A Novel Lightweight Joint Source-Channel Coding Design in Semantic Communications

Xianhua Yu, Dong Li, *Senior Member, IEEE*, Ning Zhang, *Senior Member, IEEE*, and Xuemin Shen, *Fellow, IEEE*

***Abstract*—Semantic communication has emerged as a promising solution to meet the growing demand for efficient data transmission in the information age. Unlike traditional communication methods that focus on transmitting raw data, semantic communication prioritizes preserving the meaning of transmitted information, which significantly reduces the data volume. However, implementing semantic communication systems in resource-constrained environments, such as Internet of Things (IoT) devices, remains challenging due to limited computational resources. In this letter, we propose a novel lightweight deep learning (DL) model, termed the lightweight image compression and reconstruction network (LICRnet). LICRnet leverages depthwise separable convolution (DSC) and a local and nonlocal mixture (LNLM) block to significantly reduce computational costs. Additionally, the LNLM incorporates a variable window size-based multiscale attention mechanism (VW-MSA), enabling it to effectively learn from both local detailed features and global high-level meaningful features. Extensive simulations demonstrate that LICRnet significantly reduces computational complexity while maintaining satisfactory image compression and reconstruction performance, making it highly suitable for deployment in resource-constrained environments.**

***Index Terms*—Deep learning (DL), lightweight joint-source channel coding, semantic communication.**

I. Introduction

THE RAPID advancement of the information age has driven an increasing demand for large-scale data transmission, posing significant challenges for conventional communication systems and necessitating new technologies. Semantic communication has emerged as a promising solution, focusing on preserving the meaning of transmitted information rather than merely minimizing bit error rates. Originally conceptualized by Weaver in 1949, semantic communication operates at the semantic level, aiming to convey the intended meaning rather than replicate individual bits, unlike traditional bit-level communication. By reducing the data volume while maintaining the high semantic fidelity [1], semantic communication enhances transmission efficiency, positioning itself as a viable approach for next-generation communication systems. Recent advancements in deep learning (DL), particularly in areas such as natural language processing, image processing, and speech recognition, have significantly advanced semantic communication, enabling efficient human-to-machine and machine-to-machine interactions.

The limited power, memory, and processing capabilities of IoT devices make heavy computational models impractical [2]. These constraints are especially critical in real-world applications, such as smart cameras, wearable health monitors, and remote sensors, where low latency and energy efficiency are critical. As a result, there is a pressing need for lightweight semantic communication solutions that deliver high performance under strict resource constraints. Only a few studies [3], [4], [5], [6], [7] have specifically focused on minimizing computational costs. Techniques, such as knowledge distillation [4], [7], leverage high-complexity teacher models to guide simplified student models, allowing for efficient offline processing. In [3], a lightweight transformer-based model, DeLighT, was proposed to reduce computational costs in text semantic communication. Although DeLighT reduces part of the attention operations, it still calculates attention over the entire feature set, which may lead to high computational costs. Similarly, [6] incorporates the ConvNeXt model into the semantic encoder and decoder to alleviate the computation burden through parameters reduction. In [5], an attention-based UNet architecture is proposed to achieve a low computational complexity by reducing the number of downsampling layers. However, models in [5] and [6] may exhibit suboptimal performance due to the simplistic strategy of reducing the number of layers, as performance often correlates positively with the depth of the network.

However, the aforementioned works may face challenges in achieving a satisfactory performance while maintaining a low computational complexity. To address the limitations of existing semantic communication models in resource-constrained environments, we propose a novel lightweight DL model, termed lightweight image compression and reconstruction network (LICRnet). The key contributions of this letter are summarized as follows.

1) We propose LICRnet, a lightweight DL model tailored for semantic image compression in IoT scenarios. By employing depthwise separable convolutions (DSC) throughout the architecture, LICRnet significantly reduces the computational cost and the model size compared to traditional semantic communication models.
2) We design a new local and nonlocal mixture (LNLM) block that unifies lightweight local feature extraction

Received 16 January 2025; revised 27 March 2025; accepted 29 March 2025. Date of publication 1 April 2025; date of current version 23 May 2025. This work was supported in part by the Science and Technology Development Fund, Macau, SAR, under Grant 0188/2023/RIA3. *(Corresponding author: Dong Li.)*

Xianhua Yu and Dong Li are with the School of Computer Science and Engineering, Macau University of Science and Technology, Macau, China (e-mail: xianhuacn@foxmail.com; dli@must.edu.mo).

Ning Zhang is with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON N9B 3P4, Canada (e-mail: ning.zhang@ieee.org).

Xuemin Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: sshen@uwaterloo.ca).

Digital Object Identifier 10.1109/JIOT.2025.3556909

Fig. 1. Semantic communication system.

(via DSC) and global feature extraction through a variable window multiscale attention (VW-MSA) mechanism. Unlike prior works that use fixed or shifted window attention, our VW-MSA adaptively scales window sizes based on the feature resolution, enabling fine-grained detail extraction at high resolutions and global representation at low resolutions, all while significantly reducing the attention overhead. This architecture enhances semantic representation without sacrificing computational efficiency, making it uniquely suited for real-time applications.

3) Extensive simulations under realistic wireless channel conditions demonstrate that LICRnet achieves a superior balance between compression performance and computational efficiency, outperforming existing lightweight models such as LSCC and Cnet in terms of both PSNR/MS-SSIM and inference speed, while maintaining a much smaller model size.

II. System Model

In this letter, we consider a semantic communication system comprising a trainable encoder E_{θ_e}, an untrainable physical channel, and a trainable decoder D_{θ_d} as depicted in Fig. 1. The encoder and decoder parameters are represented by E_{θ_e} and D_{θ_d}, respectively. The entire transmission process is elaborated as follows. The image data x is fed into the encoder, which encodes it to semantic code (SC), given by

$$z = E_{\theta_e}(x, R) \tag{1}$$

where R denotes the compression ratio. After encoding, the SC is transmitted to the receiver through a physical channel, as given by

$$\bar{z} = hz + w \tag{2}$$

where h is the channel gain coefficient, and w represents the noise. The noise follows an i.i.d. circularly symmetric complex Gaussian (CSCG) distribution, with $\mathbf{w} \sim \mathcal{CN}(0, \sigma_w^2\mathbf{I})$, where σ_w^2 represents the noise power. Finally, the corrupted SC, denoted as $\bar{z}$, is received by the receiver and fed into the decoder, which outputs a reconstructed image $\hat{x}$ based on the corrupted SC, as given by

$$\hat{x} = E_{\theta_d}(\bar{z}, R). \tag{3}$$

III. Proposed Deep Learning Model

In this section, we introduce the architecture of LICRnet, emphasizing the novel contributions and detailing its lightweight design.

A. Preliminary Knowledge

DSC is composed of a depthwise layer and a pointwise layer. The depthwise layer uses a single kernel for each input feature map. DSC is a popular lightweight module used to reduce redundant operations in the standard convolution. The conversion from the standard convolution to DSC can be described as: $F^{\text{out}} = C(F^{\text{in}}) \approx P(D(F^{\text{in}}))$, where F^{out} represents the output features, C represents the standard convolution, F^{in} represents the input features, D stands for the depthwise convolution, and P stands for the pointwise convolution. Depthwise convolution is the main component responsible for processing the spatial information of the input features, requiring far fewer parameters and computation compared to standard convolution with the same kernel settings [12].

B. Architecture of the Proposed LICRnet

Before illustrating the process of LICRnet, we elaborate on the core components of the model: RDSCS blocks, LNLM blocks, and RDSCU blocks.

The RDSCS block is used in the encoder to extract features while reducing spatial resolution. It consists of two consecutive DSC layers followed by an average pooling layer for downsampling. A residual connection bypasses the core block to preserve input information and stabilize training. The RDSCU block is used in the decoder to gradually reconstruct the spatial resolution of the compressed features. It employs interpolation-based upsampling followed by two DSC layers, again with a residual connection.

The proposed LNLM block, shown in Fig. 2(a), replaces standard convolutions in traditional residual blocks [13] with a DSC layer to reduce the computational cost. The block incorporates a VW-MSA module, derived from MSA [9], which partitions features into resolution-dependent windows—smaller at high resolutions, larger at low resolutions—achieving efficient global context modeling with minimal overhead. LNLM consists of two DSC layers, a feature split, an RDSC block, a VW-MSA module, a concatenation layer, and residual connections. After the first DSC, features are split into two branches: one processes local features using RDSC, and the other captures the global context through VW-MSA. The outputs are fused through concatenation and a second DSC layer, followed by a skip connection. This design enables lightweight integration of local and nonlocal features for efficient semantic reconstruction.

The architecture of the proposed LICRnet is shown in Fig. 2(b). It consists of multiple RDSCS blocks, LNLM blocks, and RDSCU blocks. The compression ratio is (1/16), with each RDSCS block downsampling the feature to half the resolution, and each RDSCU block upsampling the feature to double the resolution. The encoder of LICRnet processes the image by initially downsampling it using the RDSCS block, followed by feature extraction using DSC blocks. The model then learns local and nonlocal information from the downsampled feature using the LNLM block. After repeating the RDSCS and LNLM blocks three times, the output feature is sent to the RDSCS block to generate the SC. The decoder of LICRnet functions similarly to the encoder, with the RDSCS blocks replaced by RDSCU blocks to upsample the features.

In the proposed VW-MSA, we utilize a variable window size to efficiently learn both local detailed features and global

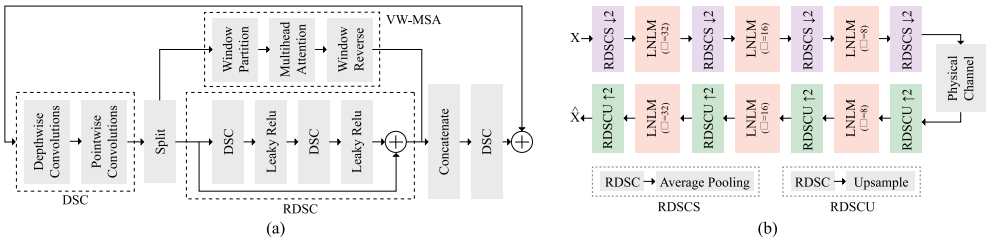

Fig. 2. (a) Architecture of the proposed LNLM. (b) Architecture of the proposed LICRnet where ↓, ↑, and □ denote downsampling, upsampling, and window size, respectively.

high-level features. Specifically, we apply a smaller window size at high resolution (coarse features) to capture fine, detailed information from the input image, and a larger window size at low resolution (fine features) to extract more abstract, high-level information. The finer, high-resolution features contain more detailed spatial information, which motivates the use of smaller windows at high resolution to learn these details more effectively. At lower resolutions, the larger window size helps capture global context and higher-level abstract features, ensuring that the VW-MSA effectively learns from both local detailed features and global, high-level meaningful features.

This variable window approach enables the VW-MSA to maintain a low computational cost while achieving a performance similar to that of the conventional W-MSA mechanism. In contrast to conventional W-MSA uses a fixed window size and facilitates feature interactions across different image partitions, VW-MSA avoids such costly operations by naturally leveraging multiresolution attention patterns. While it may omit fine-grained interactions at high resolutions, it compensates by enabling the network to learn global semantics more efficiently across scales. This leads to a significant reduction in computational complexity, while maintaining comparable reconstruction performance, as shown in Section IV. Furthermore, to mitigate the vanishing gradient problem, we introduce skip connections between the input and output of each RDSCS block, LNLM block, and RDSCU block. These additions help preserve gradient flow throughout the network and improve training stability.

C. LICRnet-Based Number Compression and Reconstruction Framework

The LICRnet-based compression and reconstruction framework comprises two phases: 1) the offline training phase and 2) the online compression and reconstruction phase.

1) Offline Training Phase: Given a training data set

$$(\mathcal{R}) = \left\{\left(\mathbf{r}^{(1)}\right), \left(\mathbf{r}^{(2)}\right), \ldots, \left(\mathbf{r}^{(n)}\right)\right\} \tag{4}$$

where $(\mathbf{r}^{(n)})$, $n \in 1, \ldots, N$ is the nth training sample of $(\mathcal{R})$. $\mathbf{r}^{(n)}$ is the input image and also serves as the ground truth. We select the mean square error (MSE) as our cost function due to its widespread use and proven effectiveness in image reconstruction tasks, which can be expressed as

$$L_{MSE}\left(E_{\theta_e}, D_{\theta_d}\right) = \frac{1}{N}\sum_{n=1}^{N}\|\mathbf{r}^{(n)} - \hat{\mathbf{r}}^{(n)}\|^2 \tag{5}$$

where $\hat{\mathbf{r}}^{(n)}$ is the reconstructed image based on the received noisy SC. Additionally, we utilize the backpropagation (BP) algorithm to progressively update the network parameters to obtain a well-trained model.

2) Online Testing Phase: Given a test dataset $\dot{\mathbf{R}}$, it is sent to the well-trained LICRnet as input to obtain the set of reconstructed images $\hat{\mathbf{R}}$

$$\hat{\mathbf{R}} = f_{\theta_e,\theta_d}\left(\dot{\mathbf{R}}\right). \tag{6}$$

IV. Simulation Results

A. Training Procedure

We adopt an end-to-end training approach for LICRnet using a combined dataset consisting of 8,000 high-resolution images (larger than 256×256) from ImageNet [14] and the entire CLIC dataset, both commonly used in image compression research. The model is trained for 200 epochs with a batch size of 64. We use the Adam optimizer due to its fast convergence and low memory requirements, which align well with our lightweight design. The initial learning rate is set to 1×10^{-4} and adjusted using a cosine annealing schedule to improve the training stability and convergence. Additionally, we simulate a realistic wireless channel with distance-dependent path loss and Rician fading (K-factor = 2.8) to ensure the robustness under practical IoT conditions.

B. Performance Analysis

We evaluate the LICRnet against two state-of-the-art lightweight semantic communication models, LSCC [3] and Cnet [6],[1] along with an internal variant, LICRnet-F, which replaces the proposed VW-MSA with conventional W-MSA. All models are evaluated on the Kodak dataset under a fixed compression ratio of (1/16), using identical training and evaluation settings for fairness.

We adopt peak signal-to-noise ratio (PSNR) and multiscale structural similarity (MS-SSIM) as the primary evaluation metrics. PSNR is a standard measure of pixel-level fidelity, while MS-SSIM reflects human visual perception and MS-SSIM. In semantic image communication, the goal is to preserve both low-level detail and high-level perceptual structure. These two metrics complement each other in evaluating the semantic and

[1]As prior studies [3], [6] have shown that semantic communication models consistently outperform traditional source-channel coding methods in terms of the reconstruction quality under similar conditions, this letter focuses on comparing LICRnet with recent semantic communication models only.

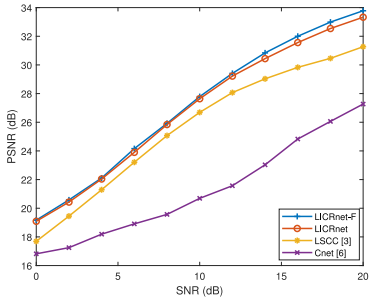

Fig. 3. PSNR performance.

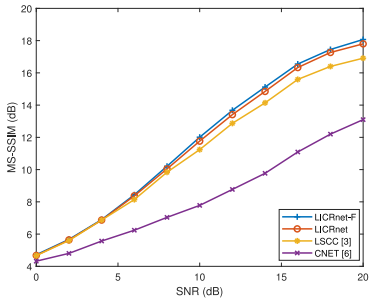

Fig. 4. MS-SSIM performance.

TABLE I
COMPARISON OF COMPUTATIONAL COMPLEXITY AND MODEL SIZE

Model	Parameter size	Processing time
LICRnet-F	289.54MB	0.217ms
LICRnet	160.39MB	0.122ms
LSCC [3]	280.91MB	0.132ms
Cnet [6]	281.45MB	0.135ms

perceptual quality of reconstruction and are widely adopted in recent works.

As shown in Figs. 3 and 4, LICRnet consistently outperforms LSCC and Cnet across various SNR levels. While LICRnet-F provides slightly better reconstruction results, the gap is marginal, validating that our proposed VW-MSA achieves competitive performance at significantly lower computational cost. This result of LICRnet can be attributed to the effectiveness of the LNLM block, which integrates local feature extraction through DSC and nonlocal features through VW-MSA. This hybrid design enables the model to capture both fine-grained spatial detail and broader semantic context, which is essential for reconstructing perceptually meaningful images under compression. The ablation comparison with LICRnet-F results in a small margin, further demonstrating that adaptive window resizing in VW-MSA is sufficient to preserve global context, validating the efficiency-oriented attention design.

To evaluate the computational complexity of the proposed LICRnet, we compare the model parameter size and inference time with other state-of-the-art methods, as shown in Table I. LICRnet achieves the lowest processing time and a smaller parameter size compared to conventional models, such as LSCC and Cnet, demonstrating its computational efficiency. This reduction is primarily attributed to the use of DSC, which decomposes standard convolutions into depthwise and pointwise operations. This design significantly reduces the number of parameters and multiplications required. Furthermore, the VW-MSA replaces the conventional shifted window attention with an adaptive attention mechanism that adjusts window sizes according to feature resolution, thereby lowering the computational burden at higher spatial scales.

These results confirm that LICRnet achieves a favorable balance between the reconstruction quality and the computational efficiency, making it well-suited for deployment in real-time, resource-constrained semantic communication systems, particularly for IoT applications.

V. CONCLUSION

In this letter, we proposed a novel lightweight method, termed LICRnet, to address the challenge of limited computational resources in IoT devices in semantic communication. In LICRnet, we replace conventional CNN layers with DSC blocks to reduce computational costs. Additionally, we introduce a novel LNLM block that captures both local and nonlocal features and utilizes VW-MSA with variable window sizes to further minimize computational complexity while enhancing model performance. Simulation results demonstrate that LICRnet achieves high-quality image compression and reconstruction performance while significantly lowering computational complexity, validating its suitability for deployment in resource-constrained environments such as IoT systems.

REFERENCES

[1] C. Liang, D. Li, Z. Lin, and H. Cao, "Selection-based image generation for semantic communication system," *IEEE Commun. Lett.*, vol. 28, no. 1, pp. 34–38, Jan. 2024.
[2] I. Ullah et al., *Future Communication Systems Using Artificial Intelligence: Internet of Things and Data Science*. Boca Raton, FL, USA: CRC Press, 2024.
[3] Y. Jia, Z. Huang, K. Luo, and W. Wen, "Lightweight joint source-channel coding for semantic communications," *IEEE Commun. Lett.*, vol. 27, no. 12, pp. 3161–3165, Dec. 2023.
[4] D. Ye, X. Wang, and X. Chen, "Lightweight generative joint source-channel coding for semantic image transmission with compressed conditional GANs," in *Proc. ICCC Workshops*, 2023, pp. 1–6.
[5] G. Ma, H. Tong, N. Yang, and C. Yin, "Attention-based UNet enabled Lightweight Image Semantic Communication System over Internet of Things," in *Proc. WCNC*, 2024, pp. 1–6.
[6] G. Chen et al., "Lightweight and robust wireless semantic communications," *IEEE Commun. Lett.*, vol. 28, no. 11, pp. 2633–2637, Nov. 2024, doi: 10.1109/LCOMM.2024.3409559.
[7] C. Liu, Y. Zhou, Y. Chen, and S.-H. Yang, "Knowledge distillation-based semantic communications for multiple users," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 7000–7012, Jul. 2024.
[8] B. Sun, Y. Zhang, S. Jiang, and Y. Fu, "Hybrid pixel-unshuffled network for lightweight image super-resolution," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, 2023, pp. 2375–2383.
[9] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
[10] J. Liu, H. Sun, and J. Katto, "Learned image compression with mixed transformer-CNN architectures," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14388–14397.
[11] X. Yu and D. Li, "Phase shift compression for control signaling reduction in IRS-aided wireless systems: Global attention and lightweight design," *IEEE Trans. Wireless Commun.*, vol. 23, no. 8, pp. 8528–8541, Aug. 2024.
[12] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1251–1258.
[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.