Networks and Heterogeneous Media

https://www.aimspress.com/journal/nhm

NHM, 21(7): 2504–2525.
DOI: 10.3934/nhm.2026093
Received: 12 May 2026
Revised: 09 October 2026
Accepted: 19 November 2026
Published: 27 November 2026

Research article

Semantic-fidelity enhanced deep joint source-channel coding for wireless visual sensing in smart-city digital twins

Yang Wang[1,2,*], **Shantong Sun**[3], **Hanjun Duan**[4] **and Ziwei Tang**[1,2]

[1] School of Artificial Intelligence, Taizhou University, Taizhou 225300, China

[2] Jiangsu Key Laboratory of Intelligent Drug Screening and Repositioning (TZU), Taizhou 225300, China

[3] School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China

[4] School of Low-Altitude Technology and Electronic Information, Guangdong University of Science and Technology, Dongguan 523349, China

* **Correspondence:** Email: wy23912@gmail.com; Tel: +8618030203344.

Abstract: Smart-city digital twins rely on visual observations collected by distributed sensing devices in heterogeneous urban environments. For camera-based sensing, wireless image transmission is an important communication-layer function because visual data must be delivered from edge devices to edge servers or digital-twin platforms for reconstruction and machine analysis. Deep joint source-channel coding (D-JSCC) improves wireless image transmission by jointly learning source and channel coding, but conventional methods mainly optimize pixel-level distortion and may not preserve task-relevant semantic information. This paper proposes a semantic-fidelity enhanced D-JSCC framework, termed S-JSCC, for wireless visual sensing. S-JSCC introduces teacher-guided semantic supervision that combines supervised contrastive representation with attribution consistency guided by gradient-weighted class activation mapping (Grad-CAM), encouraging reconstructions to preserve both global semantic structure and local task-sensitive responses. Experiments on CIFAR-10 and KITTI under additive white Gaussian noise (AWGN) and simplified Rayleigh fading channels show that S-JSCC improves peak signal-to-noise ratio (PSNR) and recognition accuracy compared with D-JSCC and contrastive JSCC baselines.

Keywords: smart-city digital twins; wireless visual sensing; deep joint source-channel coding; semantic fidelity; machine perception

1. Introduction

Smart-city digital twins [1, 2] require continuously updated sensing data to maintain virtual representations of physical urban environments [3]. In practical deployments, visual observations are often collected by distributed cameras, roadside units, mobile sensing devices, and other edge terminals under heterogeneous network and channel conditions. These visual media streams support traffic monitoring, infrastructure inspection, public-space analytics, and intelligent transportation services. Therefore, reliable wireless visual sensing is an important communication-layer component for smart-city digital twins because the quality of reconstructed visual observations directly affects the fidelity of subsequent modeling and machine perception tasks.

Among different sensing modalities, image-based visual data provide rich spatial and semantic information. However, transmitting visual data from resource-limited edge devices to edge servers or digital-twin platforms is challenging due to bandwidth constraints, channel noise, fading, and heterogeneous sensing environments. In such networked visual sensing scenarios, the reconstructed images are expected to maintain not only pixel-level visual quality but also semantic information that is useful for downstream perception models. Channel distortion may produce reconstructions that appear visually acceptable but are less reliable for machine analysis, especially when task-sensitive regions or class-discriminative features are degraded.

Deep joint source-channel coding (D-JSCC) [4, 5] has emerged as an effective end-to-end solution for wireless image transmission. Unlike conventional separate source-channel coding schemes, D-JSCC directly maps images to channel symbols and learns noise-resilient representations under given channel conditions. Most existing D-JSCC methods are optimized using pixel-level distortion metrics such as mean squared error (MSE) or peak signal-to-noise ratio (PSNR). Although these objectives are effective for improving reconstruction quality, they do not explicitly constrain whether the reconstructed image preserves task-relevant semantic information.

Such semantic degradation is undesirable in machine-oriented scenarios, including edge-assisted visual monitoring, intelligent transportation, and visual sensing for smart-city digital twins. In these applications, downstream models may rely on object contours, class-discriminative regions, or other task-sensitive visual cues. Motivated by this observation, this work investigates semantic-fidelity optimization for D-JSCC in wireless image transmission. In a smart-city digital-twin pipeline, the communication layer determines the quality of visual observations delivered to the modeling and analytics layers. This paper therefore focuses on semantic-fidelity preservation at the wireless visual-sensing transmission stage, aiming to improve reconstructed visual inputs for subsequent machine perception.

1.1. Prior work

From the perspective of networks and heterogeneous media, the problem studied in this paper can be viewed as semantic-preserving visual media delivery over noisy wireless links. The transmitted content is image-based sensing media, while the communication environment may vary due to different channel conditions, sensing locations, and edge-device constraints. Unlike task-specific semantic communication methods that transmit only labels, semantic maps, or highly compressed task features, this work focuses on reconstructable visual media with improved semantic fidelity. This setting is suitable for digital-twin visual sensing pipelines where the same reconstructed observation may be used by different downstream

analytics modules.

Research on D-JSCC has addressed reconstruction quality and channel robustness through channel feedback [6], adaptive rate or bandwidth [7, 8], and designs for multipath fading [9]. These studies demonstrate the advantage of learning-based joint source-channel coding under noisy channels, but most of them still rely on pixel-level losses as the main optimization objective. For machine perception, however, a small pixel-level distortion does not always imply that task-relevant semantic information is well preserved. D-JSCC techniques achieve robustness against channel noise by optimizing the system in an end-to-end manner, yet the MSE or PSNR objective may limit the encoder's ability to distinguish semantically important features from background redundancy. As a result, reconstructed images may still suffer from semantic degradation under low signal-to-noise ratio (SNR) conditions, even when their pixel-level quality remains acceptable.

Recent studies have introduced mutual-information-based joint source-channel coding (MI-JSCC) for wireless image transmission [10] and contrastive-learning-based JSCC schemes [11, 12] to improve channel adaptability and semantic robustness. More recent work has further explored auxiliary semantic guidance for D-JSCC [13], dynamic encoder–decoder structures for semantic communications [14], and lightweight semantic image transmission models [15]. In parallel, semantic-oriented representation learning and feature compression have been studied for Internet of Things (IoT) and resource-constrained distributed sensor networks [16, 17], while hierarchical variational autoencoders have been explored for task-oriented semantic communications in partially observable decision processes [18]. However, the joint preservation of global semantic structure and local task-sensitive regions in reconstructed images remains insufficiently explored. While task-specific compression methods, such as semantic-map transmission or segmentation-mask coding, can be efficient for a particular downstream task, this work focuses on image reconstruction with semantic-fidelity preservation, so that the same reconstructed signal can be evaluated from both visual and machine-perception perspectives.

Recent generative approaches broaden the design space of semantic image communication, as reflected in the survey by Liang and Li [19]. The multimodal semantic communication (MMSemCom) system uses image and text semantic features to select among images produced by a diffusion receiver [20]. The joint source-channel noise adding with adaptive denoising (JSCNA-AD) framework incorporates channel noise into diffusion-based reconstruction and allocates denoising steps according to regional semantic importance [21]. The multi-text transmission semantic communication (Multi-SC) system combines multiple textual descriptions with semantic segmentation labels for image recovery [22].

These generative approaches use transmitted semantic information to guide image synthesis at the receiver. Teacher-guided JSCC incorporates task information into encoder–decoder optimization. S-JSCC follows this training-based approach by combining pixel reconstruction with semantic supervision from a fixed teacher, which is removed after training.

Guidance derived from gradient-weighted class activation mapping (Grad-CAM) has also been investigated in semantic communication. Wu et al. [23] transfer class-discriminative information from classification to few-shot detection while sharing the encoder across tasks. The semantic-importance-based deep joint source-channel coding (SD-JSCC) method [24] uses task-response gradients to weight feature reconstruction errors according to semantic importance. These studies establish the relevance of task-sensitive features to semantic transmission. S-JSCC combines supervised-contrastive teacher representations with local attribution consistency to guide

image reconstruction, linking the global structure of the teacher feature space with the preservation of class-sensitive visual information.

1.2. Motivation and contributions

The rise of deep neural networks (DNNs) has promoted the development of end-to-end JSCC frameworks that integrate source coding and channel coding into a unified learning process. These methods improve the robustness of wireless image transmission by learning channel-adaptive representations. However, when the reconstructed images are used for machine analysis, preserving pixel-level fidelity alone may be insufficient. It is therefore useful to introduce semantic-fidelity constraints into the D-JSCC training process.

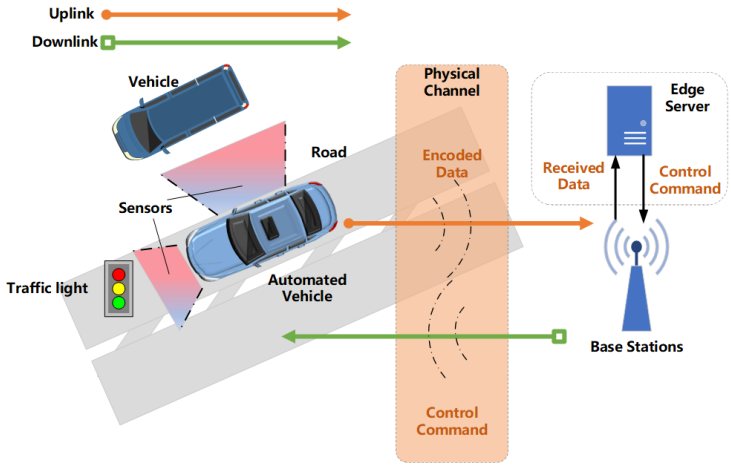


Figure 1. An edge-assisted visual sensing scenario for wireless image transmission toward smart-city digital twins.

In this paper, the considered edge-assisted visual sensing scenario is illustrated in Figure 1. A visual sensor transmits image data to an edge server through a wireless channel, where the received signal is reconstructed before being used by machine perception models or smart-city analytics services. This communication-layer setting is motivated by applications such as intelligent transportation, infrastructure monitoring, and edge-assisted visual sensing.

This paper presents a semantic-fidelity enhanced D-JSCC framework, termed S-JSCC, for machine-oriented wireless visual sensing. The central idea is to regularize the source image and its channel-distorted reconstruction in a semantic teacher space, instead of relying only on pixel-level reconstruction loss. The proposed formulation therefore encourages the encoder–decoder to preserve both visual fidelity and task-relevant semantic responses during reconstruction. The main contributions are summarized as follows:

- We formulate semantic-fidelity optimization for D-JSCC in networked wireless visual sensing. The considered setting corresponds to a communication-layer link in smart-city digital-twin pipelines, where visual sensing data are transmitted from distributed edge devices to edge servers or digital-twin platforms under noisy and fading channel conditions.
- We develop a global-local semantic supervision mechanism for D-JSCC training. A supervised-contrastive teacher provides a global semantic representation space with intra-class

compactness and inter-class separability, while the Grad-CAM-guided attribution term constrains local task-sensitive regions that strongly influence the teacher prediction. The two constraints jointly guide the reconstruction toward visual content that is more informative for downstream perception.

- We evaluate the proposed method on CIFAR-10 and KITTI wireless image transmission benchmarks under additive white Gaussian noise (AWGN) and simplified Rayleigh fading channels. The use of object-level images and driving-scene images provides two representative visual media settings for examining semantic-fidelity preservation. The results show that S-JSCC improves PSNR and classification-based semantic fidelity compared with conventional D-JSCC and contrastive JSCC baselines, with only a small increase in implementation overhead.

2. System model

We consider a point-to-point wireless image transmission system, where an image is encoded into channel symbols, transmitted through a noisy channel, and reconstructed at the receiver for both visual recovery and machine-perception evaluation. In the motivating smart-city digital-twin setting, the reconstructed image can be regarded as a visual observation delivered from the physical sensing layer to the edge or digital-twin analytics layer. Figure 2(a) illustrates the conventional D-JSCC pipeline, and Figure 2(b) shows the proposed S-JSCC framework. Since the proposed method is built upon the standard D-JSCC paradigm, we first introduce the networked visual-sensing interpretation, then describe the baseline system model, and finally present the semantic-fidelity extension.

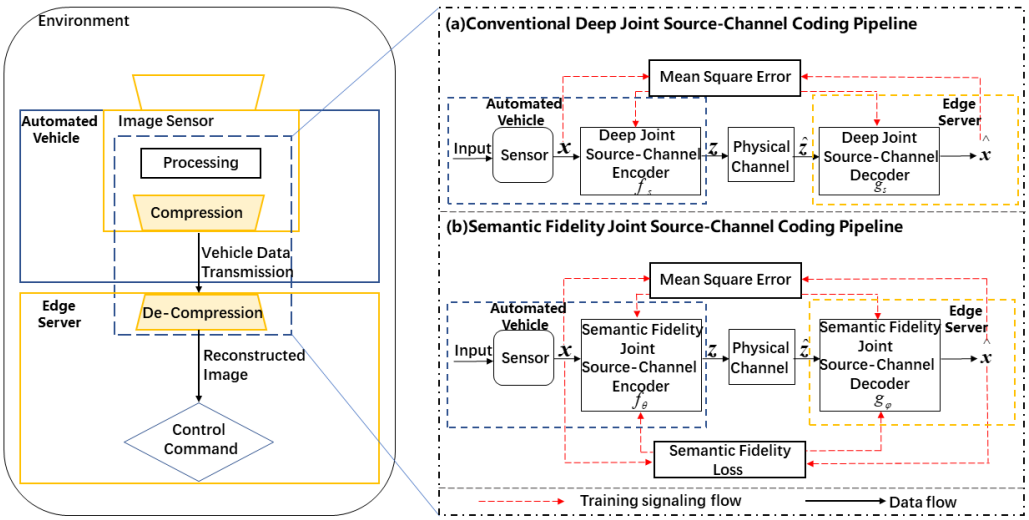


Figure 2. Overview of the considered wireless image transmission framework. (a) Conventional D-JSCC framework [4, 5]. (b) The proposed S-JSCC framework.

2.1. Networked visual-sensing interpretation

The considered transmission model represents a communication-layer building block in a networked visual sensing system. In a smart-city digital-twin scenario, multiple visual sensors may be deployed at different locations and connected to edge servers or digital-twin platforms through wireless links. Each sensing link may experience different noise levels, fading conditions, and device constraints. This paper focuses on the image transmission problem over one representative wireless link, while the same S-JSCC model can be applied to multiple distributed sensing links in a networked deployment.

The considered link-level model serves as a communication-layer building block in a broader smart-city digital-twin sensing pipeline. The objective is to improve the quality of reconstructed visual media delivered to downstream analytics modules. Therefore, the proposed method is evaluated in terms of pixel-level reconstruction quality, semantic-fidelity loss, and recognition accuracy of reconstructed images, which together reflect the usefulness of the transmitted visual observations for machine-oriented sensing applications.

2.2. D-JSCC under Gaussian and Rayleigh channels

The encoder function $f_s : \mathbb{R}^L \to \mathbb{C}^K$, implemented as a convolutional neural network (CNN) with trainable parameters S, transforms input data $\boldsymbol{x} \in \mathbb{R}^L$ into a compressed complex vector $\boldsymbol{z} \in \mathbb{C}^K$ for channel transmission. In D-JSCC systems, compression efficiency is commonly quantified by the feature reduction ratio K/L, where K denotes the dimension of encoded representations and L represents the original data dimensionality. The last layer normalizes the encoded vector to satisfy the average transmission power constraint:

$$\tilde{\boldsymbol{z}} = \sqrt{KP}\frac{\boldsymbol{z}}{\|\boldsymbol{z}\|_2}, \tag{2.1}$$

where $P > 0$ is the average symbol power and $\|\tilde{\boldsymbol{z}}\|_2^2 = KP$. For notational simplicity, the normalized channel input is still denoted by $\boldsymbol{z}$ in the following. The encoded vector is transmitted as a complex vector over the physical-layer channel. The received symbol is given by

$$\hat{\boldsymbol{z}} = \mathbf{H}\boldsymbol{z} + \boldsymbol{n}, \tag{2.2}$$

where $\mathbf{H}$ characterizes the channel gain and $\boldsymbol{n} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}_K)$ models the additive complex Gaussian noise. We analyze two classical channel models: the Rayleigh fading channel with complex Gaussian-distributed gain matrix $\mathbf{H} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_K)$ and the Gaussian channel with deterministic unit gain $\mathbf{H} = \mathbf{I}$. The average SNR at the receiver is consequently derived as

$$\rho = \frac{P}{\sigma^2}. \tag{2.3}$$

The decoder implements a reconstruction function $g_s : \mathbb{C}^K \to \mathbb{R}^L$ governed by learnable parameters S, transforming the received complex signal into the reconstructed output $\hat{\boldsymbol{x}} \in \mathbb{R}^L$. In standard D-JSCC frameworks, the MSE criterion $\mathcal{L}_{\text{mse}}$ is typically adopted as the reconstruction objective:

$$\mathcal{L}_{\text{mse}} = \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|^2, \tag{2.4}$$

where $\boldsymbol{x}_i$ and $\hat{\boldsymbol{x}}_i$ denote the i-th original and reconstructed images, respectively, and N is the batch size. Although the MSE objective is effective for reducing pixel-level reconstruction errors, it does not explicitly constrain the semantic consistency between the source and reconstructed images [25]. This motivates the S-JSCC formulation described in the next section.

2.3. Semantic-fidelity enhanced D-JSCC

As shown in Figure 2(b), the proposed S-JSCC follows the same basic transmission pipeline as conventional D-JSCC. The transmitter maps the input image into a channel-symbol vector $\boldsymbol{z} \in \mathbb{C}^K$ under the transmission power constraint. After passing through a noisy channel, the receiver obtains $\hat{\boldsymbol{z}} \in \mathbb{C}^K$ and reconstructs the image $\hat{\boldsymbol{x}} \in \mathbb{R}^L$. The main difference from conventional D-JSCC lies in the training objective, where semantic-fidelity supervision is introduced in addition to pixel-level reconstruction loss.

The training objective combines the conventional reconstruction loss and a semantic-fidelity loss:

$$\mathcal{L}_{\text{sum}}(\theta, \varphi) = \alpha \mathcal{L}_{\text{mse}} + (1 - \alpha)\mathcal{L}_{\text{sf}}, \tag{2.5}$$

where θ and φ represent the trainable parameters of the encoder and decoder and $\alpha \in (0, 1)$ is the trade-off coefficient. The semantic-fidelity loss between the reconstructed image $\hat{\boldsymbol{x}}$ and the original image $\boldsymbol{x}$ can be written in a general form as

$$\mathcal{L}_{\text{sf}} = \mathbb{E}_{p(\boldsymbol{x}, \hat{\boldsymbol{x}})} \left[d(S(\boldsymbol{x}), S(\hat{\boldsymbol{x}})) \right], \tag{2.6}$$

where $d(\cdot, \cdot)$ denotes the discrepancy between semantic responses, $p(\boldsymbol{x}, \hat{\boldsymbol{x}})$ characterizes the joint empirical distribution of original and reconstructed samples, and $S(\cdot)$ serves as a semantic-fidelity mapping. Since the true data distribution is generally unknown, this expected distortion is estimated empirically from training samples. In this work, the semantic-fidelity mapping is instantiated by a supervised contrastive teacher and a Grad-CAM-based attribution response, as described in Section 3.

2.4. Performance metrics

In the context of image transmission via D-JSCC, it is essential to evaluate both the pixel-level reconstruction quality and the preservation of semantic information. This work uses PSNR and recognition accuracy as the main quantitative metrics.

2.4.1. Peak signal-to-noise ratio (PSNR)

PSNR is widely used to assess the quality of image reconstruction. It is defined as

$$\text{PSNR} = 10 \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right), \tag{2.7}$$

where MAX specifies the dynamic range of pixel values and MSE denotes the average pixel-wise squared error between the source image $\boldsymbol{x}$ and its reconstruction $\hat{\boldsymbol{x}}$.

2.4.2. Semantic fidelity via recognition accuracy

To assess semantic consistency in reconstructed images, we use the recognition accuracy of a pre-trained DNN classifier. Given an original image $\boldsymbol{x} \in C_c$ belonging to class c, let $\hat{c}(\hat{\boldsymbol{x}})$ denote the

predicted class label for reconstruction $\hat{\boldsymbol{x}}$. The semantic preservation capability is operationalized through the recognition accuracy rate:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(c_i = \hat{c}_i), \tag{2.8}$$

where N is the number of test samples and $\mathbb{I}(\cdot)$ is the indicator function. A higher recognition accuracy indicates that the semantic content of the original image is better preserved in the reconstructed image [12, 26].

3. Methods and experimental setup

S-JSCC is a training-stage modification of D-JSCC. Section 3.1 describes the supervised-contrastive teacher representation that supplies the global semantic reference space. Section 3.2 constructs the local attribution response from clean and reconstructed feature maps and their class-response gradients. Section 3.3 combines the norm of this response with pixel reconstruction loss and summarizes codec optimization in Algorithm 1. In the stated codec objective, global guidance is supplied through the teacher representation; there is no separately weighted contrastive-loss term in Eq (3.9). The teacher is removed after optimization, so online transmission retains the encoder–channel–decoder pipeline.

3.1. Global semantic representation constraint

Traditional D-JSCC methods mainly learn noise-robust latent representations through pixel-level reconstruction objectives. To introduce semantic-fidelity supervision, we employ a pre-trained semantic model as a fixed training-stage teacher. The teacher provides a feature space in which semantic similarity can be measured more directly than in the pixel space. In this work, it provides two complementary forms of guidance: global semantic representation consistency and local task-sensitive attribution consistency.

A simple framework for contrastive learning of visual representations (SimCLR) [27] has demonstrated strong representation learning ability by maximizing the similarity between different augmented views of the same sample [11, 16]. Supervised contrastive learning has also been applied to robust wireless signal recognition, such as underwater acoustic modulation classification [28], which motivates its use as a communication-oriented representation regularizer. Let $\psi(\cdot)$ denote the normalized embedding extracted by the semantic teacher. For a reconstructed sample $\hat{\boldsymbol{x}}_i$, we write its teacher embedding as $\boldsymbol{u}_i = \psi(\hat{\boldsymbol{x}}_i)$ with $\|\boldsymbol{u}_i\|_2 = 1$.

Let $i \in I \equiv \{1, \ldots, 2N\}$ denote the index of an arbitrary augmented sample, and let $j(i)$ be the index of the other augmented sample originating from the same source sample. The contrastive learning loss is defined as

$$\mathcal{L}_{\text{contra}} = -\sum_{i\in I}\log\frac{\exp\left(\boldsymbol{u}_i^{\top}\boldsymbol{u}_{j(i)}/\tau\right)}{\sum_{a\in A(i)}\exp\left(\boldsymbol{u}_i^{\top}\boldsymbol{u}_a/\tau\right)}, \tag{3.1}$$

where τ is a temperature hyperparameter and $A(i) \equiv I \setminus \{i\}$. The index i is the anchor, $j(i)$ is the positive sample, and all other indices serve as negatives.

To utilize label information, supervised contrastive learning [29] further treats samples from the same class as positives. For a set of labeled samples $M = \{(\boldsymbol{x}^i, \boldsymbol{x}_l^i)\}_{i=1}^{2N}$, the supervised contrastive loss is

defined as

$$\mathcal{L}_{\text{sup}} = \sum_{i \in M} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\boldsymbol{u}_i^\top \boldsymbol{u}_p / \tau\right)}{\sum_{a \in A(i)} \exp\left(\boldsymbol{u}_i^\top \boldsymbol{u}_a / \tau\right)}, \tag{3.2}$$

where $P(i) \equiv \{p \in A(i) : \boldsymbol{x}_l^p = \boldsymbol{x}_l^i\}$ is the set of samples sharing the same label as sample i.

Previous studies [11, 12] indicate that contrastive learning can improve representation robustness under wireless noise. Following this motivation, we use supervised contrastive representation as a semantic teacher to regularize reconstructed images affected by channel noise. The supervised contrastive objective encourages same-class embeddings to be close relative to other classes. Its role here is to establish the teacher representation used by local attribution supervision, rather than to provide a bound for selecting the codec loss or its weight. As illustrated in Figure 3, this global constraint helps maintain the semantic structure of reconstructed samples under channel distortion.

Figure 3 illustrates the intended role of the teacher representation. Class-related samples should remain semantically close despite image distortion, while different groups should remain distinguishable. This motivates using a semantic teacher to evaluate reconstructions beyond their pixel error. The illustration does not establish a quantitative robustness bound or demonstrate detection performance.

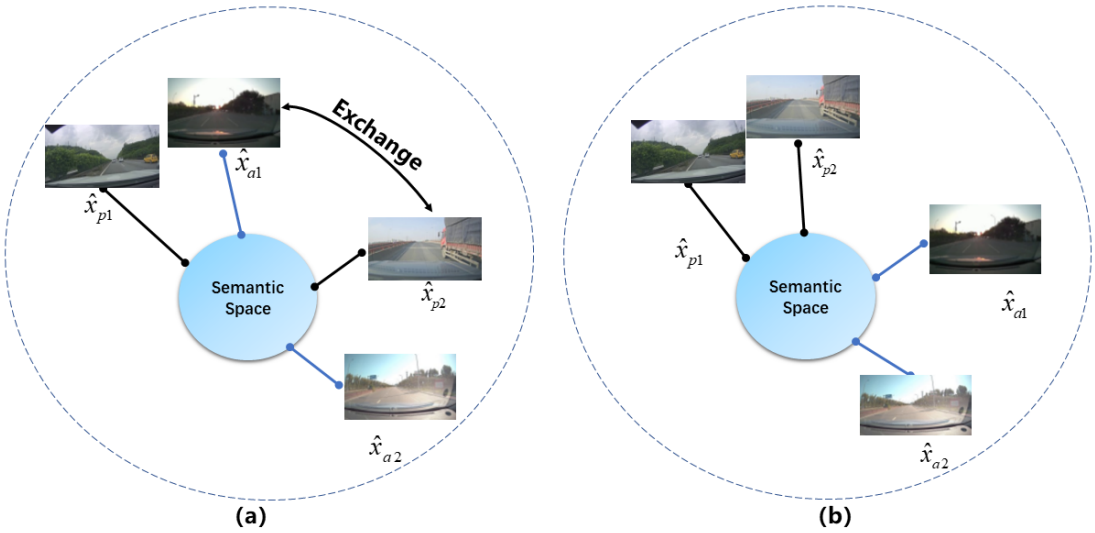


Figure 3. Conceptual illustration of semantic-space regularization. The image thumbnails represent reconstructed samples from two illustrative groups, distinguished by black and blue links to the semantic space. (a) Channel distortion may disturb their semantic relationships; the "Exchange" arrow depicts this possible displacement, not a sample-swapping operation. (b) Contrastive supervision encourages preservation of the intended group relationships. The layout is schematic and does not report measured embedding distances or a high-resolution KITTI evaluation.

3.2. Local semantic attribution constraint

The global contrastive constraint regularizes the overall semantic structure, but it does not explicitly indicate which spatial regions or feature maps are most responsible for the teacher's decision. For machine-oriented image transmission, this local information is important because channel distortion may alter class-discriminative regions even when the overall pixel-level error is small. Grad-CAM [30] provides a natural way to describe such local semantic evidence by weighting feature maps according to their contribution to the teacher's class response. Therefore, Grad-CAM is used to construct a training-stage attribution-consistency constraint between the clean image and its reconstructed counterpart.

The rationale can be described through a first-order approximation of the teacher response. Let l^c denote the teacher logit for class c, and let $\boldsymbol{A}^k$ be the k-th feature map of the teacher. When channel distortion changes the feature map from $\boldsymbol{A}^k$ to $\hat{\boldsymbol{A}}^k$, the change of the teacher response can be approximated as

$$\Delta l^c \approx \sum_k \sum_{i,j} \frac{\partial l^c}{\partial \boldsymbol{A}^k_{ij}} \left(\hat{\boldsymbol{A}}^k_{ij} - \boldsymbol{A}^k_{ij}\right). \tag{3.3}$$

Equation (3.3) indicates that distortions on feature maps with larger class-response gradients have a stronger influence on the semantic decision of the teacher. Grad-CAM summarizes this sensitivity by spatially averaging the gradients and using them as feature-map weights. Consequently, aligning Grad-CAM responses between the clean and reconstructed images provides a practical surrogate for preserving local task-sensitive semantic evidence. In digital-twin visual sensing, preserving such evidence is important because downstream analytics often depend on class-discriminative visual structures rather than uniform pixel fidelity across the entire image.

Let $\boldsymbol{x}$ denote the original image and $\hat{\boldsymbol{x}}$ its reconstructed version after passing through noisy channels. For the pre-trained teacher model $f_M(\cdot)$, we extract the corresponding feature maps:

$$\{\boldsymbol{A}^k\} = f_M(\boldsymbol{x}), \quad \{\hat{\boldsymbol{A}}^k\} = f_M(\hat{\boldsymbol{x}}). \tag{3.4}$$

For class c, the Grad-CAM weights of the clean and reconstructed images are computed as

$$w^c_k(\boldsymbol{x}) = \frac{1}{L} \sum_{i,j} \frac{\partial l^c}{\partial \boldsymbol{A}^k_{ij}}, \quad w^c_k(\hat{\boldsymbol{x}}) = \frac{1}{L} \sum_{i,j} \frac{\partial \hat{l}^c}{\partial \hat{\boldsymbol{A}}^k_{ij}}, \tag{3.5}$$

where l^c and $\hat{l}^c$ are the logits for class c derived from the clean and reconstructed features, respectively, and L is the number of spatial locations in each feature map. The gradient discrepancy is then defined as

$$\Delta w^c_k = w^c_k(\boldsymbol{x}) - w^c_k(\hat{\boldsymbol{x}}). \tag{3.6}$$

Based on this discrepancy, the semantic-fidelity attribution response is constructed as

$$\Delta \mathcal{H}^c = \mathrm{ReLU}\left(\sum_k \Delta w^c_k \left(\boldsymbol{A}^k - \hat{\boldsymbol{A}}^k\right)\right), \tag{3.7}$$

which highlights regions where the reconstruction process changes the teacher's task-sensitive semantic response. As illustrated in Figure 4, this term encourages the reconstructed image to preserve the attribution regions that contribute strongly to the teacher prediction.

Rectified linear unit (ReLU) in Eq (3.7) retains positive entries of the summed gradient–feature discrepancy and sets negative entries to zero. Its role is to define a one-sided attribution penalty, so signed discrepancy entries are not treated as equally penalized directions. This use is inspired by the positive-response filtering in Grad-CAM [30], but the filtered quantity here is a discrepancy product rather than the original class-evidence map. Consequently, a zero response need not imply identical features or identical predictions. The term is a training surrogate, not a symmetric distance or a guarantee of semantic equivalence.



Figure 4. Examples of reconstructed images and their corresponding Grad-CAM responses for local semantic-fidelity supervision.

3.2.1. Noise-robust feature alignment

The proposed semantic supervision establishes two complementary constraints. Reconstructed samples with similar semantic labels are encouraged to remain close in the teacher feature space, and each reconstructed image is encouraged to preserve the class-discriminative attribution response of the clean image. The joint optimization therefore enforces the following tendencies:

$$\begin{cases} \cos\beta_{i-j(i)} \to 1, & \text{global semantic clustering,} \\ \|\Delta\mathcal{H}^c\|_2 \to 0, & \text{local attribution consistency.} \end{cases} \tag{3.8}$$

The L2-norm minimization of $\Delta\mathcal{H}^c$ reduces the discrepancy between the local semantic responses of clean and reconstructed images. The semantic teacher is used to generate supervision during training and is removed during online transmission.

3.3. Loss function design and algorithm description

Based on the above global and local semantic constraints, the proposed S-JSCC can be viewed as a semantic-fidelity enhanced version of D-JSCC. The supervised contrastive teacher provides a global semantic representation space, while Grad-CAM provides a local attribution response that reflects class-discriminative evidence. The discrepancy between clean and reconstructed attribution responses is used as the semantic-fidelity loss. Accordingly, the overall training objective is formulated as follows:

$$\mathcal{L}_{\text{sum}}(\theta,\varphi) = \underbrace{\alpha\mathcal{L}_{\text{mse}}}_{\text{Reconstruction loss}} + \underbrace{(1-\alpha)\,\|\Delta\mathcal{H}^c\|_2}_{\text{Semantic-fidelity loss}}. \tag{3.9}$$

Equation (3.9) encourages the encoder–decoder to reduce pixel-level distortion while preserving the task-sensitive semantic responses measured by the semantic teacher. Wang and Li [31] use a weight

of 0.5 to combine a PSNR-derived reconstruction term and classification cross-entropy in semantic image transmission. Consistent with this equal-coefficient convention, we set $\alpha = 0.5$ for pixel reconstruction and teacher-guided semantic supervision. The specific algorithmic steps are described in Algorithm 1.

Algorithm 1: Training of S-JSCC.

Input: Dataset $\mathbf{x}$ and pre-trained model $f_M(\cdot)$
Output: Encoder parameters θ and decoder parameters φ

while *stopping criterion is not met* **do**
  Encode $\mathbf{x}$ into $\mathbf{z}$;
  Transmit $\mathbf{z}$ through the channel and obtain $\hat{\mathbf{z}}$;
  Decode $\hat{\mathbf{z}}$ into $\hat{\mathbf{x}}$;
  Extract $\{\mathbf{A}^k\}$ and $\{\hat{\mathbf{A}}^k\}$ using $f_M(\cdot)$;
  Calculate $\Delta\mathcal{H}^c$ using Equation (3.7);
  Calculate $\mathcal{L}_{\text{sum}}$ using Equation (3.9);
  Update θ and φ by gradient descent;
end
return θ and φ;

3.4. Model and experimental setup

The proposed S-JSCC framework is implemented by retaining the baseline D-JSCC encoder–decoder architecture illustrated in Figure 5. This setting allows a direct comparison between the conventional reconstruction-loss-based D-JSCC baseline and the proposed semantic-fidelity enhanced training objective. The D-JSCC baseline follows the learning-based image recovery framework in [4], and a contrastive-learning-based JSCC method [11] is also included as a comparison baseline.

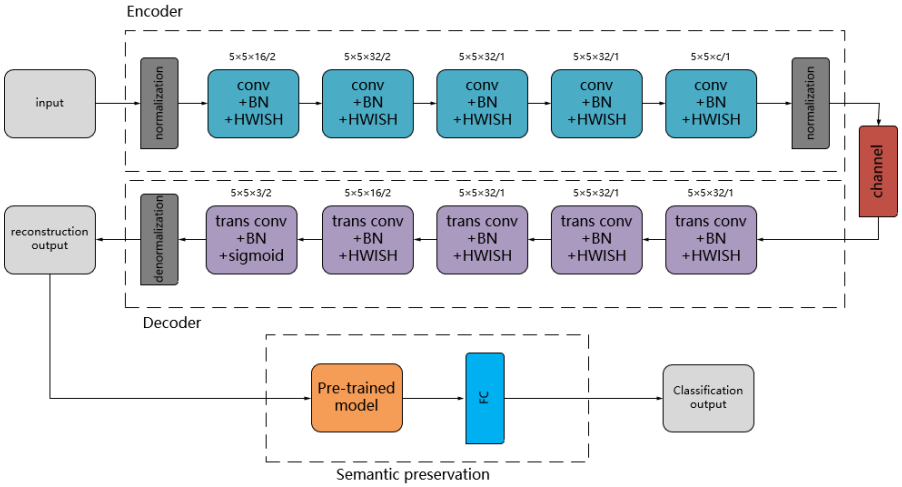


Figure 5. DNN architecture and implementation configuration.

We use CIFAR-10 [32] and KITTI [33] as image-transmission benchmarks. The original benchmark experiments use 32×32 inputs following the baseline D-JSCC configuration. We additionally evaluate full-frame red-green-blue (RGB) KITTI inputs of 256×832, with the corresponding results reported

in Table 3. CIFAR-10 provides a controlled object-level visual recognition benchmark, while KITTI provides driving-scene images relevant to urban mobility and smart-city sensing.

The learning rate is set to 0.001 using the AdamW optimizer, and the batch size is 64. The pre-trained semantic teacher $f_M(\cdot)$ is ResNet-50 [34]. The ResNet-50 model is used only during training to supervise semantic fidelity and is removed during inference, so the deployed JSCC encoder–decoder remains close to the standard D-JSCC architecture. The performance of different JSCC schemes is evaluated using PSNR, semantic-fidelity loss, classification accuracy of reconstructed images, visual comparison, and inference complexity. The specific experimental environment is shown in Table 1.

Table 1. Experimental environment.

Component	Version/Specification
Pytorch	1.7.0
Python	3.8.0
GPU	RTX2060 6GB
CPU	Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz

For classification-based semantic evaluation on CIFAR-10, reconstructed outputs undergo a standardized evaluation protocol. (1) An ImageNet-pretrained vision transformer is fine-tuned exclusively on native CIFAR-10 data. (2) Reconstructed samples are not used to train this classifier. The resulting classification accuracy on channel-distorted reconstructions is used as a quantitative measure of semantic preservation.

In addition to the AWGN channel, we employ a simplified Rayleigh fading model. In this model, the link is a single-input single-output (SISO) system experiencing flat fading, characterized by a single complex gain per transmission batch, i.e., a single path without multipath delay. This gain remains constant throughout each batch, corresponding to block fading, and no Doppler shift is considered. These channel models provide basic noisy and fading transmission settings for evaluating the proposed semantic-fidelity objective.

4. Results and discussion

4.1. Reconstruction quality under different channel conditions

We first evaluate the reconstruction quality of D-JSCC and S-JSCC under different SNR values. Figure 6 reports the reconstructed PSNR on CIFAR-10 and KITTI under both AWGN and simplified Rayleigh fading channels.

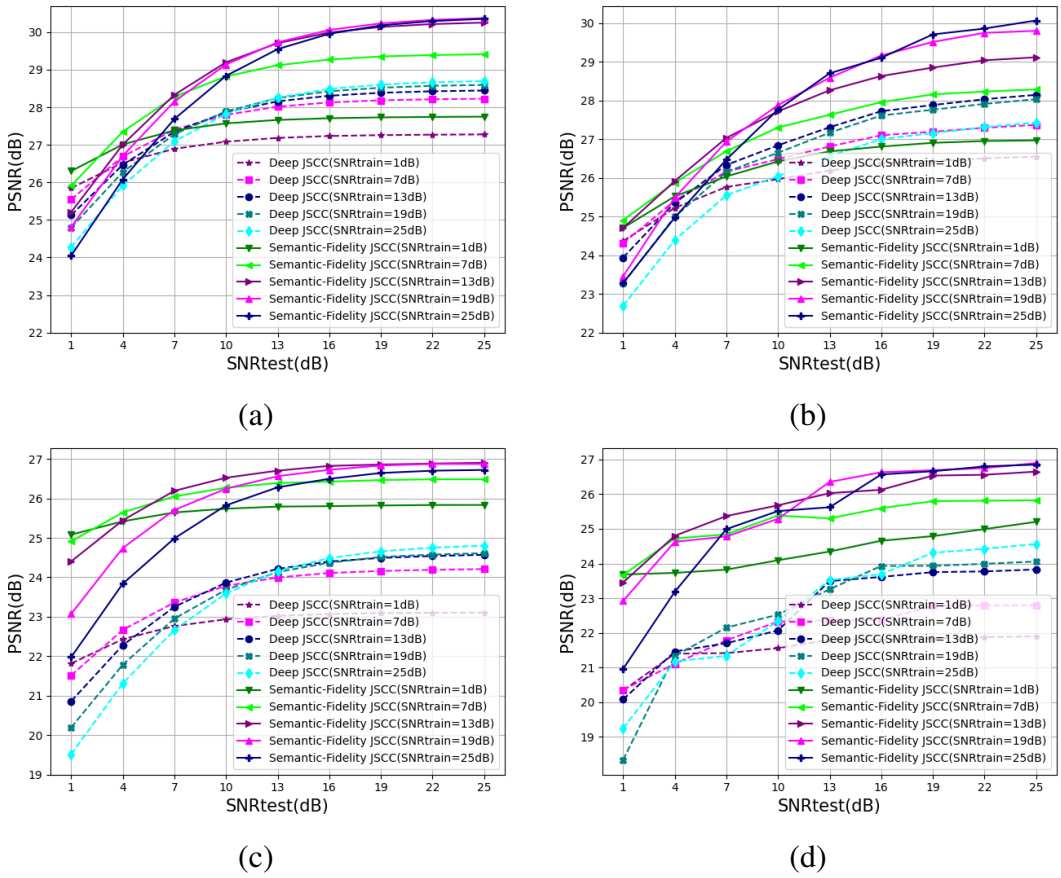
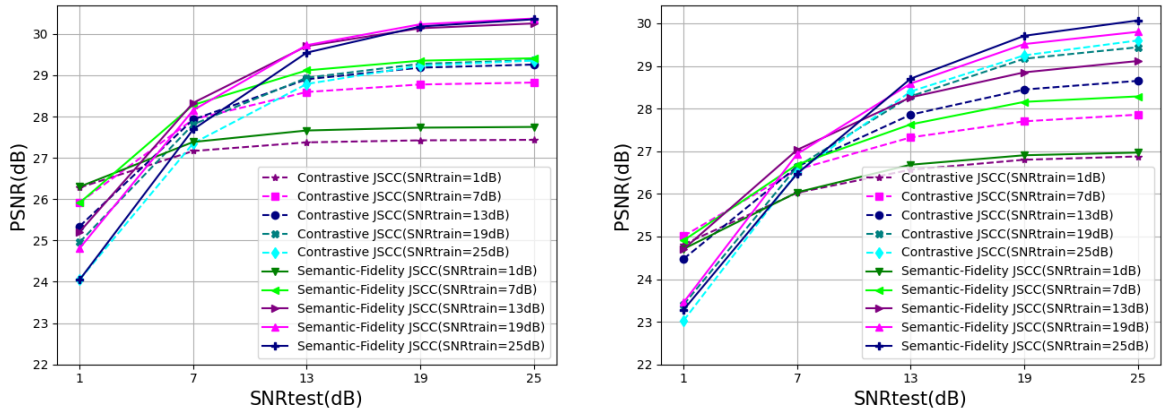


Figure 6. Reconstruction quality of D-JSCC and S-JSCC under different training SNRs ($K = 500$). (a) CIFAR-10 under AWGN; (b) CIFAR-10 under Rayleigh fading; (c) KITTI under AWGN; (d) KITTI under Rayleigh fading.

As shown in Figure 6, S-JSCC consistently improves PSNR across SNR values from 0 dB to 25 dB. Since PSNR $= 10\log_{10}(\mathrm{MAX}^2/\mathrm{MSE})$, the PSNR curves also reflect the MSE performance as a function of SNR. On the CIFAR-10 dataset, S-JSCC outperforms D-JSCC by 0.5–2 dB in AWGN and 0.3–2.8 dB in Rayleigh channels. On the KITTI dataset, it achieves average gains of 2.56 dB in AWGN and 2.91 dB in Rayleigh channels. These results indicate that introducing semantic-fidelity supervision does not compromise pixel-level reconstruction quality; instead, it improves reconstruction performance under the tested channel settings. From the perspective of networked visual sensing, the AWGN and Rayleigh settings can be interpreted as two representative wireless transmission conditions experienced by distributed edge sensors. The consistent gains under both channels indicate that the proposed semantic-fidelity supervision improves not only visual reconstruction but also the reliability of visual media delivery under heterogeneous channel conditions.

4.2. Comparison with existing JSCC baselines

To further evaluate the proposed method, we compare it with C-JSCC, a contrastive-learning-based JSCC scheme. Figure 7 compares the reconstructed PSNR of C-JSCC and S-JSCC across SNR values from 0 dB to 25 dB under AWGN and simplified Rayleigh fading channels.



(a) (b)

Figure 7. PSNR comparison between C-JSCC [11] and S-JSCC on CIFAR-10 with $K = 500$. (a) AWGN channel; (b) Rayleigh fading channel.

On the CIFAR-10 dataset, S-JSCC achieves average gains of 0.5 dB in AWGN and 0.25 dB in Rayleigh channels over C-JSCC. These results suggest that combining global contrastive guidance with local attribution guidance is beneficial under the tested settings.

4.3. Complexity analysis

Table 2 reports the average processing runtime and parameter count of D-JSCC and S-JSCC.

Table 2. Average processing runtime and parameters of different schemes.

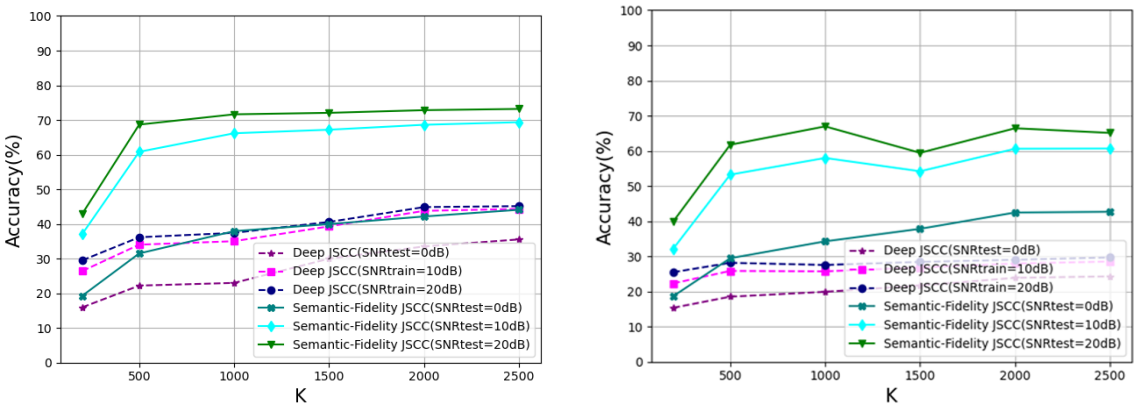
Method	Parameters (M)	Inference time (ms)
D-JSCC	0.4772	1.72
S-JSCC	0.4777	1.98

S-JSCC includes batch normalization (BN) to improve training convergence, whereas D-JSCC does not. The additional BN operations account for the inference-time increase in Table 2; the semantic teacher is removed after training and incurs no inference cost.

The ResNet-50 teacher is used only during training and is removed for inference. Table 2 reports 0.4772M parameters and 1.72 ms per image for D-JSCC, versus 0.4777M and 1.98 ms for S-JSCC. The additional inference time comes from the BN configuration, not the teacher. These measurements characterize deployment cost only; training cost has not been quantified.

4.4. Recognition accuracy and high-resolution evaluation

Figure 8 shows the classification accuracy of reconstructed images under different compression ratios and test SNR values. Compared with conventional D-JSCC, which mainly emphasizes pixel-level reconstruction, S-JSCC achieves a 10–25.5% recognition accuracy improvement on the AWGN channel and an 18–35% improvement on Rayleigh channels. These results suggest that the proposed semantic-fidelity objective helps preserve task-relevant information that may not be fully captured by pixel-level reconstruction loss.



(a) AWGN channel ($\mathbf{H} = \mathbf{I}_K$) (b) Rayleigh channel ($\mathbf{H} \sim \mathcal{CN}(0, \mathbf{I}_K)$)

Figure 8. Recognition accuracy of reconstructed images using S-JSCC trained at $\mathrm{SNR}_{\mathrm{train}} = 20$ dB under different compression ratios and test SNRs. (a) AWGN channel; (b) Rayleigh fading channel.

To examine performance beyond the low-resolution benchmarks, Table 3 reports KITTI validation results using full-frame RGB inputs of 256×832. D-JSCC, S-JSCC, and MI-JSCC [10] are trained at 13 dB and evaluated at 13 dB, with separate training for AWGN and Rayleigh channels and a matched transmission budget of $K = 100040$ for all three methods.

Table 3. KITTI validation with full-frame RGB inputs of 256×832. All three methods use $K = 100040$ and are trained and evaluated at 13 dB; AWGN and Rayleigh models are trained separately.

Channel	Method	PSNR (dB)	Accuracy (%)
AWGN	S-JSCC	27.26	97.59
AWGN	D-JSCC	26.38	97.08
AWGN	MI-JSCC	26.99	97.33
Rayleigh	S-JSCC	26.40	96.62
Rayleigh	D-JSCC	26.21	96.00
Rayleigh	MI-JSCC	26.10	96.12

Under AWGN, S-JSCC achieves 27.26 dB PSNR and 97.59% classification accuracy, compared with 26.38 dB and 97.08% for D-JSCC and 26.99 dB and 97.33% for MI-JSCC. Under Rayleigh fading, S-JSCC achieves 26.40 dB and 96.62%, compared with 26.21 dB and 96.00% for D-JSCC and 26.10 dB and 96.12% for MI-JSCC. Thus, S-JSCC retains a reconstruction and classification advantage in this higher-resolution setting, including PSNR gains of 0.27 dB and 0.30 dB over MI-JSCC under AWGN and Rayleigh fading, respectively.

Table 4 compares D-JSCC, S-JSCC, and D-JSCC with perceptual supervision based on the 16-layer Visual Geometry Group network (VGG16) [35]. The PSNR and classification results use the two SNR settings specified in the table. S-JSCC has a 0.45 dB higher PSNR than the VGG baseline in the PSNR setting, while its classification accuracy is 1 percentage point lower in the classification setting.

Table 4. Comparison with VGG-based perceptual supervision on CIFAR-10 over AWGN ($K = 500$). The two metrics use different SNR settings.

Method Train/test: 1/1 dB	PSNR (dB) Train/test: 0/0 dB	Accuracy (%)
D-JSCC	≈ 25.8	≈ 22.3
S-JSCC	≈ 26.3	≈ 31.6
D-JSCC+VGG ($\alpha = 0.5$)	≈ 25.85	≈ 32.6

The three methods in Table 4 compare different forms of supervision: pixel reconstruction in D-JSCC, additional VGG feature reconstruction in D-JSCC+VGG, and teacher-guided semantic supervision in S-JSCC. In S-JSCC, supervised-contrastive learning shapes the teacher representation, while local attribution consistency uses that representation to guide reconstruction. These components therefore act together within the training formulation. The comparison evaluates the combined formulation against pixel- and feature-level alternatives; it does not isolate the contribution of either semantic component. Component-wise attribution remains a limitation of the present evaluation.

4.5. Visual comparison under similar PSNR and SSIM



Figure 9. Visual comparison of S-JSCC and D-JSCC under similar PSNR/SSIM values. The text below each image reports PSNR/SSIM; blue boxes denote S-JSCC results, and the last row denotes D-JSCC results.

Figure 9 compares reconstructed outputs of S-JSCC and D-JSCC under similar PSNR and structural similarity index (SSIM) values, which shows that images with lower semantic-fidelity loss tend to preserve clearer task-relevant visual structures, even when traditional pixel-level metrics are close.

4.6. Limitations and scope

The original 32×32 benchmarks provide controlled low-resolution comparisons, while the additional 256×832 KITTI evaluation in Table 3 directly examines larger driving-scene inputs. S-JSCC retains higher PSNR and classification accuracy than D-JSCC and MI-JSCC at the evaluated 13 dB training and validation SNR under both channel models. Higher-resolution feature maps and codec activations increase memory and computational requirements. A fixed channel-symbol budget imposes stronger compression, whereas a fixed channel-use ratio requires more transmitted symbols. The results support the effectiveness of the combined semantic supervision in the evaluated higher-resolution setting; broader resolution, SNR, and downstream-task coverage remains future work.

Category discrimination is a foundational capability shared by image classification, object detection, and semantic segmentation. The region-based convolutional neural network (R-CNN) and fully convolutional networks demonstrate the use of classification-pretrained representations for detection and segmentation through task-specific adaptation [36, 37]. We therefore use classification as a tractable starting point for evaluating semantic preservation in S-JSCC. This choice simplifies task-specific supervision and evaluation by using image-level labels, allowing the study to focus on whether reconstructed images retain category-discriminative information under channel distortion.

AWGN and Rayleigh fading provide foundational models of additive noise and multipath-induced fading. The Rayleigh model considered here is SISO and flat block-fading, with a single complex gain held constant within each transmission batch. This controlled setting abstracts away multipath delay spread and Doppler variation. A broadband orthogonal frequency-division multiplexing (OFDM) extension [9] can incorporate frequency-selective multipath and time-varying fading, together with cyclic-prefix and pilot overhead, channel estimation, equalization, and Doppler-induced inter-carrier interference.

5. Conclusions

This paper presented S-JSCC, a semantic-fidelity enhanced D-JSCC framework for networked wireless visual sensing in smart-city digital-twin scenarios. The proposed method focuses on the communication-layer problem of transmitting visual sensing data from distributed edge devices to edge servers or digital-twin platforms under noisy and fading channel conditions. Unlike conventional D-JSCC methods that mainly optimize pixel-level reconstruction loss, S-JSCC introduces teacher-guided semantic supervision that combines global supervised contrastive representation and local Grad-CAM-guided attribution consistency.

Experiments on CIFAR-10 and KITTI under AWGN and simplified Rayleigh fading channels show that S-JSCC improves reconstruction quality and classification-based semantic fidelity compared with D-JSCC and contrastive JSCC baselines under the tested settings. These results indicate that semantic-fidelity supervision can improve the usefulness of reconstructed visual media for downstream machine perception in distributed sensing pipelines. The complexity results further show that the additional implementation overhead is small in the reported configuration because the semantic teacher is removed during online transmission. Overall, the proposed semantic-fidelity supervision provides a practical training strategy for reliable visual media delivery in networked smart-city sensing scenarios.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the High-level Talent Research Initiation Fund Project of Taizhou University under Grant TZXYQD2024A007.

Conflict of interest

The authors declare there is no conflict of interest.

Author contributions

Yang Wang contributed to Conceptualization, Investigation, Funding acquisition, Methodology, and Writing—review & editing. Shantong Sun contributed to Formal analysis and Validation. Hanjun Duan contributed to Data collection and Validation. Ziwei Tang contributed to Software and Writing—review & editing. All authors read and approved the final manuscript.

Availability of data and materials

The datasets used in this study are publicly available from the CIFAR-10 repository (https://www.cs.toronto.edu/~kriz/cifar.html) and the KITTI repository (http://www.cvlibs.net/datasets/kitti/).

References

1. M. Yessef, Y. Hakam, M. Tabaa, M. M. Alammar, Z. M. S. Elbarbary, Digital twin technology in smart cities: A step toward intelligent urban management, *Energy Rep.*, **14** (2025), 5539–5557. https://doi.org/10.1016/j.egyr.2025.11.097
2. M. Alvi, H. Dutta, R. Minerva, N. Crespi, S. M. Raza, M. Herath, Global perspectives on digital twin smart cities: Innovations, challenges, and pathways to a sustainable urban future, *Sustain. Cities Soc.*, **126** (2025), 106356. https://doi.org/10.1016/j.scs.2025.106356
3. T. A. Jauhar, S. Han, T. H. Khan, C. Noh, A. Lavrinec, An implementation of sensor network to assess near real-time digital twin fidelity, *Cyber-Phys. Syst.*, **12** (2026), 366–389. https://doi.org/10.1080/23335777.2026.2613415
4. R. Prenger, R. Valle, B. Catanzaro, Waveglow: A flow-based generative network for speech synthesis, in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, 3617–3621. https://doi.org/10.1109/ICASSP.2019.8683143

5. E. Bourtsoulatze, D. B. Kurka, D. Gündüz, Deep joint source-channel coding for wireless image transmission, *IEEE Trans. Cogn. Commun. Netw.*, **5** (2019), 567–579. https://doi.org/10.1109/TCCN.2019.2919300

6. D. B. Kurka, D. Gündüz, DeepJSCC-F: Deep joint source-channel coding of images with feedback, *IEEE J. Sel. Areas Inf. Theory*, **1** (2020), 178–193. https://doi.org/10.1109/JSAIT.2020.2987203

7. M. Yang, H. S. Kim, Deep joint source-channel coding for wireless image transmission with adaptive rate control, in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, 2022, 5193–5197. https://doi.org/10.1109/ICASSP43922.2022.9746335

8. D. B. Kurka, D. Gündüz, Bandwidth-agile image transmission with deep joint source-channel coding, *IEEE Trans. Wireless Commun.*, **20** (2021), 8081–8095. https://doi.org/10.1109/TWC.2021.3090048

9. M. Yang, C. Bian, H. S. Kim, OFDM-guided deep joint source channel coding for wireless multipath fading channels, *IEEE Trans. Cogn. Commun. Netw.*, **8** (2022), 584–599. https://doi.org/10.1109/TCCN.2022.3151935

10. N. Huang, W. Hua, L. Chen, MI-based joint source channel coding for wireless image transmission, in *2023 IEEE/CIC International Conference on Communications in China (ICCC)*, Dalian, China, 2023, 1–6. https://doi.org/10.1109/ICCC57788.2023.10233378

11. W. Hua, D. Chen, J. Fang, L. Chen, J. F. C. Mota, X. Hong, Semantics-guided contrastive joint source-channel coding for image transmission, in *2022 14th International Conference on Wireless Communications and Signal Processing (WCSP)*, Nanjing, China, 2022, 505–510. https://doi.org/10.1109/WCSP55476.2022.10039111

12. S. Tang, Q. Yang, L. Fan, X. Lei, Y. Deng, A. Nallanathan, Contrastive learning based semantic communication for wireless image transmission, in *2023 IEEE 98th Vehicular Technology Conference (VTC2023-Fall)*, Hong Kong, 2023, 1–6. https://doi.org/10.1109/VTC2023-Fall60731.2023.10333392

13. J. Feng, G. Cao, J. Liu, Z. Wang, Z. Liu, Y. Ding, Deep joint source-channel coding for wireless image transmission with auxiliary semantics, *Phys. Commun.*, **73** (2025), 102836. https://doi.org/10.1016/j.phycom.2025.102836

14. A. D. Raha, A. Adhikary, M. Gain, Y. Park, W. Saad, C. S. Hong, DD-JSCC: Dynamic deep joint source-channel coding for semantic communications, in *ICC 2025 - IEEE International Conference on Communications*, Montreal, QC, Canada, 2025, 3754–3759. https://doi.org/10.1109/ICC52391.2025.11162089

15. X. Li, D. Ban, Z. Ruan, X. Yue, H. Chen, Y. Sun, A star modulation network for wireless image semantic transmission, *Sci. Rep.*, **15** (2025), 31127. https://doi.org/10.1038/s41598-025-16753-4

16. W. Hua, L. Xiong, S. Liu, L. Chen, X. Hong, J. F. C. Mota, Classification-driven discrete neural representation learning for semantic communications, *IEEE Internet Things J.*, **11** (2024), 16061–16073. https://doi.org/10.1109/JIOT.2024.3354312

17. L. Xiong, W. Hua, X. Hong, X. Cheng, Semantic-oriented feature compression for resource-constrained distributed sensor networks, in *2024 IEEE Wireless Communications and Networking Conference (WCNC)*, Dubai, United Arab Emirates, 2024, 1–6. https://doi.org/10.1109/WCNC57260.2024.10570935

18. D. Chen, W. Hua, Hierarchical VAE based semantic communications for POMDP tasks, in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, 2024, 5540–5544. https://doi.org/10.1109/ICASSP48485.2024.10445833

19. C. Liang, D. Li, Generative AI-enabled semantic communication: State-of-the-art, applications, and the way ahead, *IEEE Commun. Surv. Tutor.*, **28** (2026), 3976–4015. https://doi.org/10.1109/COMST.2025.3649707

20. C. Liang, D. Li, Image generation with supervised selection based on multimodal features for semantic communications, *IEEE Trans. Commun.*, **73** (2025), 14469–14485. https://doi.org/10.1109/TCOMM.2025.3615798

21. C. Liang, D. Li, Joint source-channel noise adding with adaptive denoising for diffusion-based semantic communications, *IEEE Internet Things J.*, **12** (2025), 45909–45912. https://doi.org/10.1109/JIOT.2025.3589398

22. P. Huang, D. Li, Exploring textual semantics diversity for image transmission in semantic communication systems, *IEEE Trans. Commun.*, **74** (2026), 10602–10611. https://doi.org/10.1109/TCOMM.2026.3704182

23. Q. Wu, F. Liu, H. Xia, T. Zhang, Semantic transfer between different tasks in the semantic communication system, in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, Austin, TX, USA, 2022, 566–571. https://doi.org/10.1109/WCNC51071.2022.9771768

24. Q. Sun, C. Guo, Y. Yang, J. Chen, R. Tang, C. Liu, Deep joint source-channel coding for wireless image transmission with semantic importance, preprint, arXiv:2302.02287. https://doi.org/10.48550/arXiv.2302.02287

25. Q. Hu, G. Zhang, Z. Qin, Y. Cai, G. Yu, G. Y. Li, Robust semantic communications with masked VQ-VAE enabled codebook, *IEEE Trans. Wireless Commun.*, **22** (2023), 8707–8722. https://doi.org/10.1109/TWC.2023.3265201

26. F. Kalantari, K. Faez, H. Amindavar, S. Nazari, Improved image reconstruction from brain activity through automatic image captioning, *Sci. Rep.*, **15** (2025), 4907. https://doi.org/10.1038/s41598-025-89242-3

27. T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in *Proceedings of the 37th International Conference on Machine Learning*, **119** (2020), 1597–1607. Available from: https://proceedings.mlr.press/v119/chen20j.html.

28. D. Gao, W. Hua, W. Su, Z. Xu, K. Chen, Supervised contrastive learning-based modulation classification of underwater acoustic communication, *Wirel. Commun. Mob. Comput.*, **2022** (2022), 3995331. https://doi.org/10.1155/2022/3995331

29. P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, et al., Supervised contrastive learning, in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*, 2020, 18661–18673. Available from: https://dl.acm.org/doi/abs/10.5555/3495724.3497291.

30. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, Dhruv Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, 618–626. https://doi.org/10.1109/ICCV.2017.74

31. Z. Wang, H. Li, A multi-task training semantic communication system for image reconstruction and classification tasks, *J. Imaging*, **12** (2026), 367. https://doi.org/10.3390/jimaging12080367

32. A. Krizhevsky, *Learning Multiple Layers of Features from Tiny Images*, University of Toronto, Technical Report, 2009. Available from: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

33. A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: The KITTI dataset, *Int. J. Robot. Res.*, **32** (2013), 1231–1237. https://doi.org/10.1177/0278364913491297

34. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, 770–778. https://doi.org/10.1109/CVPR.2016.90

35. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in *International Conference on Learning Representations*, 2015. Available from: https://mlanthology.org/iclr/2015/simonyan2015iclr-very/.

36. R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, 580–587. https://doi.org/10.1109/CVPR.2014.81

37. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965



AIMS Press

© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)