# Joint ROI Guidance and Spatial Analysis for Task-Aware Distributed Deep Joint Source-Channel Coding

Wenkai Tian†, Biao Dong†, and Bin Cao†‡
†School of Electronics and Information Engineering, Harbin Institute Of Technology, Shenzhen, China
‡Research Center of Networks and Communications, Peng Cheng Laboratory, Shenzhen, China
Email: 22s152098@stu.hit.edu.cn, 23b952012@stu.hit.edu.cn, caobin@hit.edu.cn

***Abstract*—In this paper, we investigate the system performance of deep joint source-channel coding (JSCC) for task-oriented transmission in the Wyner-Ziv scenario, i.e., a distributed coding scenario, aiming to improve the image reconstruction performance and task accuracy. Unlike existing deep JSCC based methods, we introduce regions of interest (ROI), which facilitates the effective utilization of side information for enhancing task performance. Meanwhile, we incorporate a spatial analysis mechanism to fuse the side information. By integrating these two mechanisms, we propose a novel distributed deep JSCC scheme that further leverages task relevance within the side information. Simulation results show that our proposed scheme outperforms the benchmark in terms of image reconstruction performance and task accuracy. The code is available on the project website[1].**

## I. Introduction

The rapid development of the Internet of Things (IoT) [1] and edge computing [2] poses challenges to traditional communication systems. To support the explosive growth of interconnected smart devices and artificial intelligence (AI) services, the sixth-generation communications (6G) must meet demanding requirements, including low latency, multi-user support, and intelligence. In recent times, semantic communications have emerged as a promising technology to address these challenges. Unlike traditional bit-level communication systems, semantic communications extract, transmit, and utilize the semantic features of information at the semantic level. With its exceptional feature extraction capabilities, deep learning (DL) provides robust support for the design of semantic communications.

Recently, numerous semantic communication technologies based on deep learning (DL) have been developed. Among these, joint source-channel coding technology has sparked broad interest and research enthusiasm [3]. By using an autoencoder architecture to extract compact representations, deep joint source-channel coding (DeepJSCC) can improve content awareness and enhance compression performance [4]. In point-to-point communication scenarios, many mature schemes have been proposed for channel adaptation [5], adaptive rate allocation [6], and multi-task semantic communication [7]. In addition, Yilmaz *et al.* [8] studied a distributed communication scenario, and presented a deep joint source-channel coding (JSCC) communication framework for the Wyner-Ziv scenario (DeepJSCC-WZ) [9], wherein only the decoder uses side information for image decoding. By exploiting the correlation among correlated sources, DeepJSCC-WZ employs multi-scale feature fusion at the decoder, achieving enhanced image reconstruction performance. However, DeepJSCC-WZ utilizes side information in a relatively coarse manner, resulting in limited performance improvement. To fully exploit the advantages of distributed communication systems, the design and utilization of side information have become critical questions.

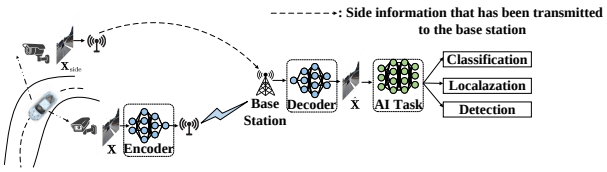



Fig. 1: Task-oriented deep JSCC scheme with side information(the Wyner-Ziv scenario).

It is noticed that, task-oriented semantic communications (TOSC), which focus on extracting and transmitting essential information for downstream tasks, have attracted growing attention from both academia and industry [3], [7], [10]. Tong *et al.* [11] introduced the information bottleneck principle into TOSC, shifting the focus from image reconstruction to the execution of semantic tasks. Hu *et al.* [7] proposed prioritizing the transmission of more important semantic features by ranking their importance, improving the system's scalability under varying channel conditions. Building upon the advanced deep JSCC architecture [6], Tan *et al.* [12] utilized regions of interest (ROI) to adapt the entropy model and enhance task performance, albeit at the cost of increased encoder complexity. However, these works solely focus on improving task performance in point-to-point communication scenarios, without exploring the utilization of correlated sources in distributed scenarios.

In this paper, we propose a kind of ROI assisted deep JSCC scheme for image reconstruction and semantic tasks in Wyner-Ziv scenarios (see Fig. 1). The main contributions of this work are summarized as follows:

[1]https://github.com/tianwenkye/DeepJSCC-WZ-ROI

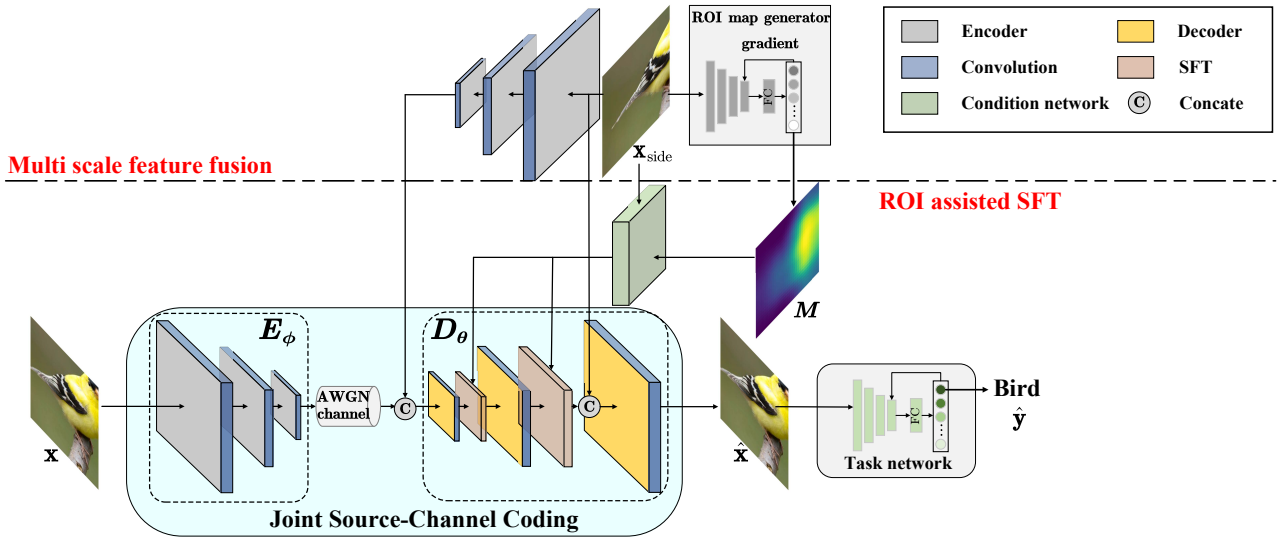



Fig. 2: Architecture of our proposed system of ROI-assisted distribution deep JSCC. The dashed line above represents the multi-scale feature fusion proposed in [8], and the dashed line below represents the proposed ROI assisted SFT modules.

- A novel distributed deep JSCC scheme is proposed by introducing ROI maps and the Spatial Feature Transform (SFT), achieving a more refined fusion of side information and enhancing image reconstruction performance.
- The proposed scheme leverages the SFT module to fuse the task-related information embedded in ROI maps, guiding image reconstruction and further improving task performance.

## II. System Model

We consider the problem of task-oriented wireless image transmission with side information only at the decoder. As shown in Fig. 1, the transmitter consists of a stereo camera, where the left and right cameras transmit highly correlated images to the base station, which subsequently engages in machine tasks. To align with prior works [8], [13], [14], we assume that the image from the left camera is firstly transmitted, and the base station uses this received image to decode the image from the right camera.

The input image from the right camera, denoted as $\mathbf{x} \in \mathbb{R}^{C\times W\times H}$, where $W$, $H$ and $C$ represent the width, height and the number of channels of images, respectively, is first encoded by a neural network encoder $E_\phi$ as $\mathbf{z} \in \mathbb{C}^k$, where $\phi$ denotes the parameters of the encoder and $k$ is the length of the encoded symbols, as given by

$$\mathbf{z} = E_\phi(\mathbf{x}). \tag{1}$$

In this work, we consider the system is power constrained, hence we perform power normalization on the encoded symbols $\mathbf{z}$:

$$\frac{1}{k}\|\mathbf{z}\|_2^2 \le P_{\text{avg}}, \tag{2}$$

where $P_{\text{avg}}$ denotes the average power of the encoded symbols. The bandwidth ratio (BR) is defined as follows:

$$\rho \triangleq \frac{k}{CWH} \quad \text{channel symbols/pixel.} \tag{3}$$

The channel inputs are then transmitted over an additive white Gaussian noise (AWGN) channel with noise variance $\sigma^2$, i.e., $\tilde{\mathbf{z}} = \mathbf{z} + \mathbf{n}$, where $\mathbf{n} \in \mathbb{C}^k$ consists of independent and identically distributed (i.i.d) samples with the distribution $\mathcal{CN}\left(0, \sigma^2\mathbf{I}\right)$.

The receiver then uses the side information to jointly decode the received signals as follows:

$$\hat{\mathbf{x}} = D_\theta(\tilde{\mathbf{z}}, \mathbf{x}_{\text{side}}), \tag{4}$$

where $\hat{\mathbf{x}} \in \mathbb{R}^{C\times W\times H}$ is the estimation of the input images $\mathbf{x}$, $\theta$ denotes the parameters of the decoder, and $\mathbf{x}_{\text{side}} \in \mathbb{R}^{C\times W\times H}$ is the side information, which is highly correlated with $\mathbf{x}$ and has been transmitted to the receiver. Finally, the decoded image is sent to a task execution network to obtain the final predicted output:

$$\hat{\mathbf{y}} = F_\eta(\hat{\mathbf{x}}), \tag{5}$$

where $F_\eta$ is the task execute network with parameters $\eta$, and $\hat{\mathbf{y}}$ denotes the task execution result of the receiver.

## III. Proposed Method

In this section, we present the proposed distributed deep JSCC scheme. We first describe the overall structure of the architecture. Next, we introduce the ROI map generation method and the SFT module [15].

### A. Overview

Our DNN architecture is based on the DeepJSCC-WZ model proposed in [8]. As shown in Fig. 2, the distributed deep JSCC structure employed in our system is similar to DeepJSCC-WZ, both utilizing an autoencoder-based architecture. With its extensive set of parameters, the deep JSCC framework can learn a compact representation of the input images. The encoder maps the input image to a low-dimensional latent vector, while the decoder reconstructs the original image from this latent representation. Through this process, the network learns a compressed representation aimed at preserving essential semantic information from the input image.

The encoder consists of three convolutional blocks, and performs two downsampling operations. Each convolutional block includes a convolutional layer with a kernel size of $5 \times 5$ and $256$ channels, followed by the generalized divisive normalization function and a PReLU activation layer. The decoder is built with three deconvolutional blocks and performs two upsampling operations. Meanwhile, the work in [8] incorporates a multi-scale feature fusion structure, where features from $\mathbf{x}_{\text{side}}$ are extracted using an extractor that shares parameters with the encoder $E_\phi$. These extracted features are then strategically integrated with the received compressed features at various stages and scaled along the channel dimension. In this context, we adopt the similar approach and further explore the utilization of side information by introducing the ROI map and SFT modules.

### B. ROI Map Generation Mechanism

Inspired by Gradient-weighted Class Activation Mapping (Grad-CAM) [16], we extract ROI maps from side information to guide the decoder in reconstructing of the image. The idea behind Grad-CAM is to use gradient information activated within the network to quantify the importance of each pixel for the target concept through a weighted approach. In a well-trained classification network, Grad-CAM generates an ROI map by backpropagating gradients related to the target concept. This map indicates which regions of the image contribute significantly to the classification task. Next, we describe the ROI map generation process.

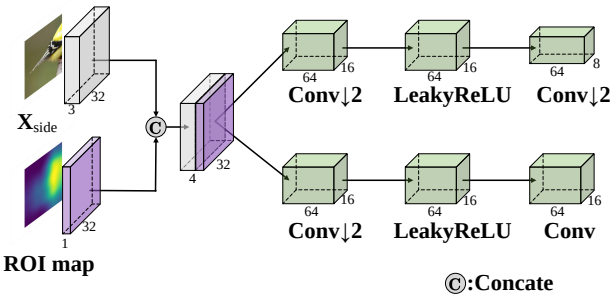
First, we feed the image into a pre-trained classification network to obtain the highest score $o^c$, where $c$ denotes the class $c$. We then calculate the gradients of the activation feature maps $\boldsymbol{a}^l$ in the last layer with respect to $o^c$ as $\frac{\partial o^c}{\partial \boldsymbol{a}^l}$, where $l$ denotes the $l$th feature map. Next, we perform global average pooling on the backpropagated gradients across the channel dimension to obtain the weights $w_l^c$ for the feature maps as follows:

$$w_l^c = \frac{1}{W \times H} \sum_i \sum_j \frac{\partial o^c}{\partial \boldsymbol{a}_{ij}^l}, \tag{6}$$

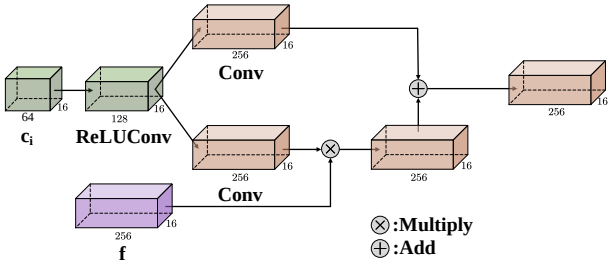
where $W$ and $H$ represent the width and the height of the feature maps, respectively. This weight signifies the relative contribution of different feature maps to the final classification decision. Using the obtained weights and the feature maps themselves, we perform a weighted combination and apply a ReLU layer to obtain the ROI map $\boldsymbol{M}$:

$$\boldsymbol{M} = \text{ReLU}\left(\sum_l w_l^c \boldsymbol{a}^l\right). \tag{7}$$

Finally, the ROI map undergoes image interpolation to restore its scale to match the original image size. Fig. 2 provides a sample ROI map generated by Grad-CAM from the side information $\mathbf{x}_{\text{side}}$. Despite transformations applied to the image, the ROI map still captures the approximate location of pixels corresponding to the target.




(a) Condition network




(b) SFT layer

Fig. 3: Illustration of the Condition network and Spatial Feature Transform (SFT) layer [15]. ↓ indicates down-sampling.

### C. Spatial Feature Transform Mechanism

To further leverage the side information at the decoder, we introduce the Spatial Feature Transform (SFT) module [15]. The SFT module is derived from adaptive feature transformation, initially proposed for style transfer [17]. Unlike traditional methods that rely on normalization of intermediate feature maps, SFT employs direct affine transformations [18]. This approach enhances the adaptability and flexibility of feature extraction in response to different inputs or contexts.

The introduced SFT module is illustrated in Fig. 3, where the Condition network utilizes external priors to generate appropriate inputs for the SFT layer. Initially, ROI map $\boldsymbol{M}$ is extracted from the side information using Grad-CAM. After concatenating $\boldsymbol{M}$ with the side information $\mathbf{x}_{\text{side}}$ along the channel dimension, the combined input is fed into the Condition network to obtain intermediate conditional semantic features $\mathbf{c_i}$ at different scales, where $\mathbf{c_i}$ denotes the $i$th condition

semantic features. The intermediate features are then processed by distinct SFT layers to apply scale adjustments to the intermediate representations of the decoder. Each SFT layer uses $\mathbf{c_i}$ to generate a set of affine transformation parameters ($\gamma$, $\beta$). These parameters are applied to each element in the intermediate feature maps $\mathbf{f}$ produced by the decoder. Finally, an affine transformation is applied to the intermediate feature maps $\mathbf{f}$ based on ($\gamma$, $\beta$), given by:

$$\mathrm{SFT}(\mathbf{f}, \mathbf{c_i}) = \gamma \odot \mathbf{f} + \beta, \tag{8}$$

where $\odot$ denotes element-wise multiplication.

**Algorithm 1** Image reconstruction training algorithm

**Input:** An image dataset $\{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$ with $n$ images, side information dataset $\{\mathbf{x}^1_{\text{side}}, \ldots, \mathbf{x}^n_{\text{side}}\}$.
**Output:** The parameterized networks $E_\phi$, $D_\theta$.

Initialize encoder and decoder's parameters $W(\phi$ and $\theta)$;
ROI map generation: $\mathbf{x}_{\text{side}} \xrightarrow{(6)(7)} \boldsymbol{M}$;
**while** Stop criterion is not met **do**
  **Forward:** $\mathbf{x} \xrightarrow{E_\phi(\cdot)} \mathbf{z} \xrightarrow{channel} \tilde{\mathbf{z}} \xrightarrow{D_\theta(\cdot, \mathbf{x}_{\text{side}})} \hat{\mathbf{x}}$;
  **Loss:** $\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}) = \sum_{(\mathbf{x}, \mathbf{x}_{\text{side}}) \in \mathcal{D}_{\text{train}}} \mathrm{MSE}(\mathbf{x}, \hat{\mathbf{x}})$;
  **Back-propagation:** $\mathcal{L} \to \frac{\partial \mathcal{L}}{\partial W}$;
  **Update parameters:** $W = W - lr \frac{\partial \mathcal{L}}{\partial W}$;
**end while**

### D. Training Loss

To validate the effectiveness of the proposed structure for both image reconstruction and task execution, we evaluate our algorithm from two aspects: Algorithm 1 and Algorithm 2.

Algorithm 1 illustrates the entire training process for image reconstruction. It undergoes end-to-end training in an unsupervised manner, generating AWGN channel samples randomly throughout the training process. The objective is to reconstruct the input image by minimizing a specified distortion measure. The loss function for Algorithm 1 is as follows:

$$\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}) = \sum_{(\mathbf{x}, \mathbf{x}_{\text{side}}) \in \mathcal{D}_{\text{train}}} \mathrm{MSE}(\mathbf{x}, \hat{\mathbf{x}}), \tag{9}$$

where $\mathrm{MSE}(\mathbf{x}, \hat{\mathbf{x}}) \triangleq \frac{1}{m} \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2$ is the mean squared error (MSE) loss and $m = CWH$.

Algorithm 2 is designed for task-oriented semantic communication scenarios, where the reconstructed images are further sent to a task execution network, as shown in Fig. 1. The evaluation metrics encompass both the distortion in image reconstruction and the accuracy of the task. Therefore, following the reasoning in [11], we design a new loss function that combines weighted losses, as follows:

$$\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}) = \sum_{(\mathbf{x}, \mathbf{x}_{\text{side}}) \in \mathcal{D}_{\text{train}}} \mathrm{MSE}(\mathbf{x}, \hat{\mathbf{x}}) + \lambda \cdot \mathrm{CE}(\mathbf{y}, \hat{\mathbf{y}}), \tag{10}$$

where $\mathbf{y}$ denotes the label corresponding to the image $\mathbf{x}$, $\mathrm{CE}(\mathbf{y}, \hat{\mathbf{y}})$ represents the cross-entropy loss, and $\lambda$ controls the trade-off between reconstruction quality and task performance. Details regarding the specific training procedure are provided in Algorithm 2.

**Algorithm 2** TOSC training algorithm

**Input:** An image dataset $\{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$ with $n$ images, side information dataset $\{\mathbf{x}^1_{\text{side}}, \ldots, \mathbf{x}^n_{\text{side}}\}$.
**Output:** The parameterized networks $E_\phi$, $D_\theta$.

Train a task network $F_\eta$ with the input dataset;
ROI map generation: $\mathbf{x}_{\text{side}} \xrightarrow{(6)(7)} \boldsymbol{M}$;
Train $E_\phi$, $D_\theta$ using Algorithm 1;
Load pre-trained network's parameters $\phi, \theta$ ($W$) and $\eta$, freeze $\eta$;
**while** Stop criterion is not met **do**
  **Forward:** $\mathbf{x} \xrightarrow{E_\phi(\cdot)} \mathbf{z} \xrightarrow{channel} \tilde{\mathbf{z}} \xrightarrow{D_\theta(\cdot, \mathbf{x}_{\text{side}})} \hat{\mathbf{x}} \xrightarrow{F_\eta(\cdot)} \hat{\mathbf{y}}$;
  **Loss:** $\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}) = \sum_{(\mathbf{x}, \mathbf{x}_{\text{side}}) \in \mathcal{D}_{\text{train}}} \mathrm{MSE}(\mathbf{x}, \hat{\mathbf{x}}) + \lambda \cdot \mathrm{CE}(\mathbf{y}, \hat{\mathbf{y}})$;
  **Back-propagation:** $\mathcal{L} \to \frac{\partial \mathcal{L}}{\partial W}$;
  **Update parameters:** $W = W - lr \frac{\partial \mathcal{L}}{\partial W}$;
**end while**

## IV. Numerical Results

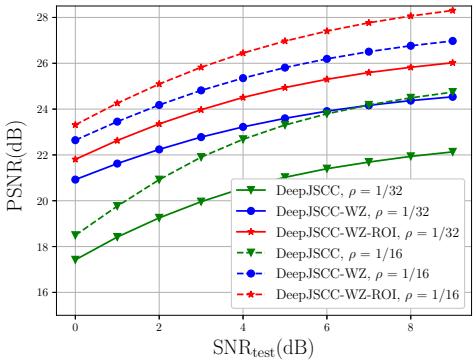
In this section, we conduct a series of experiments based on the two algorithms proposed in the previous section, focusing on two performance metrics: image reconstruction quality and task accuracy.

### A. Settings

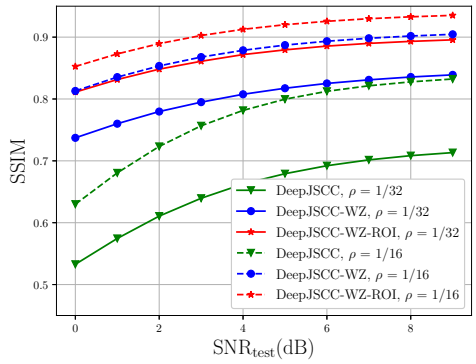
*1) Dataset:* We perform the experiments on the CIFAR-10 dataset, which comprises 50,000 training images and an additional 10,000 validation images, each with a size of $32 \times 32$ pixels, across a total of 10 image classes. To generate correlated images to serve as side information, we adopt a method for constructing correlated sources similar to that described in [19].

*2) Metrics:* We employ a classification task as the artificial intelligence task, using ResNet50 [20] as the classification network. The performance of this task is measured by classification accuracy. The quality of reconstruction is assessed using peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM), a metric more aligned with human perception of images.

*3) Benchmarks:* Firstly, in terms of image reconstruction, we use the proposed ROI-assisted distributed deep JSCC (marked as DeepJSCC-WZ-ROI) as the baseline, and compare it with methods including DeepJSCC [4] and DeepJSCC-WZ [8]. All three approaches are trained using Algorithm 1. Secondly, for task-oriented semantic communications, the methods comppared include DeepJSCC and DeepJSCC-WZ which are trained using Algorithm 2. In the ablation study, the models DeepJSCC-WZ-zeros-ROI and DeepJSCC-WZ-norms-ROI are trained using the proposed method. It needs to be pointed out that, during the testing phase, the decoder receives ROI maps that consist entirely of zeros or random numbers following a normal distribution. DeepJSCC-WZ-SFT is trained using the structure shown in Fig. 2, where the input




(a) PSNR.




(b) SSIM.

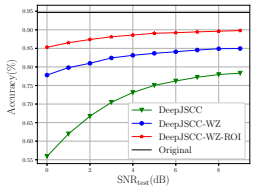
Fig. 4: Comparisons of different methods on different evaluation indicators in image reconstruction scenarios for BR $\rho = 1/32$ and $1/16$.

to the Conditional network is only the side information $\mathbf{x}_{\text{side}}$, meaning that no ROI map is used during training.

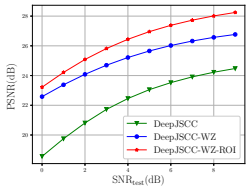
*4) Implementation Details:* The learning rate is set to $1 \times 10^{-4}$. The batch size is 64. The average power constraint $P_{\text{avg}}$ is set to 1.0. The Adam optimizer is employed to minimize the training loss in (9) and (10). The signal-to-noise ratio (SNR) of the AWGN channel is maintained at 5 dB during training, while the test SNR ranges from 0 to 9 dB. The trade-off factor $\lambda$ is set to $1 \times 10^{-3}$.

### *B. Experiment Results and Analysis*

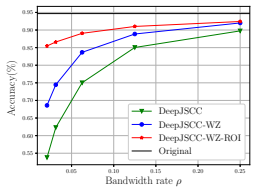
*1) Reconstruction Performance:* Fig. 4 illustrates the image reconstruction performance under various SNR conditions. From Fig. 4(a), it can be observed that the proposed scheme achieves a performance gain in terms of PSNR. The improvement of the DeepJSCC-WZ over DeepJSCC is attributed to the utilization of correlated information from the side information. The performance gain of the proposed scheme over DeepJSCC-WZ is partly due to the introduction of the ROI map, which captures part of the original image's contour information. Moreover, the SFT module provides more delicate adjustments to the intermediate features of the decoder. In Fig. 4(b), it is evident that the proposed approach not only



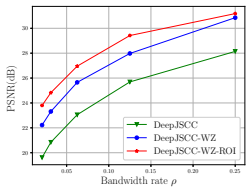

(a) Accuracy for BR $\rho = 1/16$.




(b) PSNR for BR $\rho = 1/16$.




(c) Accuracy for SNR = 5dB.




(d) PSNR for SNR = 5dB.

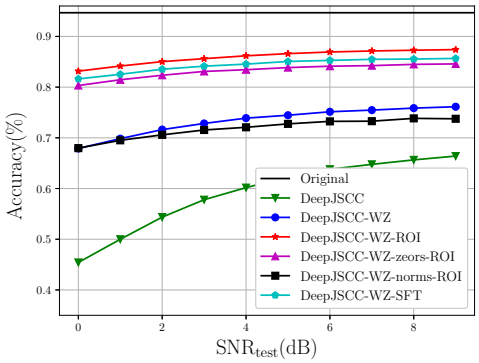
Fig. 5: Comparisons of different methods on classification and reconstruction in TOSC scenarios.

enhances PSNR performance but also improves metrics that better align with human perception of image quality.

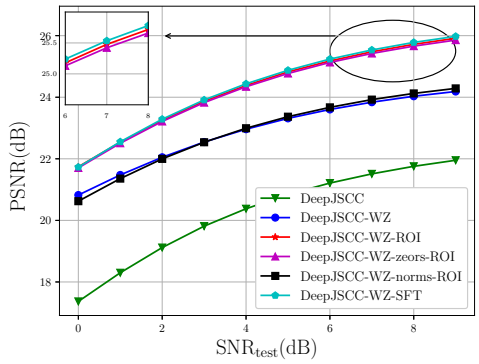
*2) Task Execution Performance:* Fig. 5 depicts the variation in task performance and image reconstruction performance of the different methods across various SNR conditions and BR values. From Fig. 5(a), it can be observed that DeepJSCC-WZ-ROI provides a significant improvement in task performance compared to the other schemes. DeepJSCC-WZ shows an enhancement in task performance over DeepJSCC, which is attributed to the use of side information, enhancing image reconstruction quality. DeepJSCC-WZ-ROI further improves task performance compared to DeepJSCC-WZ, highlighting the effectiveness of incorporating ROI maps and the SFT module. Fig. 5(b) illustrates that the reconstruction quality of the proposed scheme still surpasses that of other methods in TOSC scenarios.

Fig. 5(c) and Fig. 5(d) demonstrate that the proposed scheme achieves higher compression performance, especially at lower BR values, where the performance gains are more noticeable. As the BR values increase, both image reconstruction and task execution performance tend to approach saturation due to the increased amount of transmitted information.

*3) Ablation Study:* Fig. 6 presents ablation experiments to validate the effectiveness of introducing the ROI map and SFT modules. Firstly, Fig. 6(a) shows that DeepJSCC-WZ-zeros-ROI results in a decrease in task performance. This decline is attributed to the absence of a meaningful ROI map. DeepJSCC-WZ-norms-ROI causes performance to degrade to the level of DeepJSCC-WZ. This is because the incorrectly distributed ROI map leads the SFT layers to guide the decoder's intermediate features in a direction that is detrimental to performance. Additionally, Fig. 6(b) reveals that introducing an all-zero ROI map results in only a minimal decline in PSNR performance, contrasting with the more significant decrease in task performance. This observation further underscores the importance of the ROI map obtained from Grad-CAM for task




(a) Test Accuracy.




(b) PSNR.

Fig. 6: Ablation study results for BR $\rho = 1/32$.

relevance.

Based on the two figures, DeepJSCC-WZ-SFT exhibits slightly better image reconstruction performance compared to DeepJSCC-WZ-ROI, and there is a relatively noticeable difference in task accuracy. This indicates that the SFT modules achieve a more precise fusion of side information, thereby enhancing image reconstruction quality. Additionally, by further introducing ROI maps as an external condition for the SFT module, the task performance can be further improved. These results indicate that the performance gains of our proposed scheme stem from the integration of the ROI map and SFT module.

## V. Conclusion

In this paper, we devise a ROI-assisted deep JSCC scheme for machine task image transmission in distributed scenarios. To more effectively utilize the correlation among correlated sources and better serve downstream tasks, our approach incorporates side information and employs ROI maps along with SFT modules. This combination ensures superior image reconstruction quality and enhances task accuracy. Experimental results confirm the effectiveness of our method in achieving nuanced information fusion for image reconstruction and improving task performance in TOSC scenarios. We demonstrate its robustness and flexibility, and highlight its potential for diverse applications in distributed communication scenarios.

## References


[1] H. Allioui and Y. Mourdi, "Exploring the full potentials of IoT for better financial growth and stability: A comprehensive survey," *Sensors*, vol. 23, no. 19, pp. 8015–8083, Sep. 2023.

[2] H. Hua, Y. Li, T. Wang, N. Dong, W. Li, and J. Cao, "Edge computing with artificial intelligence: A machine learning perspective," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, Jan. 2023.

[3] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE J. Select. Areas Commun.*, vol. 41, no. 1, pp. 5–41, Jan. 2022.

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

[5] J. Xu, B. Ai, W. Chen, A. Yang, P. Sun, and M. Rodrigues, "Wireless image transmission using deep source channel coding with attention modules," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2315–2328, Apr. 2022.

[6] J. Dai, S. Wang, K. Tan, Z. Si, X. Qin, K. Niu, and P. Zhang, "Nonlinear transform source-channel coding for semantic communications," *IEEE J. Select. Areas Commun.*, vol. 40, no. 8, pp. 2300–2316, Aug. 2022.

[7] J. Hu, F. Wang, W. Xu, H. Gao, and P. Zhang, "Scalable multi-task semantic communication system with feature importance ranking," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2023, pp. 1–5.

[8] S. F. Yilmaz, E. Özyilkan, D. Gündüz, and E. Erkip, "Distributed deep joint source-channel coding with decoder-only side information," *arXiv preprint arXiv:2310.04311*, 2023.

[9] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 1, pp. 1–10, Jan. 1976.

[10] S. Xie, S. Ma, M. Ding, Y. Shi, M. Tang, and Y. Wu, "Robust information bottleneck for task-oriented communication with digital modulation," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2577–2591, Aug. 2023.

[11] W. Tong, F. Liu, Z. Sun, Y. Yang, and C. Guo, "Image semantic communications: An extended rate-distortion theory based scheme," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2022, pp. 1723–1728.

[12] K. Tan, J. Dai, S. Wang, K. Yang, and K. Niu, "Learned image transmission toward machine-type semantic communications," in *Proc. IEEE 34th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, 2023, pp. 1–6.

[13] S. Ayzik and S. Avidan, "Deep image compression using decoder side information," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 699–714.

[14] N. Mital, E. Özyilkan, A. Garjani, and D. Gündüz, "Neural distributed image compression with cross-attention feature alignment," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2023, pp. 2498–2507.

[15] M. Song, J. Choi, and B. Han, "Variable-rate deep image compression through spatially-adaptive feature transform," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 2380–2389.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.

[17] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 1501–1510.

[18] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 606–615.

[19] P.-h. Li, S. K. Ankireddy, R. P. Zhao, H. Nourkhiz Mahjoub, E. Moradi Pari, U. Topcu, S. Chinchali, and H. Kim, "Task-aware distributed source coding under dynamic bandwidth," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2024, pp. 406–417.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.