SEMANTIC COMMUNICATION SYSTEM FOR VEHICLE DETECTION BASED ON CNN ENCODER AND SUPERNET ARCHITECTURE

Qikai Wang

Belarusian State University of Informatics and Radioelectronics
Minsk, Republic of Belarus

Jun Ma – Assistant

Abstract. This paper proposes a semantic communication system for vehicle detection tasks over noisy wireless channels. A CNN-based semantic encoder is designed to compress image features at multiple compression ratios (CH4, CH8, CH16), transmitting only task-relevant information rather than raw pixel data. To reduce model redundancy, a Supernet architecture with shared encoder weights is introduced, supporting all three compression ratios within a single model trained using the Sandwich Rule strategy. Experiments are conducted under AWGN channel conditions across five SNR levels (0-30 dB), evaluated using six detection metrics via the YOLOv12 framework on a multi-class vehicle dataset of 5,171 training images. Results demonstrate that at 0 dB SNR, the traditional transmission method suffers catastrophic performance collapse, while the semantic CH16 encoder retains mAP@0,5 = 0,414 (a reduction of only 8,2% from its 30 dB performance). Furthermore, Supernet_K16 achieves mAP@0,5 = 0,618 at 0 dB, compared to 0,627 for the independently trained CH16, a gap of only 1,4%, demonstrating that a single Supernet can effectively replace three independent models with negligible performance loss.

Keywords. Semantic communication, vehicle detection, YOLOv12, CNN encoder, supernet, AWGN channel, compression ratio.

Introduction. The rapid proliferation of intelligent transportation systems (ITS) has driven growing demand for reliable vehicle detection over wireless communication channels. In bandwidth-constrained or high-noise environments, conventional image transmission methods degrade significantly because they transmit the full raw pixel stream regardless of its task relevance. When channel quality is poor, compressed JPEG artifacts and AWGN noise corrupt pixel values irreversibly before reaching the inference engine, causing catastrophic detection failure.

Semantic communication offers a fundamentally different paradigm: instead of transmitting the image, only the features necessary to accomplish the downstream task are encoded and sent. This approach can provide strong robustness to channel noise because the encoder learns compact, task-aware representations that are far less sensitive to moderate channel distortion than raw RGB values.

In this work, we design a CNN-based semantic encoder for vehicle detection and systematically evaluate its robustness across five SNR levels under AWGN conditions. We further introduce a Supernet architecture that unifies three compression ratios (CH4, CH8, CH16) into a single model, reducing deployment overhead while preserving detection performance [1, 4].

System Model. The proposed semantic communication pipeline consists of four components: (i) a CNN semantic encoder E, (ii) an AWGN channel simulation, (iii) a CNN semantic decoder D, and (iv) a pre-trained and frozen YOLOv12 object detector. Given an input image x, the encoder maps it to a compact feature vector z where $C \in \{4, 8, 16\}$ is the channel dimension corresponding to compression ratios CH4, CH8, and CH16 respectively. The compressed feature is then corrupted by additive white Gaussian noise: $\hat{z} = z + n$, $n \sim N(0, \sigma^2 I)$ (1), where the noise variance σ^2 is determined by the target SNR level. The decoder D reconstructs a feature representation which is passed to the frozen YOLOv12 detector to produce bounding box predictions. For the traditional baseline, the raw image is JPEG-compressed, pixel values are directly corrupted by AWGN, and the noisy image is forwarded to YOLOv12 without any semantic encoding.

Supernet Architecture. Training three independent semantic encoders for CH4, CH8, and CH16 is computationally redundant. We therefore design a SupernetSemanticCodec that shares encoder and decoder backbone weights across all three sub-networks, differing only in the final channel projection layer[3,4]. Training is performed using the Sandwich Rule strategy: in each iteration, the largest sub-network (K=16), the smallest sub-network (K=4), and a randomly sampled intermediate sub-network (K=8) are each evaluated and their gradients accumulated before a single optimizer step. This ensures all sub-networks receive balanced gradient updates. The Supernet was trained for 50 epochs on an NVIDIA RTX 4090, completing training in approximately 53 minutes – the same wall-clock time as a single independent encoder.

Experiments. 1. Dataset and Setup: Experiments are conducted on a multi-class vehicle aerial image dataset containing 5,171 training images with annotations for cars, trucks, buses, SUVs, and motorcycles. Each semantic encoder is trained for 50 epochs. Evaluation is performed on a held-out validation set under five SNR conditions: {30, 20, 10, 5, 0} dB. Six metrics are reported: mAP@0,5, mAP@0,5:0,95, Precision, Recall, F1 Score, and Average Detection Count.

2. Semantic vs. Traditional Transmission: Figure 1 shows the mAP@0,5 curves for Traditional transmission and the three semantic encoders across SNR levels. The traditional method achieves the highest mAP@0,5 at 30 dB (0,511), but collapses sharply below 10 dB, reaching only 0,003 at 0 dB. In contrast, CH16 degrades gracefully, retaining 0,414 at 0 dB – a relative decrease of only 8,2% from its 30 dB performance (0,451).

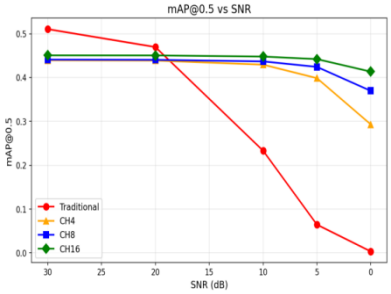

Figure 1 – mAP@0.5 vs SNR: Traditional vs semantic encoders (CH4/CH8/CH16)

Table 1 presents the complete quantitative evaluation. At 0 dB, CH16 achieves mAP@0.5 = 0.414 compared to Traditional 0.003, a 138× improvement. CH8 and CH4 show progressively lower performance as the compression ratio increases.

Table 1 – Full evaluation results across SNR levels

Method	SNR(dB)	mAP@0.5	Prec.	Recall	F1
Traditional	30	0.511	0.713	0.852	0.776
Traditional	10	0.233	0.741	0.465	0.572
Traditional	0	0.003	0.787	0.009	0.018
CH16	30	0.451	0.718	0.802	0.758
CH16	10	0.448	0.721	0.798	0.757
CH16	0	0.414	0.721	0.758	0.739

3. Supernet vs. Independent Models: Figure 2 compares the Supernet sub-networks (dashed lines) against independently trained encoders (solid lines) across all SNR levels. Supernet_K16 tracks CH16_Indep closely, with a maximum gap of 1.4% at 0 dB (0.618 vs 0.627). Table 2 provides a focused comparison at 0 dB SNR. The Supernet replaces three independent models with one, at approximately 67% parameter cost reduction, with minimal performance degradation.

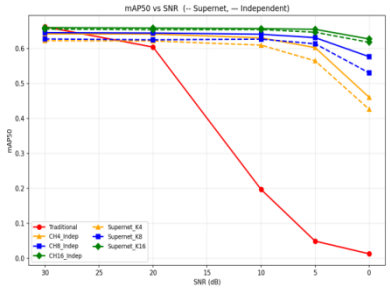

Figure 2 – mAP@0.5 vs SNR: Supernet sub-networks vs independent encoders

Table 2 – Comparison at 0 dB SNR: Supernet vs. independent encoders

Model	SNR(dB)	mAP@0.5	Prec.	Recall	F1
CH4_Indep	0	0.461	0.581	0.444	0.503
Supernet_K4	0	0.426	0.547	0.416	0.473
CH16_Indep	0	0.627	0.679	0.580	0.626
Supernet_K16	0	0.618	0.686	0.567	0.621

Conclusion. This paper presents a semantic communication system for vehicle detection that demonstrates strong robustness to AWGN channel noise. The proposed CNN semantic encoder (CH16) reduces mAP@0.5 degradation from 99.4% (Traditional) to 8.2% at 0 dB SNR. The Supernet architecture with Sandwich Rule training further consolidates three compression-ratio variants into a single model, with Supernet_K16 achieving only 1.4% lower mAP@0.5 than its independently trained counterpart at 0 dB. Future work will investigate adaptive compression ratio selection based on estimated channel quality, incorporation of attention mechanisms in the semantic encoder, and end-to-end joint optimization of the encoder-decoder-detector pipeline.

References:

1. Aoki T., Ohka S., Shiozawa D., Ogawa Y., Sakagami T. // Engineering Proceedings. 2019. Vol. 51. P. 44–55.
2. Hu Y., Xu L., Shen X., Jin L. // Applied Optics. 2021. Vol. 60. P. 9396–9403.
3. Wang D., Li Y., Pu Y. // Sensors. 2024. Vol. 24. P. 1307–1321.
4. Sun Y., Chen W., Li F., Gu Z., Feng L., Guo D. and Cai H. // IEEE Transactions on Instrumentation and Measurement. 2023 Feb. Vol. 72. P. 1–11.