ES-ISC Demo: An Explainable Semantics-based Image Semantic Communication System for 6G

Dongshan Ye*, Rundong Chen*, Xijun Wang*, Chenyuan Feng†, Xiang Chen*, and Tony Q. S. Quek‡
* School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China
† Department of Communication Systems, EURECOM, Biot 06410, France
‡ Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore
Email: {yedsh3,chenrd6}@mail2.sysu.edu.cn, {wangxijun,chenxiang}@mail.sysu.edu.cn, Chenyuan.Feng@eurecom.fr, tonyquek@sutd.edu.sg

***Abstract*—Semantic Communication (SeCom) has garnered widespread attention due to its effectiveness and intelligence as an emerging technique. The current image SeCom system, based on Deep Joint Source-Channel Coding (JSCC), requires joint model training at both ends, leading to coupled models that need concurrent updates. Additionally, the semantics transmitted via joint training are feature vectors, which lack interpretability. To address these issues, we design an Explainable Semantics-based Image Semantic Communication (ES-ISC) demo. This demo transforms images into explainable semantic texts and segmentation maps for transmission. By leveraging the universality of these semantic carriers, we facilitate the decoupling of transmitter (TX) and receiver (RX) training without substantially impacting performance. Moreover, the interpretability of the designed semantic carriers supports multiple downstream tasks. Experimental results demonstrate that our system can transmit images with over 100-fold compression while maintaining high-quality reconstruction at the RX.**

***Index Terms*—Semantic Communication, Explainable Semantics, Ultra-high compression rate, Image Transmission, hardware friendly**

I. Introduction

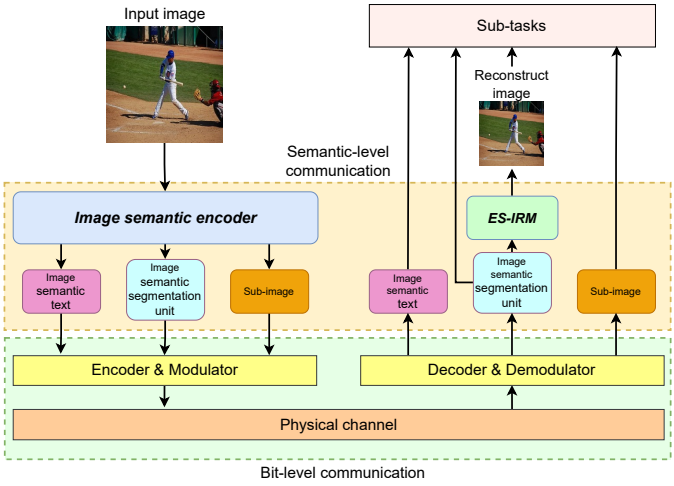
Semantic Communication (SeCom) emphasizes the semantics-level transmission between transmitter (TX) and receiver (RX), in contrast to standard digital communication systems that prioritize bit-level correctness. However, image semantic extraction faces substantial challenges due to its richness and ambiguity. For efficient image transmission, Eirina et al. proposed a Deep Joint Source-Channel Coding (JSCC) scheme by using a unified neural network to extract and encode image semantics [1]. This method requires joint model optimization at both the TX and RX, and its superiority in image reconstruction quality is verified by simulation experiments. However, it still faces the following significant limitations: i) the models at the TX and RX require joint training and simultaneous updates, which complicates deployment; ii) the semantic feature vectors generated by the existing scheme are uninterpretable, which hinders human understanding and extended applications; iii) the existing scheme assumes that any complex-valued signal can be transmitted in the channel, and directly maps the image into continuous signals for transmission, which is not only inconsistent with the hardware design, but also difficult to be compatible with the existing communication system. Therefore, we propose an Explainable Semantics-based Image Semantic Communication (ES-ISC) system that converts images into discrete explainable semantics for transmission using segmentation mapping and text extraction techniques. ES-ISC system maintains exceptionally high transmission efficiency while supporting a multitude of downstream functions. The discrete representational characteristics of these explainable semantics ensure their compatibility with modern digital communication systems. Our proposed ES-ISC demo is displayed in Fig. 1.

This work was supported in part by the National Key Research and Development Program of China under Grant 2024YFE0200300, and by the National Natural Science Foundation of China under Grant 62271513.

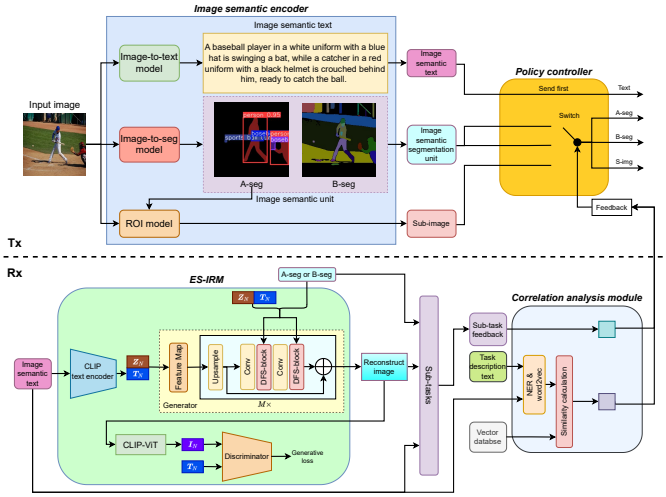
Fig. 1. An Explainable Semantics-based Image Semantic Communication (ES-ISC) System. This demo system runs on two host computers equipped with Ubuntu22.04 system and Nvidia RTX4060 graphics card. For on-site presentation, we need a desk to provide enough space for two host computers, two displays, keyboard and mouse, and also four three-legged sockets to power the computers and displays.

II. Technical Description

As illustrated in Fig. 2(a), the key components of the proposed ES-ISC demo are the image semantic encoder, channel encoder and modulator, channel decoder and demodulator, and Explainable Semantic Image Reconstruction Module (ES-RIM). At the TX, the image to be sent is converted using image-to-text and image semantic segmentation technologies into image semantic text, image semantic segmentation unit, and additional image subgraph. The channel encoder and modulator at the TX and the channel decoder and demodulator at the RX are consistent with those in modern digital



(a) Overall Architecture



(b) Design Details

Fig. 2. Architecture of the proposed ES-ISC demo, where A-seg and B-seg denote the segmentation maps based on semantic segmentation and based on Segment Anything, respectively, T_N and I_N denote the generated text and image feature vector, Z_N denotes the additional noise for generative model training, NER stands for Named Entity Recognition, and DFS-block stands for Deep text-image-segmentation Fusion Block.

communication systems. These modules convert the semantics extracted by the image semantic encoder into analog signals for transmission through the physical channel, and vice versa. As shown in Fig. 2(b), the implementation details are listed as follows:

- Image semantic encoder, mainly consists of one image-to-text model and one image-to-segmentation model. The role of the former is to generate semantic text corresponding to the input image, and the latter is used to generate two types of image semantic segmentation maps. It is worth mentioning that image semantic segmentation includes two key sub-modulator: i) Type-A segmentation maps (A-seg) based on semantic segmentation, primarily containing the category label, overall boundary and instance box of each segmentation object; ii) Type-B segmentation maps (B-seg) based on Segment Anything, mainly containing the detail boundaries of all segmentation instances. These two types of semantics can be directly understandable by humans and be used for inference tasks. To facilitate the image reconstruction task, image semantic encoder also extracts sub-images (S-img) corresponding to various objects in the Region of Interest (ROI) based on the semantic segmentation map in A-seg. Both semantic text and image semantic segmentation maps can be directly transmitted over physical channels using modern digital communication systems.
- Explainable Semantic Reconstruction Image Module (ES-RIM) is designed to ensure the high-quality image reconstruction at the RX and to fully utilize the two types of explainable semantics. It is based on Contrastive Language-Image Pre-training (CLIP) model [2] and Generative Adversarial CLIP model [3]. The core architecture of ES-RIM consists of one CLIP text encoder and its corresponding generator, as well as one CLIP image encoder (CLIP-ViT) and its corresponding discriminator. A key feature is the Deep text-image-segmentation Fusion Block (DFS-Block) within the generator, which integrates the segmentation map into the image reconstruction process at the feature map level. To enhance consistency between generated and original images, we incorporate a spatial normalization layer that considers segmentation map textures. This approach ensures that images generated from text maintain structural consistency with the original images.

III. Results and Applications

The proposed demo addresses key limitations of traditional Deep JSCC by transforming images into explainable semantic texts and semantic segmentation maps. This approach resolves the issues of non-decoupled models at the TX and RX, as well as the uninterpretable nature of transmitted semantics. The generated semantics offer two significant advantages: their discrete representation facilitates integration with current communication systems, and their universality allows for decoupled training of both ends without significantly compromising performance. Consequently, this demo supports a variety of downstream sub-tasks with extremely high transmission efficiency. Real-world experiments demonstrate the demo's capability to transmit images with a compression ratio exceeding 100 times, while simultaneously accomplishing multiple sub-tasks such as image captioning, semantic segmentation, and reconstruction. This performance represents a substantial improvement over conventional schemes.

References

[1] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. Cognitive Commun. Networking*, vol. 5, no. 3, pp. 567–579, 2019.

[2] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Machine Learning (ICML)*. PMLR, July 2021, pp. 8748–8763.

[3] M. Tao *et al.*, "Galip: Generative adversarial clips for text-to-image synthesis," in *Proc. Conf. Computer Vision Pattern Recognition (CVPR)*, June 2023, p. 14214 – 14223.