m|} \sum_{k \in S_m} d_{ik} d_{jk}}}{\sum_{j=1}^C e^{\frac{1}{|S_m|} \sum_{k \in S_m} d_{ik} d_{jk}}}. \quad (3)$$

The attention output is equal to:

$$\varphi_{ij} = f_{ij} + \sum_{k=1}^C \psi_{ik}^{S_m(j)} f_{kj}, \quad (4)$$

followed by semantic-aware normalization:

$$z_{ij} = \frac{\varphi_{ij} - \mathbb{E}[\varphi_{ij}^{S_m(j)}]}{\sqrt{\text{Var}[\varphi_{ij}^{S_m(j)}]}} w_{im} + b_{im}, \quad (5)$$

where  $w, b \in \mathbb{R}^{C \times M}$  are learnable parameters. Assuming region-wise feature consistency, this reduces parameters by a factor  $M/L$ . The output is reshaped to  $Z \in \mathbb{R}^{C \times H/G \times W/G}$  and passed to *SNRMod*. After *SNRMod*, a *Conv2D-K* layer projects features to  $K$  channels, followed by power normalization and mapping to complex symbols  $X_R \in \mathbb{C}^k$ , with  $k = (K/2) \times H/G \times W/G$ . The channel is modeled as AWGN:

$$Y_R = X_R + n, \quad n \sim \mathcal{CN}(0, \sigma^2 I_k), \quad (6)$$

with  $\text{SNR} = 10 \log_{10}(1/\sigma^2)$  dB.

The bandwidth compression ratio is  $R = k/n$ , with  $n = 3HW$ , controlled via  $K$ . At the receiver,  $Y_R$  is mapped back to a latent tensor and processed by the channel and source decoders  $D_C$  and  $D_S$ , mirroring  $E_C$  and  $E_S$ , to reconstruct  $Y$ . The module is trained end-to-end to minimize the reconstruction Mean Square Error (MSE):

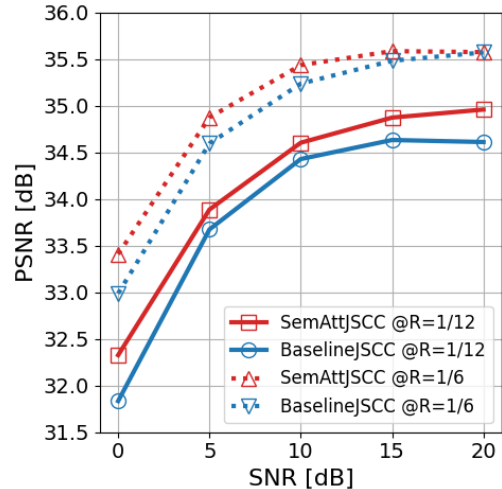
$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N \|X_i - Y_i\|^2, \quad (7)$$

where  $N$  is the number of training samples.

### III. EXPERIMENTAL RESULTS

We evaluate the proposed JSCC framework for key-frame transmission on the *Cityscapes* dataset [17], a standard benchmark for semantic urban scene understanding. The dataset contains 2975 training images and 500 validation images with resolution  $1024 \times 2048$  and pixel-level annotations across 34 semantic classes.

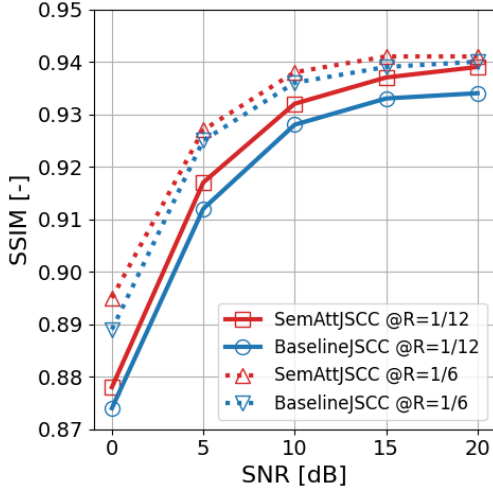
The *SemAttJSCC* model is trained with the Adam optimizer [18] (learning rate  $5 \times 10^{-4}$ , batch size 2) on a single NVIDIA A100-SXM4-40GB GPU (CUDA 12.2). Early stopping with a patience of 14 epochs is applied, yielding convergence after roughly 35 epochs. During training, the SNR is uniformly sampled in  $[0, 20]$  dB. Two bandwidth compression ratios are considered:  $R = 1/12$  and  $R = 1/6$ . The DynamicJSCC baseline [13] is trained under identical conditions. Crucially, since this baseline already incorporates the *SNRMod* modulation mechanism, the performance comparison presented in this section effectively serves as an ablation study, isolating the specific contribution of the proposed *SemAtt* module.



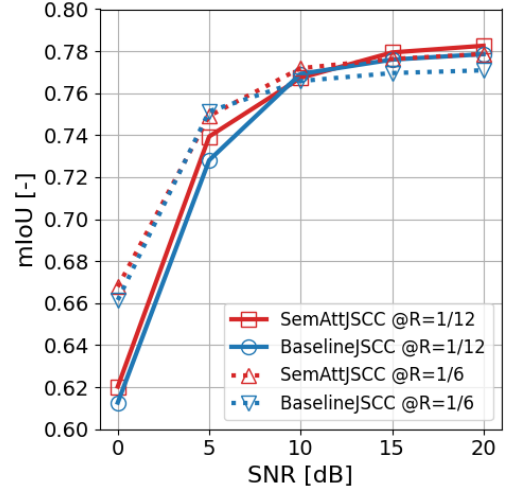
**Fig. 3:** PSNR comparison between *SemAttJSCC* and the baseline for varying SNR and  $R$ .

#### A. Comparison with Baseline JSCC Scheme

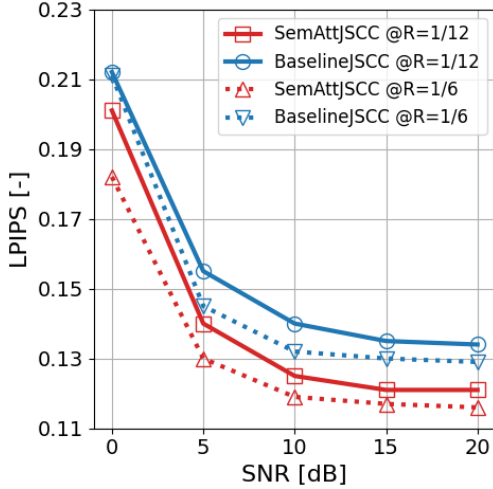
Performance is evaluated using both fidelity-based and task-oriented metrics. Peak Signal-to-Noise Ratio (PSNR)



**Fig. 4:** SSIM comparison between *SemAttJSCC* and the baseline for varying SNR and  $R$ .



**Fig. 6:** mIoU comparison between *SemAttJSCC* and the baseline for varying SNR and  $R$ .



**Fig. 5:** LPIPS comparison between *SemAttJSCC* and the baseline for varying SNR and  $R$ .

and Structural Similarity Index Measure (SSIM) quantify pixel-level reconstruction quality, while LPIPS [19] measures perceptual similarity. Ground-truth segmentation maps are used to isolate transmission effects, and downstream semantic performance is evaluated via a pretrained HRNetV2+OCR model [9], [10], [20], computing mIoU on reconstructed images.

Figures 3 and 4 show that *SemAttJSCC* consistently outperforms the baseline across all SNR values and both compression ratios. Gains are more pronounced at stronger compression ( $R = 1/12$ ) and low SNR, confirming improved robustness under adverse channel conditions. A maximum PSNR gain of 1.5% occurs at  $SNR = 0$  dB and  $R = 1/12$ , while the largest SSIM improvement is 0.7% at  $SNR = 0$  dB and  $R = 1/6$ .

Figure 5 shows that *SemAttJSCC* achieves consistently lower LPIPS scores than the baseline across all channel conditions, indicating superior perceptual quality. The maximum reduction is 13.7% at  $SNR = 0$  dB and  $R = 1/6$ . Figure 6 demonstrates higher mIoU values, particularly under strong noise and compression, with the largest gain of 1.6% at  $R = 1/12$  and  $SNR = 5$  dB. These results confirm that semantic-aware encoding effectively preserves class-discriminative information for downstream tasks.

### B. Comparison with Conventional Coding Schemes

We further evaluate the proposed *SemAttJSCC* framework against conventional separate source and channel coding pipelines, following [7]. In these reference schemes, images are first compressed using JPEG2000 or BPG [21], then protected via LDPC coding [22] and digitally modulated. We consider LDPC code rates of 1/3, 1/2, and 2/3, with BPSK, 4-QAM, 16-QAM, and 64-QAM modulation. Perfect Channel State Information (CSI) is assumed at the receiver for log-likelihood ratio (LLR) computation [23]. For fair comparison, the compressed file sizes are matched to the number of transmitted symbols used by *SemAttJSCC*, including the overhead associated with transmitting the compressed SGM. It is worth noting that this overhead is minimal, amounting on average to 0.4% of the compression bandwidth ratio  $R$ .

Figures 7 and 8 highlight that JSCC-based methods inherently avoid the digital cliff effect typical of conventional digital pipelines. In terms of perceptual quality (measured via LPIPS), *SemAttJSCC* consistently outperforms JPEG2000-based schemes across all SNR values. Compared to BPG-based pipelines, the proposed method achieves superior reconstruction fidelity in low-SNR regimes, where digital schemes often fail due to unrecoverable bit errors. Specifically, at  $SNR = 0$  dB, *SemAttJSCC* reduces LPIPS by 5.2% and increases mIoU by 6.3% relative to the BPG+1/3LDPC+BPSK

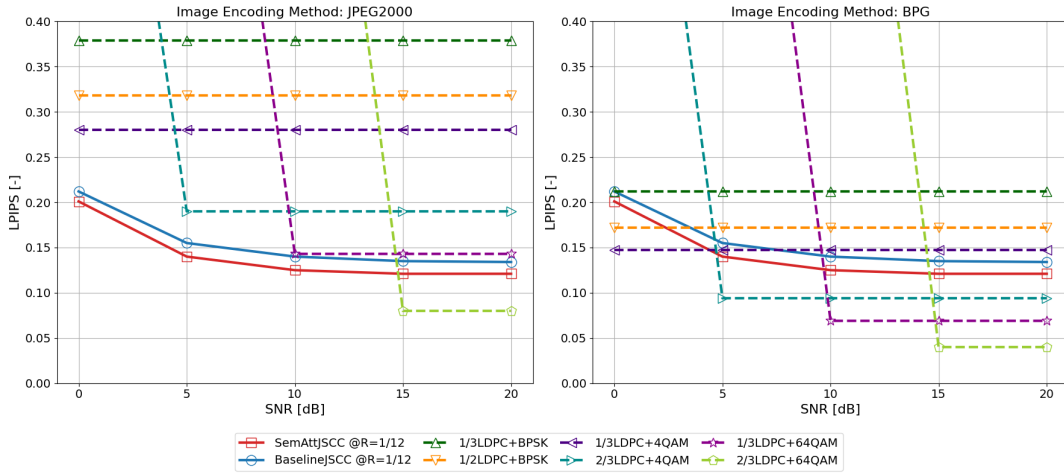


Fig. 7: LPIPS comparison with conventional coding schemes under varying SNR values.

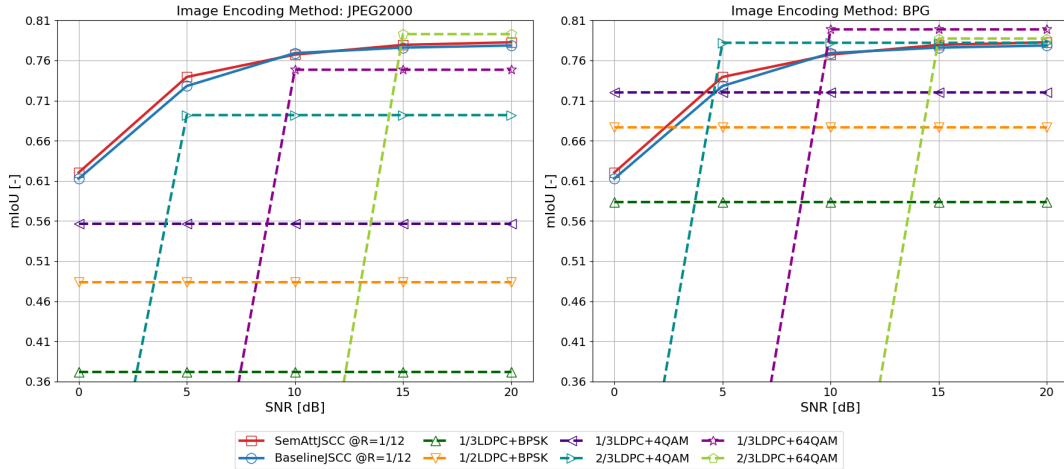


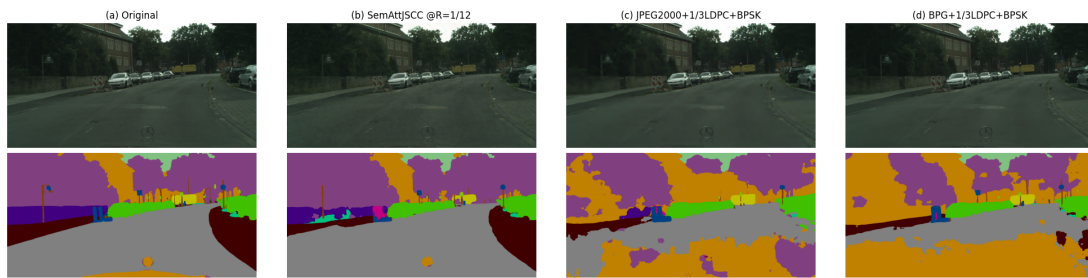
Fig. 8: mIoU comparison with conventional coding schemes under varying SNR values.

Architecture	GFLOPs	Parameters (M)	Max. LPIPS Reduction vs. Baseline
Baseline JSCC	2079	6.57	–
<i>SemAttJSCC</i>	2285	6.64	13.7%
Baseline JSCC + Convolutions [7]	3326	11.33	11.0%

**TABLE I:** Computational complexity and number of trainable parameters for the proposed *SemAttJSCC*, the baseline JSCC, and the convolution-enhanced JSCC model [7], evaluated at bandwidth compression ratio  $R = 1/12$  for key-frames transmission

configuration. While BPG demonstrates competitive performance at high SNR due to efficient source coding, this advantage primarily stems from optimized compression rather than transmission robustness, which is the focus of our study. Additionally, the high-resolution images ( $1024 \times 2048$ ) pose additional challenges for conventional and learned methods [7],

[23]. Future work could explore increasing the latent channel dimensionality  $K$  to better balance spatial compression and feature richness under a fixed bandwidth constraint. Similar trends are observed for downstream semantic tasks. Figure 8 shows that *SemAttJSCC* significantly outperforms both JPEG2000- and BPG-based pipelines in terms of mIoU, particularly at low SNR, confirming the effectiveness of semantic-aware encoding in preserving task-relevant information. A qualitative example is provided in Figure 9, showing reconstructed key frames and corresponding segmentation outputs at  $SNR = 0$  dB. The proposed method better preserves fine structures and semantic boundaries compared to conventional schemes, even under severe channel noise. In addition to improved performance, *SemAttJSCC* introduces only modest computational overhead. As summarized in Table I, GFLOPs and trainable parameters increase by only 9.9% and 1.1%, respectively, relative to the baseline JSCC. By contrast, the convolution-enhanced model in [7] incurs substantially higher overhead (60.0% GFLOPs and 72.5% parameters) while



**Fig. 9:** Visual comparison of reconstructed key frames (top row) and corresponding segmentation outputs (bottom row) at  $SNR = 0$  dB.

achieving a smaller maximum LPIPS reduction (11.0% vs. 13.7%). These results, albeit obtained on datasets of different resolutions, underscore the efficiency of integrating compact semantic priors directly within the JSCC framework.

#### IV. CONCLUSION

We introduced *SemAttJSCC*, a semantic-aware DJSCC framework for efficient key-frame transmission over bandwidth-limited, noisy wireless channels. By integrating compact semantic priors (SGMs) via lightweight attention mechanisms, the system guides feature allocation and channel adaptation. Experiments on high-resolution urban scenes show consistent gains over baseline JSCC in reconstruction (PSNR, SSIM), perceptual quality (LPIPS), and downstream semantic accuracy (mIoU), particularly under low SNR and strong compression. Compared to conventional digital pipelines (JPEG2000/BPG + LDPC), *SemAttJSCC* avoids the cliff effect and maintains semantic integrity. Future work will explore more realistic channel models, adaptive GoP structures, key-frame selection under strict latency and bandwidth constraints, and generative schemes for reconstructing non-key frames from semantically transmitted key frames..

#### ACKNOWLEDGEMENTS

This work has been supported by the SNS JU project 6G-GOALS under the EU's Horizon program Grant Agreement No 101139232

#### REFERENCES

- [1] V. N. I. Cisco, "Cisco visual networking index: Forecast and methodology, 2016–2021," *CISCO White paper*, vol. 2022, 2017.
- [2] 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.
- [3] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [4] E. Boursoulatzé, D. B. 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.
- [5] D. Huang, F. Gao, X. Tao, Q. Du, and J. Lu, "Toward semantic communications: Deep learning-based image semantic coding," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 55–71, 2022.
- [6] 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.
- [7] Q. Du, Y. Duan, Q. Yang, X. Tao, and M. Debbah, "Object-attribute-relation representation based video semantic communication," *IEEE Journal on Selected Areas in Communications*, 2025.
- [8] P. Jiang, C.-K. Wen, S. Jin, and G. Y. Li, "Wireless semantic communications for video conferencing," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 230–244, 2023.
- [9] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.
- [10] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Journal on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.
- [11] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 173–190.
- [12] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [13] M. Yang and H.-S. Kim, "Deep joint source-channel coding for wireless image transmission with adaptive rate control," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5193–5197, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:238583132>
- [14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [15] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.
- [16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [17] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [18] K. D. B. J. Adam *et al.*, "A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, vol. 1412, no. 6, 2014.
- [19] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.
- [20] Y. Yuan, X. Chen, X. Chen, and J. Wang, "Segmentation transformer: Object-contextual representations for semantic segmentation," 2021. [Online]. Available: <https://arxiv.org/abs/1909.11065>
- [21] F. Bellard. Better portable graphics. Available at: <https://bellard.org/bpg/>. [Online]. Available: <https://bellard.org/bpg/>
- [22] R. Gallager, "Low-density parity-check codes," *IRE Transactions on Information Theory*, vol. 8, no. 1, pp. 21–28, 1962.
- [23] M. Yang, C. Bian, and H.-S. Kim, "Deep joint source channel coding for wireless image transmission with ofdm," in *ICC 2021-IEEE International Conference on Communications*. IEEE, 2021, pp. 1–6.