FSCM-Net: Frequency-Spatial Collaborative Modeling for Traffic Scene Semantic Segmentation

Wei Zhao[1]*, Yi Dong[2], Lingchao Wang[3], Qiang Ai[4]
College of Information Engineering, Liaoning Institute of Science and Engineering, Jinzhou, China[1,2,3]
School of Information Science and Engineering, Lanzhou University, Lanzhou, China[4]

***Abstract*—To address the challenges of insufficient global semantic modeling and blurred boundaries in urban traffic scene segmentation, this study proposes a frequency-spatial collaborative framework based on DeepLabV3+. A Spectral Decoupling Adaptive Modulation (SDAM) module enhances low-frequency semantics and high-frequency details in the frequency domain. A Hierarchical Spatial Dependency Modeling (HSDM) module captures local consistency and global semantic dependencies, while a Structure-guided Adaptive Multi-scale Fusion (SAMF) module dynamically integrates multi-scale features using structural priors. Experiments on Cityscapes and CamVid demonstrate improvements of 3.6% and 1.6% over DeepLabV3+, respectively, while maintaining real-time performance.**

Keywords*—Frequency-domain modeling; spatial dependency modeling; local-global attention; structure-guided fusion; feature enhancement

I. Introduction

With the rapid advancement of autonomous driving, Intelligent Transportation Systems (ITS), and Advanced Driver Assistance Systems (ADAS), robust perception of traffic scenes has become a prerequisite for achieving automated driving. As a fundamental task in visual perception, semantic segmentation aims to perform fine-grained parsing of critical road elements, such as vehicles and pedestrians, thereby providing reliable support for environmental understanding and decision-making processes [1].

Current research on traffic scene semantic segmentation has mainly focused on feature enhancement and contextual modeling. Zhao et al. [2] proposed PSPNet, which captures multi-scale global contextual information through pyramid pooling and significantly improves segmentation accuracy in complex scenarios. Song et al. [3] developed MPNet based on DeepLabV3+, which effectively integrates multi-level feature representations through a multi-scale feature fusion architecture, thereby enhancing the perception capability for objects of varying scales and achieving notable improvements in traffic scene segmentation. However, these approaches primarily rely on spatial-domain feature fusion and do not explicitly exploit frequency-domain information for feature modeling and optimization. To compensate for the limitations of spatial modeling, several studies have investigated image representation from the frequency-domain perspective. Shan et al. [4] demonstrated that images can be decomposed into different frequency components through frequency analysis, where low-frequency components encode global structures and semantic information, while high-frequency components contain texture and boundary details, enabling a more comprehensive representation of visual information. Building upon this observation, Zhang et al. [5] proposed an adaptive semantic segmentation method based on frequency-domain analysis, which decouples style and semantic representations via Fast Fourier Transform (FFT) and enhances robustness under domain shifts. To further incorporate frequency information into fine-grained spatial modeling, especially for object boundary representation, Ge et al. [6] introduced SF-Net, which decomposes spatial features into high- and low-frequency components using the Haar Wavelet Transform and employs a Mamba-based enhancement module to strengthen global contextual representations. Although improvements in boundary accuracy were achieved, the method is prone to introducing frequency-domain noise in complex environments. Yang et al. [7] proposed the Frequency Feature Rectification (FFR) framework, which explicitly compensates for high-frequency components within the decoder to recover object boundaries, thereby improving segmentation quality in weakly supervised scenarios. Furthermore, Huang et al. [8] proposed FSDR, which performs randomization operations in the frequency domain to improve domain generalization capability and empirically validates the effectiveness of frequency-domain representations against domain shifts. Nevertheless, existing frequency-based methods mainly focus on feature enhancement or domain adaptation, while the deep collaborative modeling between frequency-domain and spatial-domain representations remains insufficient.

To further exploit frequency information for deep feature modeling and improve boundary recognition of small objects, increasing attention has been devoted to integrating attention mechanisms and multi-scale modeling strategies. Sima et al. [9] proposed SPCONet, which balances accuracy and efficiency through lightweight branches by jointly modeling spatial and contextual information, providing an effective solution for real-time traffic scene segmentation. To enhance the perception of small objects, Yu et al. [10] introduced BiSeNet, which adopts a dual-branch architecture to fuse local details and global semantics while strengthening shallow representations, thereby improving overall segmentation performance. Wang et al. [11] addressed the challenge of segmenting distant small objects in road scenes by integrating global contextual modeling and multi-scale feature fusion, effectively enhancing the representation of low-resolution targets. Wu et al. [12] developed a frequency-aware attention mechanism based on wavelet transforms, which theoretically benefits from multi-resolution analysis and provides a new paradigm for incorporating frequency-domain information. To improve generaliza-

*Corresponding author.

tion capability in complex traffic scenarios, Zhang et al. [13] incorporated a frequency modulation mechanism into the Transformer framework, where Fourier Transform is employed for frequency-domain modeling and adaptive modulation units are designed to fuse frequency and spatial representations, thereby strengthening feature expressiveness. Moreover, Niu et al. [14] emphasized that maintaining semantic consistency and stylistic diversity is crucial for enhancing model generalization, while such consistency heavily relies on accurate modeling of spatial relationships. Bai et al. [15] proposed a multi-scale dependency modeling strategy that improves scene understanding accuracy by constructing complex dependency networks at the expense of additional computational overhead. Li et al. [16] developed an occupancy network-based approach for 3D image semantic segmentation, which improves segmentation performance, particularly for multi-scale objects, through feature pyramids, channel attention mechanisms, and optimized backbone architectures.

The proposed method is not a straightforward combination of "frequency enhancement + spatial modeling + multi-scale fusion". Instead, it formulates an optimization paradigm in which frequency-domain structures explicitly guide spatial propagation and subsequently influence scale selection. The central innovation of this work is to elevate frequency information from an auxiliary enhancement cue to an intermediate structural representation for semantic reasoning. Specifically, amplitude-phase decoupling is first performed on encoded features to explicitly separate frequency-domain structures, where amplitudes are utilized for semantic energy modeling and phases serve as spatial structural constraints, while controlled perturbations are introduced to preserve geometric consistency. Subsequently, these structured frequency representations are not directly fed into the decoder but are instead employed as inputs to the Hierarchical Spatial Dependency Modeling (HSDM) module, which constrains local structural consistency and drives global graph-based semantic propagation. Finally, structural responses extracted from the original image are used to construct scale-selection signals independent of semantic features, thereby enabling structure-aware competitive fusion among multi-scale branches.

II. Related Theoretical Foundations

A. Encoder-Decoder-Based Framework

In semantic segmentation tasks, models are required not only to extract high-level semantic representations but also to preserve spatial location information and boundary details to achieve pixel-wise classification. Although conventional convolutional neural networks can enhance semantic representation capability through successive downsampling operations, the spatial resolution of feature maps is inevitably reduced, resulting in the loss of fine details and edge structures. Therefore, Encoder–Decoder architectures have been widely adopted in semantic segmentation to jointly accomplish semantic feature extraction and spatial information recovery through coordinated encoding and decoding processes.

The Encoder is primarily responsible for extracting deep semantic representations. The convolution operation can be formulated as:

$$\mathbf{Y} = \sigma(\mathbf{W} * \mathbf{X} + \mathbf{b}), \tag{1}$$

where, $\mathbf{X}$ and $\mathbf{Y}$ denote the input and output feature maps, respectively, $\mathbf{W}$ represents the convolution kernel parameters, $\mathbf{b}$ denotes the bias term, $*$ indicates the convolution operation, and $\sigma(\cdot)$ is a nonlinear activation function.

As the network depth increases, the Encoder acquires a larger receptive field and stronger semantic representation capability. However, repeated downsampling operations weaken high-frequency textures and boundary information, thereby degrading the segmentation performance for small objects and regions with complex boundaries.

The Decoder is designed to restore the spatial resolution of feature maps and reconstruct object structures. Common approaches include upsampling, transposed convolution, and interpolation operations. In addition, low-level detail features are often fused with high-level semantic representations to enhance boundary localization and fine-grained feature expression.

Encoder–Decoder architectures have become the dominant paradigm in semantic segmentation. Nevertheless, conventional convolution operations are inherently constrained by local receptive fields and struggle to model long-range dependencies effectively. Moreover, repeated downsampling may lead to the loss of high-frequency details. Therefore, enhancing global feature interactions while preserving fine structural information remains an important research direction in semantic segmentation.

B. DeepLab Series Networks

The DeepLab series improves pixel-level prediction accuracy by enlarging the receptive field through atrous convolutions while alleviating the loss of feature resolution.

1) Atrous convolution: Traditional convolutional networks typically enlarge the receptive field using pooling operations or strided convolutions. However, continuous downsampling often leads to the loss of boundary information and fine details of small objects. Atrous convolution expands the sampling range by introducing holes between convolution kernel elements, thereby increasing the receptive field without introducing additional parameters. Its operation can be expressed as

$$y(i) = \sum_{k} x(i + r \cdot k) w(k), \tag{2}$$

where, x and y denote the input and output feature maps, respectively, w represents the convolution kernel, and r is the dilation rate. When $r = 1$, atrous convolution degenerates into standard convolution. As the dilation rate increases, the receptive field expands accordingly.

Compared with conventional convolutions, atrous convolutions can capture richer contextual information while maintaining the spatial resolution of feature maps, thereby enhancing the representation capability for multi-scale objects and complex scenes.

2) Atrous spatial pyramid pooling: To improve the representation capability for objects of different scales, DeepLabV3 introduces the Atrous Spatial Pyramid Pooling (ASPP) module. ASPP employs parallel atrous convolution branches with different dilation rates to model multi-scale contextual information.

Different receptive fields enable ASPP to simultaneously capture local details and global semantic information. Smaller dilation rates focus on texture and boundary characteristics, whereas larger dilation rates aggregate broader contextual cues, thereby improving the adaptability to scale variations. A typical ASPP module consists of multiple atrous convolution branches with different dilation rates, a 1×1 convolution branch, a global average pooling branch, and a multi-scale feature fusion operation. The fusion process can be formulated as

$$F_{\mathrm{ASPP}} = \mathrm{Concat}\left(F_1, F_2, \ldots, F_n\right), \tag{3}$$

where, F_i denotes the feature map extracted by the i-th branch and $\mathrm{Concat}(\cdot)$ represents the channel-wise concatenation operation.

Although ASPP effectively enhances multi-scale contextual modeling, its feature extraction mechanism still primarily relies on local convolutional operations, limiting its ability to capture long-range dependencies. In addition, its capability to preserve high-frequency textures and recover small-object details remains insufficient.

3) DeepLabV3+ architecture: DeepLabV3+ extends DeepLabV3 by incorporating an Encoder–Decoder architecture to integrate high-level semantic representations with low-level spatial details, thereby improving segmentation performance for boundary regions and small objects.

In the Encoder stage, a backbone network is employed to extract high-level semantic features, while the ASPP module is adopted to enrich multi-scale contextual representations. Atrous convolutions replace part of the downsampling operations to enlarge the receptive field while preserving more spatial information.

In the Decoder stage, high-level features are first upsampled and then fused with shallow low-level features to enhance boundary refinement and spatial localization capability.

Although DeepLabV3+ has demonstrated superior performance in complex semantic segmentation scenarios, its feature modeling process still primarily relies on convolution operations, limiting its ability to capture long-range dependencies and global contextual interactions. Furthermore, repeated downsampling and sparse sampling induced by atrous convolutions may result in the loss of high-frequency details, thereby affecting segmentation accuracy for small objects and regions with intricate boundaries.

C. Visual Feature Representation Based on Spectral Analysis

1) Spatial-domain and frequency-domain representations: Traditional convolutional neural networks mainly perform feature extraction in the spatial domain, which is effective for modeling local textures and semantic information. However, their ability to characterize global frequency distributions remains limited.

In contrast, frequency-domain representations describe image structures from a global perspective. Different frequency components correspond to different visual characteristics. Low-frequency components capture overall semantic structures, whereas high-frequency components encode boundary, texture, and fine-detail information. Therefore, frequency-domain analysis provides complementary information for modeling both global semantics and fine-grained structures.

2) Two-dimensional discrete fourier transform: The two-dimensional Discrete Fourier Transform (DFT) is a fundamental tool for frequency-domain image analysis. It transforms spatial-domain images into the frequency domain and can be expressed as

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \exp\left[-j2\pi\left(\frac{ux}{M} + \frac{vy}{N}\right)\right], \tag{4}$$

where, $f(x,y)$ denotes the spatial-domain image, $F(u,v)$ represents its frequency-domain representation, and (u,v) denote frequency coordinates. Through DFT, images can be decomposed into different frequency components to characterize semantic and structural information from the frequency perspective.

3) High-frequency and low-frequency information: In frequency-domain representations, low-frequency information mainly reflects the global structure and semantic content of images, whereas high-frequency information contains boundary, texture, and local detail characteristics. High-frequency components are particularly important for object boundary recovery and small-object segmentation.

Therefore, jointly modeling low-frequency semantic information and high-frequency detail information is essential for improving segmentation performance in complex scenarios.

D. Spatial Dependency Modeling Theory

1) Importance of spatial contextual information: Semantic segmentation relies not only on local texture features but also on spatial relationships and contextual dependencies among pixels. In complex scenes, different semantic categories often exhibit structural correlations, such as road continuity, building consistency, and the spatial distributions of vehicles and pedestrians.

Therefore, relying solely on local features is insufficient for accurately interpreting complex scenes. Effective spatial contextual modeling can enhance scene understanding capability and improve the consistency and discriminability of segmentation results.

2) Limitations of convolutional networks: Conventional convolutional neural networks primarily perform local feature extraction through convolution operations. Although stacking multiple layers enlarges the receptive field, the underlying mechanism remains local in nature and is limited in modeling long-range spatial dependencies.

In complex scenarios, insufficient utilization of contextual information may impair the model's understanding of global structural relationships.

3) Existing spatial dependency modeling methods: To strengthen spatial contextual modeling, various approaches, including atrous convolutions and attention mechanisms, have been proposed. Atrous convolutions enlarge the sampling range to improve contextual perception, whereas attention mechanisms establish feature correlations to enhance information interactions across different regions.

However, while these global modeling strategies improve feature interactions, they may also introduce cross-region interference, resulting in semantic ambiguity and reduced boundary discriminability.

4) Spatial consistency problem: In complex scenes, small-object regions are susceptible to background interference, while global feature propagation may destabilize local feature representations. Excessive emphasis on global information may weaken local structural consistency, whereas relying solely on local features limits the utilization of contextual information.

Therefore, achieving collaborative modeling between local and global information while preserving structural consistency remains an important challenge in spatial dependency modeling.

E. Multi-Scale Feature Fusion Theory

1) Importance of multi-scale features: Features extracted from different network layers contain complementary visual information. Shallow features preserve boundary textures and local details due to their higher spatial resolution, whereas deep features possess larger receptive fields and stronger semantic representation capability.

Since features at different scales are complementary, relying on a single-level representation is insufficient to simultaneously capture global semantics and local details. Consequently, multi-scale feature fusion enhances the representation capability for complex scenes and objects with scale variations.

2) Classical multi-scale feature fusion methods: Various multi-scale fusion strategies have been proposed to improve adaptability to objects of different sizes. Feature Pyramid Networks (FPN) perform feature aggregation through top-down pathways and lateral connections, thereby enhancing multi-scale representations.

ASPP extracts contextual information at multiple scales using atrous convolutions with different dilation rates. Pyramid Scene Parsing Networks (PSPNet) aggregate global contextual information through multi-scale pooling structures to improve semantic representation in complex environments.

In addition, U-Net utilizes skip connections to fuse shallow spatial details with deep semantic features, effectively improving boundary recovery and local structure representation.

3) Feature fusion strategies: Current multi-scale fusion methods commonly adopt feature concatenation, weighted fusion, and attention-based fusion strategies.

Feature concatenation preserves information from different levels by channel-wise aggregation. Weighted fusion dynamically adjusts the importance of different features to strengthen critical representations. Attention-based fusion further employs channel-wise or spatial attention mechanisms to adaptively select informative features, thereby improving fusion effectiveness.

Through these strategies, models can achieve collaborative modeling of semantic and spatial information to a certain extent.

4) Limitations of existing multi-scale fusion methods: Although existing multi-scale fusion methods improve the representation capability for objects at different scales, several limitations remain. Due to significant differences between feature levels in terms of semantic content and spatial details, semantic conflicts may arise during fusion, thereby affecting feature consistency.

Moreover, small-object representations can be overwhelmed by dominant large-scale semantic information during deep propagation, leading to insufficient fine-grained discrimination capability. Existing fusion approaches also generally lack structural priors to guide feature selection and aggregation, making it difficult to fully exploit spatial structural information within scenes.

In addition, most methods still lack effective adaptive feature selection mechanisms and cannot dynamically adjust the importance of different scales according to varying scenarios. Therefore, achieving structure-guided adaptive multi-scale feature fusion has become an important direction for improving semantic segmentation performance in complex environments.

III. Methodology

The conventional DeepLabV3+ framework, built upon deep convolutional networks and Atrous Spatial Pyramid Pooling (ASPP), is capable of enlarging the receptive field and extracting multi-scale semantic representations, thereby achieving promising performance in traffic scene semantic segmentation tasks. However, the discrete sampling mechanism of atrous convolutions often leads to discontinuous responses for small objects and blurred object boundaries. In addition, single-scale feature representations struggle to simultaneously preserve local details and maintain global structural consistency, while the capability of cross-region semantic propagation remains limited. To address these issues, this study proposes a Frequency-Spatial Collaborative Segmentation framework (FSCS), which integrates frequency-domain enhancement with spatial collaborative modeling.

The overall architecture of the proposed method is illustrated in Fig. 1. First, a Spectral Decoupling Adaptive Modulation (SDAM) module is introduced after the encoder output. Through amplitude–phase decoupling and adaptive spectral modulation, SDAM enhances the responses of high-frequency features associated with small objects while preserving spatial structural continuity. Subsequently, a Hierarchical Spatial Dependency Modeling (HSDM) module is incorporated after the ASPP module. HSDM maintains fine-grained structural consistency through local neighborhood weighting and simultaneously achieves cross-region semantic propagation

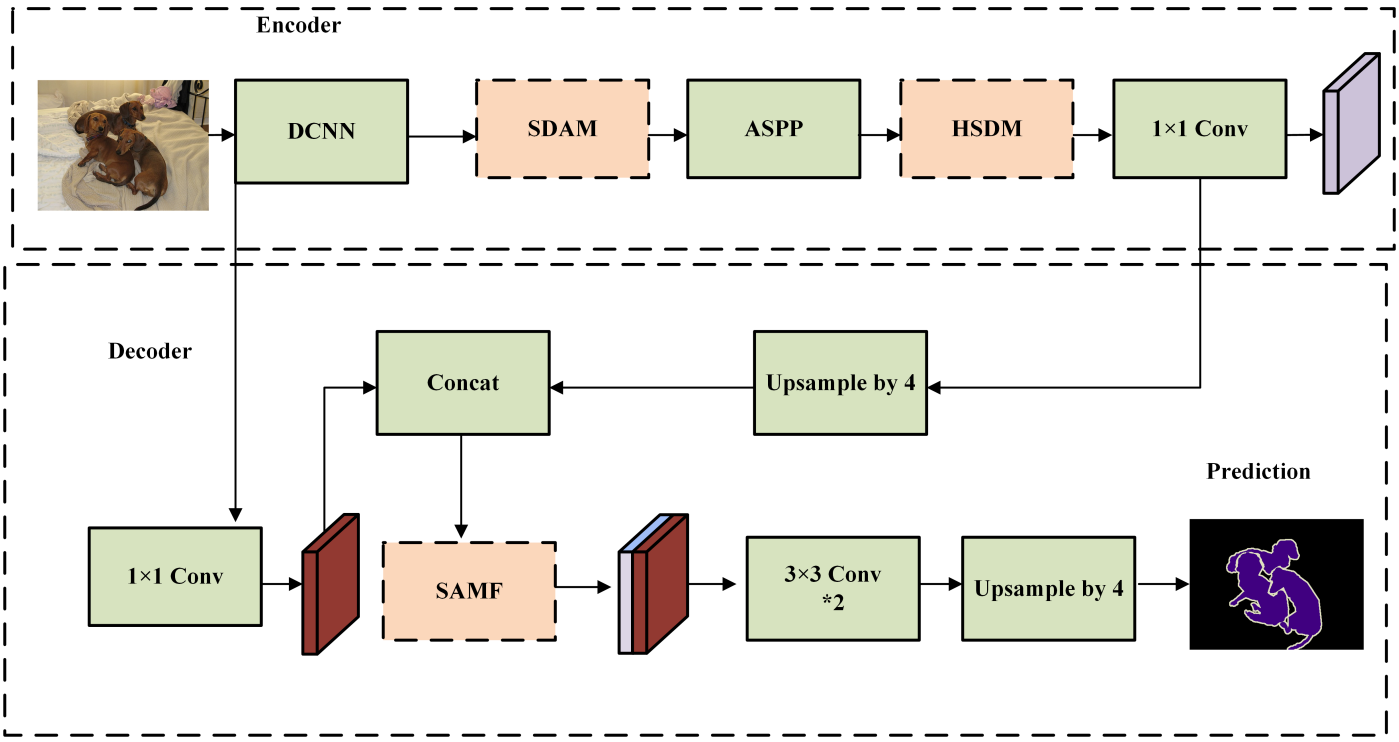

Fig. 1. Overall architecture of the proposed FSCS framework.

via global graph diffusion, thereby balancing local structural information and global semantic representations. Finally, a Structure-guided Adaptive Multi-scale Fusion (SAMF) module is embedded into the low-level feature fusion stage of the decoder. By utilizing structural responses generated from the original input image, SAMF dynamically adjusts the weights of features from different scales, enabling both boundary refinement and the preservation of semantic consistency for large-scale structures.

The three proposed modules are sequentially deployed after the encoder output, following the ASPP module, and within the decoder fusion stage, respectively. Together, they establish a complete processing pipeline ranging from frequency-domain enhancement to spatial dependency modeling and multi-scale feature fusion. Through collaborative optimization in both frequency and spatial domains, the proposed framework significantly improves semantic segmentation performance in complex traffic scenes.

A. Spectral-Decoupled Adaptive Modulation Module

Traditional convolutional neural networks primarily perform local feature extraction in the spatial domain, where the receptive field is inherently constrained by the kernel size. Consequently, such networks struggle to effectively capture long-range dependencies across spatial regions, limiting their ability to represent complex textures and fine-grained semantic structures. In traffic scene semantic segmentation, roads typically exhibit large-scale and spatially continuous structures, whereas lane markings, pedestrians, and traffic signs are characterized by prominent high-frequency components. This disparity in frequency characteristics makes it difficult for purely spatial-domain modeling approaches to simultaneously capture information distributed across different frequency bands.

The DeepLab series enlarges the receptive field through atrous convolutions, thereby improving multi-scale contextual representation. However, the discrete sampling mechanism of atrous convolutions may still result in discontinuous responses for small objects and produce “semantic holes” around object boundaries. As a consequence, the representation capability for fine structures and high-frequency details remains limited. Therefore, it is necessary to complement spatial-domain modeling with frequency-domain representations to enhance the expressive power of segmentation models in complex traffic environments.

To this end, a Spectral-Decoupled Adaptive Modulation (SDAM) module is introduced to perform frequency-domain enhancement on the encoded features. The overall architecture of SDAM is illustrated in Fig. 2. Specifically, the module transforms spatial-domain features into the frequency domain, where amplitude and phase components are explicitly decoupled. The amplitude spectrum is utilized to characterize the energy distribution associated with semantic representations at different frequency bands, while the phase spectrum preserves structural information and spatial relationships. By performing adaptive spectral modulation on the decomposed frequency components, SDAM selectively enhances informative high-frequency responses related to small objects and boundary details while retaining the global semantic consistency encoded in low-frequency components. Subsequently, the modulated spectral representations are transformed back into the spatial domain to generate refined feature maps.

Through explicit frequency-domain modeling and adaptive

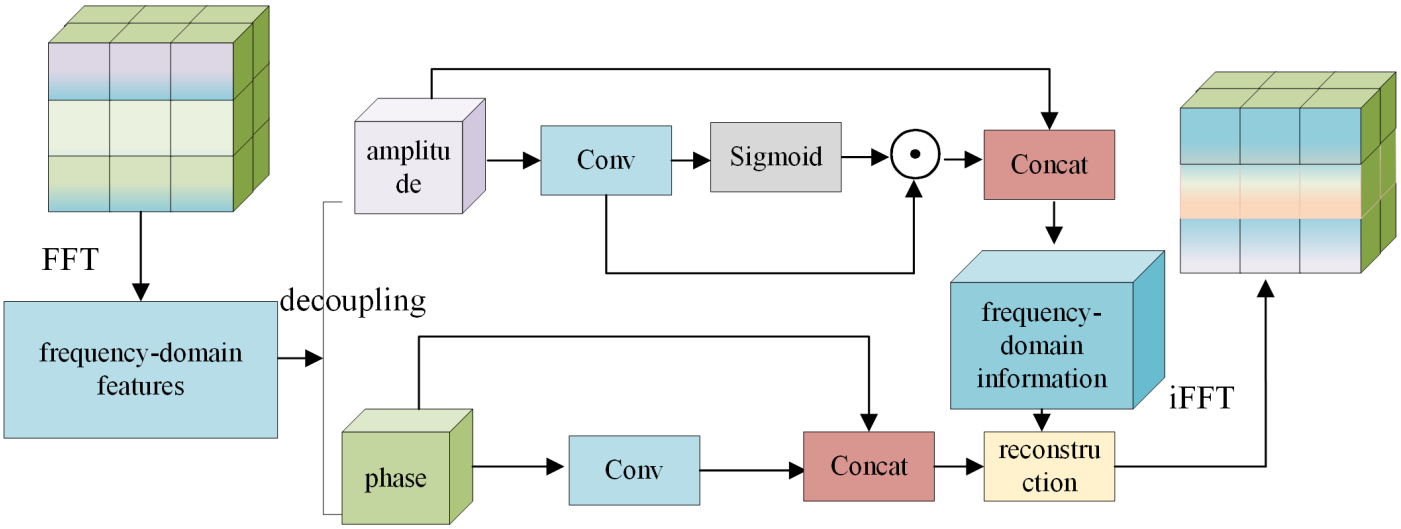

Fig. 2. Architecture of the proposed Spectral-Decoupled Adaptive Modulation (SDAM) module.

spectral interaction, SDAM effectively compensates for the limitations of conventional convolutional operations in capturing fine-grained structures. The proposed module enhances the discrimination capability for small targets and boundary regions while preserving the continuity of large-scale semantic structures, thereby providing more representative features for subsequent spatial dependency modeling and multi-scale feature fusion.

Given an input feature map $\mathbf{F} \in \mathbb{R}^{H\times W\times C}$, a two-dimensional Discrete Fourier Transform (DFT) is first applied to project the spatial-domain representation into the frequency domain:

$$\hat{\mathbf{F}} = \mathcal{F}(\mathbf{F}), \tag{5}$$

where, $\hat{\mathbf{F}} \in \mathbb{C}^{H\times W\times C}$ denotes the complex-valued frequency-domain representation containing responses from different frequency components.

To improve the controllability of frequency-domain modeling, the spectral representation is further decoupled into amplitude and phase components:

$$\mathbf{A} = |\hat{\mathbf{F}}|, \quad \mathbf{\Phi} = \angle\hat{\mathbf{F}}, \tag{6}$$

where, $\mathbf{A} \in \mathbb{R}^{H\times W\times C}$ represents the amplitude spectrum that characterizes the energy distribution across different frequency bands, while $\mathbf{\Phi} \in \mathbb{R}^{H\times W\times C}$ denotes the phase spectrum that encodes spatial structural information and plays a critical role in preserving boundaries and geometric configurations.

Subsequently, an adaptive modulation mechanism is introduced to dynamically redistribute spectral energy within the amplitude component:

$$\mathbf{A}' = \mathbf{A} \odot (1 + \sigma\left(\mathrm{Conv}(\mathbf{A})\right)), \tag{7}$$

where, $\mathrm{Conv}(\cdot)$ denotes a channel-wise convolution operation, $\sigma(\cdot)$ is the Sigmoid activation function, and $\odot$ represents element-wise multiplication. This residual scaling formulation enables adaptive reweighting of frequency components. On the one hand, the original spectral energy is preserved to prevent information loss; on the other hand, learnable weights are employed to enhance responses in informative frequency bands while suppressing redundant spectral components.

For the phase component, a residual modeling strategy is adopted to avoid disrupting spatial structural continuity:

$$\mathbf{\Phi}' = \mathbf{\Phi} + \Delta\mathbf{\Phi}, \tag{8}$$

where, the learnable phase perturbation is defined as

$$\Delta\mathbf{\Phi} = \mathrm{Conv}(\mathbf{\Phi}). \tag{9}$$

Here, $\Delta\mathbf{\Phi}$ introduces fine-grained adjustments to the phase spectrum while maintaining the continuity of the original structural representation, thereby preventing frequency-domain operations from degrading spatial geometric information.

From the perspective of frequency-domain reconstruction, the inverse Fourier transform is highly sensitive to phase variations. Consequently, uncontrolled perturbations may induce global structural distortions in the spatial domain. To alleviate this issue, the proposed residual phase modulation can be interpreted as a constrained optimization process within the local neighborhood of the original phase distribution. The magnitude of the phase update is regulated through explicit normalization and feature-scale constraints, ensuring that the perturbations remain within a controllable low-amplitude range.

Furthermore, from the viewpoint of frequency-domain noise propagation, random noise is typically manifested as the accumulation of inconsistent high-frequency phase responses. The residual formulation anchors phase updates to the original structural representation, effectively introducing structural priors into the optimization process. As a result, the perturbation

term primarily influences uncertain local boundary regions without amplifying spectral instability on a global scale. In addition, the normalization operation following the inverse transform further rebalances the spectral energy distribution, thereby suppressing distribution shifts induced by local phase perturbations and improving the stability of the frequency enhancement process.

After amplitude and phase modulation, the refined spectral components are recombined to reconstruct the enhanced frequency-domain representation:

$$\hat{\mathbf{F}}' = \mathbf{A}' \cdot e^{j\mathbf{\Phi}'}, \tag{10}$$

where, j denotes the imaginary unit.

The enhanced spectral representation is subsequently transformed back into the spatial domain using the inverse Fourier transform:

$$\mathbf{F}' = \mathcal{F}^{-1}(\hat{\mathbf{F}}'), \tag{11}$$

followed by normalization and point-wise convolution for feature reconstruction:

$$\mathbf{F}_{\text{out}} = \text{Conv}_{1\times1}\left(\text{Norm}(\mathbf{F}')\right), \tag{12}$$

where, $\text{Norm}(\cdot)$ denotes the normalization operation, which stabilizes the feature distribution and facilitates channel-wise feature reorganization.

Through amplitude–phase decoupling and controlled spectral modulation, SDAM enables fine-grained modeling of different frequency components. By enhancing informative semantic frequency bands while preserving the stability of structural representations, the proposed module provides structurally consistent and semantically enriched features for subsequent spatial dependency modeling and multi-scale feature fusion.

B. Hierarchical Spatial Dependency Modeling Module

In traffic scene semantic segmentation, spatial relationships are reflected not only in the structural continuity within local neighborhoods but also in semantic consistency across distant regions. Therefore, relying solely on frequency-domain enhancement or local convolution operations is insufficient to simultaneously enforce local structural constraints and capture long-range semantic dependencies. To address this issue, a Hierarchical Spatial Dependency Modeling (HSDM) module is proposed to progressively model spatial relationships from local to global levels and achieve unified feature representation through an adaptive fusion mechanism. The overall architecture of HSDM is illustrated in Fig. 3. The core idea of the proposed module is to first establish structurally consistent feature representations through local dependency modeling and subsequently perform global semantic propagation. Such a hierarchical strategy effectively alleviates the structural ambiguity introduced by direct global interactions, thereby achieving collaborative optimization between structural consistency and semantic coherence.

The input feature map of HSDM is obtained from the output of the SDAM module and can be represented as

$$\mathbf{F} \in \mathbb{R}^{H\times W\times C}, \tag{13}$$

where, H, W, and C denote the spatial dimensions and the number of feature channels, respectively. To facilitate global dependency modeling, the feature map is reshaped into a matrix form:

$$\mathbf{X} \in \mathbb{R}^{N\times C}, \quad N = H \times W, \tag{14}$$

where, each spatial location is treated as a graph node, transforming the spatial dependency modeling problem into a graph signal processing task.

1) Local structural consistency modeling: To preserve boundaries and fine-grained structures, local structural consistency is first established within neighborhood regions. For an arbitrary spatial location p, let $\mathcal{N}(p)$ denote its local neighborhood. The local feature response is obtained through weighted aggregation:

$$\mathbf{F}_{\text{loc}}(p) = \sum_{q\in\mathcal{N}(p)} \omega(p,q)\mathbf{F}(q), \tag{15}$$

where, the weighting coefficient $\omega(p,q)$ is defined as

$$\omega(p,q) = \frac{\exp\left(-\frac{\|\phi(\mathbf{F}(p))-\phi(\mathbf{F}(q))\|_2^2}{\sigma^2} - \lambda d(p,q)\right)}{\sum_{q'\in\mathcal{N}(p)}\exp\left(-\frac{\|\phi(\mathbf{F}(p))-\phi(\mathbf{F}(q'))\|_2^2}{\sigma^2} - \lambda d(p,q')\right)}, \tag{16}$$

where, $\phi(\cdot)$ denotes a 1×1 convolutional projection function, $d(p,q)$ represents the spatial distance between positions p and q, σ controls the feature similarity scale, and λ is the structural regularization coefficient. This weighted aggregation process is equivalent to bilateral filtering in graph feature space, enabling effective local structural preservation.

2) Global semantic modeling: Based on locally aligned structural representations, semantic information is further propagated over the entire graph. The locally consistent features are regarded as graph signals, and global semantic diffusion is performed over a weighted graph:

$$\mathbf{F}_{\text{pair}}(i) = \sum_{j=1}^{N} W_{ij}\mathbf{F}_{\text{loc}}(j), \tag{17}$$

where, the affinity matrix $\mathbf{W}$ is defined as

$$W_{ij} = \frac{\exp\left(-\frac{\|\phi(\mathbf{F}_{\text{loc}}(i))-\phi(\mathbf{F}_{\text{loc}}(j))\|_2^2}{\sigma^2}\right)}{\sum_{j'=1}^{N}\exp\left(-\frac{\|\phi(\mathbf{F}_{\text{loc}}(i))-\phi(\mathbf{F}_{\text{loc}}(j'))\|_2^2}{\sigma^2}\right)}. \tag{18}$$

This formulation corresponds to a graph diffusion process that facilitates global semantic propagation while reducing the risk of information leakage across structurally inconsistent regions.

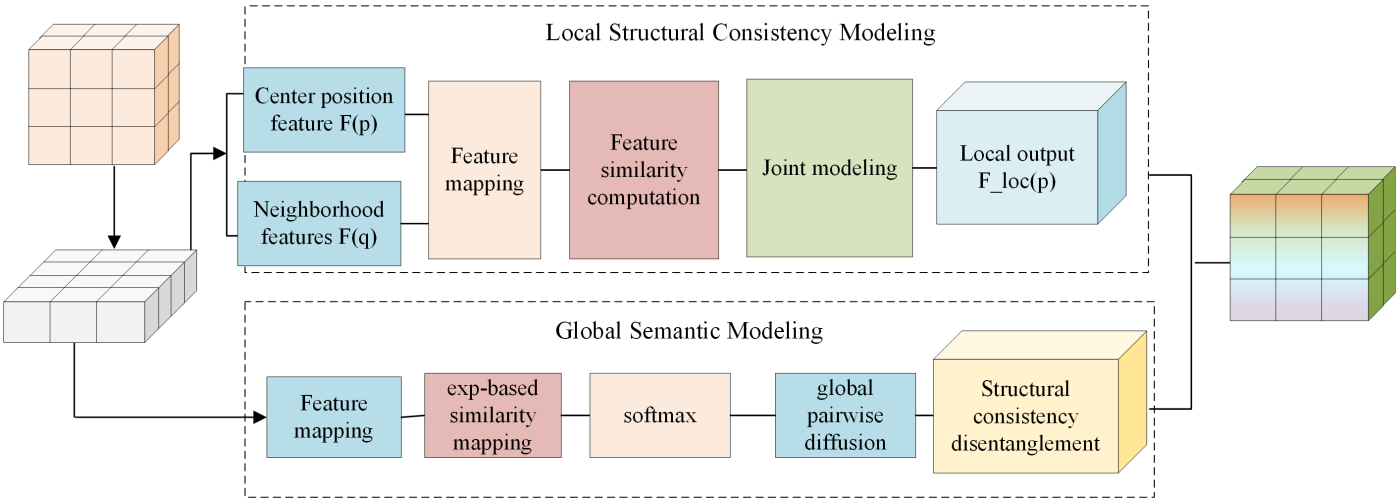

Fig. 3. Architecture of the proposed Hierarchical Spatial Dependency Modeling (HSDM) module.

3) Structure-semantic decoupled modeling: To mitigate the over-smoothing effect caused by global propagation, a decoupling strategy is introduced to separate the global response into relational and unary components:

$$\mathbf{F}_{\text{glob}} = \mathbf{F}_{\text{pair}} + \mathbf{F}_{\text{unary}}, \tag{19}$$

where, the unary term is defined as

$$\mathbf{F}_{\text{unary}}(i) = \mathbf{W}_u \mathbf{F}(i), \tag{20}$$

with, $\mathbf{W}_u$ denoting learnable transformation parameters. The unary component preserves the intrinsic saliency of each spatial position and corresponds to the unary term in graphical models, whereas $\mathbf{F}_{\text{pair}}$ represents the pairwise term responsible for modeling cross-region semantic relationships.

4) Adaptive fusion: Considering that different spatial regions exhibit varying dependencies on local structural information and global semantic context, a spatially adaptive fusion mechanism is further introduced:

$$\boldsymbol{\alpha} = \sigma\left(\text{Conv}_{1\times1}\left([\mathbf{F}_{\text{loc}}, \mathbf{F}_{\text{glob}}]\right)\right), \tag{21}$$

where, $[\cdot,\cdot]$ denotes channel-wise concatenation and $\sigma(\cdot)$ represents the Sigmoid activation function.

The final output feature representation is obtained by

$$\mathbf{F}_{\text{out}} = \boldsymbol{\alpha} \odot \mathbf{F}_{\text{loc}} + (1 - \boldsymbol{\alpha}) \odot \mathbf{F}_{\text{glob}}, \tag{22}$$

where, $\odot$ denotes element-wise multiplication.

The proposed adaptive fusion strategy utilizes local structural representations to guide feature integration. Specifically, the contribution of $\mathbf{F}_{\text{loc}}$ is emphasized in boundary regions to preserve structural details, whereas the contribution of $\mathbf{F}_{\text{glob}}$ is strengthened in semantically homogeneous regions to improve contextual consistency. Consequently, HSDM achieves a dynamic balance between structural information and semantic representations, providing more discriminative features for subsequent decoding and segmentation tasks.

C. Structure-Guided Adaptive Multi-Scale Fusion Module

In traffic scenes, different semantic categories exhibit not only scale variations but also highly heterogeneous spatial distributions and contextual dependencies. Small-scale objects, such as pedestrians and distant vehicles, are often located near scene boundaries or within occluded regions, making their representation highly dependent on local details and boundary sensitivity. In contrast, large-scale structures, such as buildings and road regions, usually present spatial continuity and require global contextual information to maintain semantic consistency. Moreover, variations in object density, viewpoint, and spatial arrangement further exacerbate the imbalance between local details and global semantics, making it difficult for single-scale representations to simultaneously capture fine-grained structures and overall scene context.

Under such circumstances, naive multi-scale feature fusion may introduce undesirable interference. Specifically, excessive global information may dilute local structural cues in complex boundary regions, whereas overemphasizing local features in large homogeneous regions may undermine semantic consistency. Therefore, a structure-guided adaptive fusion strategy is required to dynamically determine the contribution of features at different scales according to spatial locations and contextual characteristics. Such a mechanism can effectively balance local discriminative capability and global semantic coherence, thereby providing robust representations for semantic segmentation in complex traffic scenarios. The overall architecture of the proposed Structure-guided Adaptive Multi-scale Fusion (SAMF) module is illustrated in Fig. 4.

1) Multi-scale feature construction and alignment: Let the input feature representation be defined as

$$\mathbf{X} \in \mathbb{R}^{H\times W\times C}. \tag{23}$$

Multi-scale feature branches are constructed through scale transformation operations:

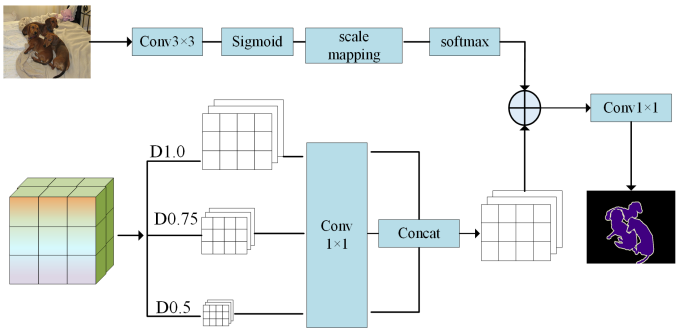

Fig. 4. Architecture of the proposed Structure-guided Adaptive Multi-scale Fusion (SAMF) module.

$$\mathbf{X}_s = D(\mathbf{X}, s), \quad s \in \{1, 0.75, 0.5\}, \tag{24}$$

where, $D(\cdot)$ denotes the scale transformation operation. Different scales correspond to different receptive field sizes. Larger scales preserve fine-grained details, whereas smaller scales facilitate the modeling of global contextual information.

To eliminate discrepancies in channel distributions across different scales, point-wise convolutions are employed for feature alignment:

$$\mathbf{X}'_s = \mathrm{Conv}_{1\times1}(\mathbf{X}_s). \tag{25}$$

Subsequently, all scale-specific features are upsampled to a unified spatial resolution:

$$\tilde{\mathbf{X}}_s = \mathrm{Up}(\mathbf{X}'_s), \tag{26}$$

resulting in a set of spatially aligned multi-scale representations.

2) Structure-guided signal modeling from the original input: To avoid information coupling among multi-scale representations during the fusion process, the proposed method does not directly derive structural responses from the fused features. Instead, the original input image is introduced as an external structural prior:

$$\mathbf{X}_0 \in \mathbb{R}^{H\times W\times C_0}, \tag{27}$$

where, C_0 denotes the number of channels in the original input.

A local convolution operation is then applied to capture structural variations:

$$\mathbf{S} = \sigma\left(\mathrm{Conv}_{3\times3}(\mathbf{X}_0)\right), \tag{28}$$

where, $\mathbf{S} \in \mathbb{R}^{H\times W\times1}$ represents the structural response map, and $\sigma(\cdot)$ denotes the Sigmoid activation function used to constrain the output range and improve training stability.

The structural response characterizes the intensity of edge and texture variations in the input image, thereby serving as a guidance signal independent of high-level semantic representations. This strategy effectively reduces mutual interference among multi-scale features during fusion.

3) Structure-guided scale response modeling: Based on the structural response map, the structure information is projected into the scale selection space:

$$\tilde{\mathbf{w}}(p) = \phi\left(\mathbf{S}(p)\right), \tag{29}$$

where, $\tilde{\mathbf{w}}(p) \in \mathbb{R}^K$ denotes the unnormalized scale response at spatial position p, K is the number of scale branches, and $\phi(\cdot)$ is implemented using a 1×1 convolution. This operation converts structural cues into scale-selection logits, establishing an explicit mapping from structural characteristics to receptive field preferences.

Through this mechanism, the model is able to adaptively adjust the importance of different scales according to local structural variations.

4) Scale normalization and competition mechanism: To ensure comparability among scale responses, a normalization operation is performed across the scale dimension:

$$w_s(p) = \frac{\exp\left(\tilde{w}_s(p)\right)}{\sum_{s'}\exp\left(\tilde{w}_{s'}(p)\right)}, \tag{30}$$

where, $w_s(p)$ denotes the relative importance of scale s at position p.

The normalization process introduces a competition mechanism among different scales, enabling the model to dynamically select the most appropriate receptive field according to spatial characteristics, thereby improving the adaptability of feature representations.

5) Structure-guided multi-scale fusion: Based on the learned scale weights, the aligned multi-scale features are aggregated as

$$\mathbf{X}_o(p) = \sum_s w_s(p)\cdot\tilde{\mathbf{X}}_s(p), \tag{31}$$

where, $\tilde{\mathbf{X}}_s$ denotes the feature representation corresponding to scale branch s.

Although all features have been spatially aligned to a unified resolution, their receptive field characteristics remain distinct. Consequently, the fusion process can be interpreted as an adaptive combination of contextual information with different spatial extents under a common coordinate system.

Since the fusion weights are derived from the original input rather than the feature representations themselves, the proposed strategy effectively alleviates feature coupling across scales and improves structural consistency in the fused representation.

Finally, a lightweight nonlinear transformation is applied to obtain the output feature representation:

$$\mathbf{X}_{\mathrm{out}} = \Psi\left(\mathbf{X}_o\right), \tag{32}$$

where, $\Psi(\cdot)$ consists of a point-wise convolution followed by an activation function for channel reorganization and feature enhancement.

Through the proposed structure-guided adaptive fusion mechanism, SAMF dynamically balances the contributions of different receptive fields according to structural characteristics in the input image. As a result, local discriminative information is preserved in fine-grained regions, while global semantic consistency is maintained in large homogeneous areas, thereby yielding more robust feature representations for traffic scene semantic segmentation.

IV. Experiments

A. Datasets and Evaluation Metrics

1) Datasets:

a) Cityscapes [17]: Cityscapes is a standard benchmark dataset for urban scene understanding and has been widely adopted in autonomous driving and semantic segmentation research. The dataset contains complex street-view images collected from 50 different cities. In total, it provides 5,000 finely annotated images with pixel-level labels, which are divided into 2,975 training images, 500 validation images, and 1,525 testing images. All images are extracted from vehicle-mounted stereo video sequences and maintain a consistent spatial resolution, ensuring comparability in semantic segmentation tasks. Cityscapes originally provides annotations for more than 30 semantic categories, covering roads, buildings, vehicles, pedestrians, and natural backgrounds. Following the common evaluation protocol, 19 representative categories are selected as the standard benchmark subset for fair comparison among different methods.

b) CamVid [18]: CamVid is one of the earliest semantic segmentation datasets designed for driving scene understanding. It is primarily used for pixel-level semantic annotation in road environments. The dataset consists of 701 manually annotated images extracted from driving videos, including 367 training images, 101 validation images, and 233 testing images. All images have a resolution of 960×720. CamVid defines 32 semantic categories covering roads, buildings, vehicles, pedestrians, traffic signs, and other common traffic scene elements.

2) Evaluation metrics: To comprehensively evaluate the effectiveness and efficiency of the proposed method, four widely used metrics are adopted, including Mean Intersection over Union (MIoU), Mean Accuracy (Mean Acc), model parameters (Params), floating-point operations (FLOPs), and Frames Per Second (FPS).

a) Mean Intersection over Union (MIoU): MIoU is the primary metric for evaluating semantic segmentation accuracy. It measures the overlap ratio between predicted segmentation results and ground-truth annotations by averaging the Intersection over Union (IoU) across all categories:

$$\mathrm{MIoU} = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i + FN_i}, \tag{33}$$

where, N denotes the number of semantic categories, and TP_i, FP_i, and FN_i represent the numbers of true positive, false positive, and false negative pixels for the i-th category, respectively.

b) Mean Accuracy (Mean Acc): Mean Accuracy evaluates the average classification accuracy over all categories and is defined as

$$\mathrm{MeanAcc} = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FN_i}, \tag{34}$$

where, TP_i denotes the number of correctly classified pixels belonging to category i, and FN_i represents the number of pixels from category i that are incorrectly classified into other categories.

c) Parameters (Params): Params quantify the structural complexity and storage requirements of a model, corresponding to the total number of learnable parameters. For a convolutional layer, the number of parameters can be expressed as

$$\mathrm{Params} = K^2 \times C_{\mathrm{in}} \times C_{\mathrm{out}}, \tag{35}$$

where, K is the kernel size, and C_{in} and C_{out} denote the numbers of input and output channels, respectively.

Although a larger number of parameters generally indicates stronger representation capability, it may also increase storage costs and the risk of overfitting.

d) Floating Point Operations (FLOPs): FLOPs are used to measure the computational complexity during the forward inference stage. For convolution operations, the computational cost is defined as

$$\mathrm{FLOPs} = H \times W \times K^2 \times C_{\mathrm{in}} \times C_{\mathrm{out}}, \tag{36}$$

where, H and W denote the spatial dimensions of the output feature map. This formulation indicates that computational complexity is jointly determined by kernel size, channel dimensions, and feature resolution.

e) Frames Per Second (FPS): FPS is adopted to evaluate inference efficiency and reflects the number of images processed per second. It is defined as

$$\mathrm{FPS} = \frac{\text{Number of processed images}}{\text{Total inference time (s)}}. \tag{37}$$

A higher FPS value indicates better real-time performance.

B. Experimental Settings

During training, the Stochastic Gradient Descent (SGD) optimizer was employed to update the network parameters, while the Cross-Entropy Loss function was adopted for optimization. The batch size was set to 8, and the number of data loading workers was fixed at 4. Considering that different network architectures exhibit varying sensitivities to input resolutions, the crop size for each model was adjusted according to its

original implementation settings. In addition, the weight decay coefficient was set to 1×10^{-4}, the initial learning rate was initialized to 0.01, and the momentum parameter was fixed at 0.9 to accelerate convergence and improve training stability (Table I).

TABLE I. EXPERIMENTAL ENVIRONMENT CONFIGURATION

Category	Parameter Name	Setting Value
Hardware Environment	Operating System	Windows 11
	CPU	Intel i7-14650HX
	GPU	NVIDIA RTX 4090 (24 GB)
	Memory	24 GB
Software Environment	Programming Language	Python
	Development Environment	PyCharm

C. Ablation Studies

A series of ablation experiments were conducted on the Cityscapes dataset to comprehensively evaluate the effectiveness of each component in the proposed FSCS framework. The experimental results are summarized in Table II. As shown in the table, the complete FSCS model achieves the best segmentation performance while introducing only a slight increase in computational cost.

Specifically, incorporating the Spectral-Decoupled Adaptive Modulation (SDAM) module independently improves the mIoU and Mean Acc by approximately 1.44% and 1.21%, respectively, compared with the baseline model. This observation indicates that frequency-domain enhancement effectively strengthens the representation capability for fine-grained structures and small objects. When the Hierarchical Spatial Dependency Modeling (HSDM) module is introduced, the model further benefits from local structural consistency and global semantic propagation, achieving improved segmentation accuracy while maintaining favorable real-time performance. The Structure-guided Adaptive Multi-scale Fusion (SAMF) module introduces structural priors into the multi-scale fusion process and obtains the highest Mean Acc among the single-module variants, reaching 0.7612. Although the mIoU improvement brought by SAMF alone is relatively limited, it effectively enhances feature consistency and boundary discrimination.

Overall, the proposed FSCS framework achieves the best trade-off between segmentation accuracy, computational efficiency, and real-time inference capability. By integrating frequency-domain enhancement, hierarchical spatial dependency modeling, and structure-guided multi-scale fusion, the complete model improves segmentation performance while preserving the lightweight characteristics required for practical traffic scene applications.

TABLE II. ABLATION STUDY RESULTS ON THE CITYSCAPES DATASET

Algorithm	Params (M)	GFLOPs	mIoU	Mean Acc	FPS
Baseline	5.23	17.00	0.6781	0.7414	231.40
Baseline + SDAM	4.81	19.66	0.6925	0.7535	200.51
Baseline + HSDM	4.39	20.16	0.6891	0.7591	210.11
Baseline + SAMF	4.21	21.23	0.6879	0.7612	208.10
Ours (FSCS)	4.96	22.50	0.7146	0.7665	210.23

D. Comparison Experiments

To systematically evaluate the generalization capability and segmentation performance of the proposed method in different urban scenarios, comparative experiments were conducted on the Cityscapes and CamVid datasets. Representative semantic segmentation models, including FCN-8s, U-Net, PSPNet, and DeepLabV3+, were selected as baseline approaches. The input resolutions were set to 2048×1024 for Cityscapes and 960×720 for CamVid. To ensure fairness and reproducibility, all models were trained using identical data preprocessing procedures and optimization strategies.

1) Comparison on the cityscapes dataset: The Cityscapes dataset contains high-resolution images with complex urban structures, posing significant challenges for semantic segmentation models. As reported in Table III, the proposed FSCS framework consistently outperforms conventional methods in terms of segmentation accuracy.

Specifically, FSCS achieves an mIoU of 0.7146, outperforming FCN-8s, U-Net, and PSPNet by approximately 10.75%, 9.41%, and 7.49%, respectively. Compared with DeepLabV3+, FSCS yields an absolute mIoU improvement of 3.65%, demonstrating the effectiveness of integrating frequency-domain enhancement, hierarchical spatial modeling, and structure-guided multi-scale fusion.

Regarding model complexity, FSCS contains only 4.96M parameters, which is lower than that of DeepLabV3+ and substantially smaller than those of FCN-8s and PSPNet. Although the computational cost increases slightly compared with DeepLabV3+, the proposed method still maintains a high inference speed of approximately 210 FPS. These results indicate that FSCS achieves a favorable balance between segmentation accuracy and computational efficiency, making it suitable for real-time semantic segmentation tasks in autonomous driving applications.

TABLE III. PERFORMANCE COMPARISON ON CITYSCAPES DATASET

Algorithm	Params (M)	GFLOPs	mIoU	Mean Acc	FPS
FCN-8s [18]	134.46	223.25	0.6071	0.7275	34.52
U-Net [19]	26.36	228.24	0.6205	0.7310	56.16
PSPNet [2]	51.44	190.34	0.6397	0.7365	64.70
DeepLabV3+ [20]	5.23	17.00	0.6781	0.7414	231.40
Ours (FSCS)	4.96	22.50	0.7146	0.7665	210.23

2) Comparison on the camvid dataset: The CamVid dataset contains fewer training samples but exhibits a relatively balanced category distribution, making it suitable for evaluating the generalization ability of segmentation models under limited data conditions. As shown in Table IV, the proposed FSCS framework consistently achieves superior performance compared with existing approaches.

In particular, FSCS demonstrates enhanced segmentation capability for small objects, such as pedestrians and traffic signs, as well as slender structures including curbs and poles. The proposed method attains the highest mIoU and Mean Acc among all competing methods, achieving values of 0.6342 and 0.7325, respectively. Compared with DeepLabV3+, FSCS yields absolute improvements of 1.65% in mIoU and 1.41% in Mean Acc, indicating that the proposed frequency-spatial collaborative framework remains effective even on relatively small-scale datasets.

In terms of computational efficiency, FSCS maintains a lightweight architecture with only 4.96M parameters and

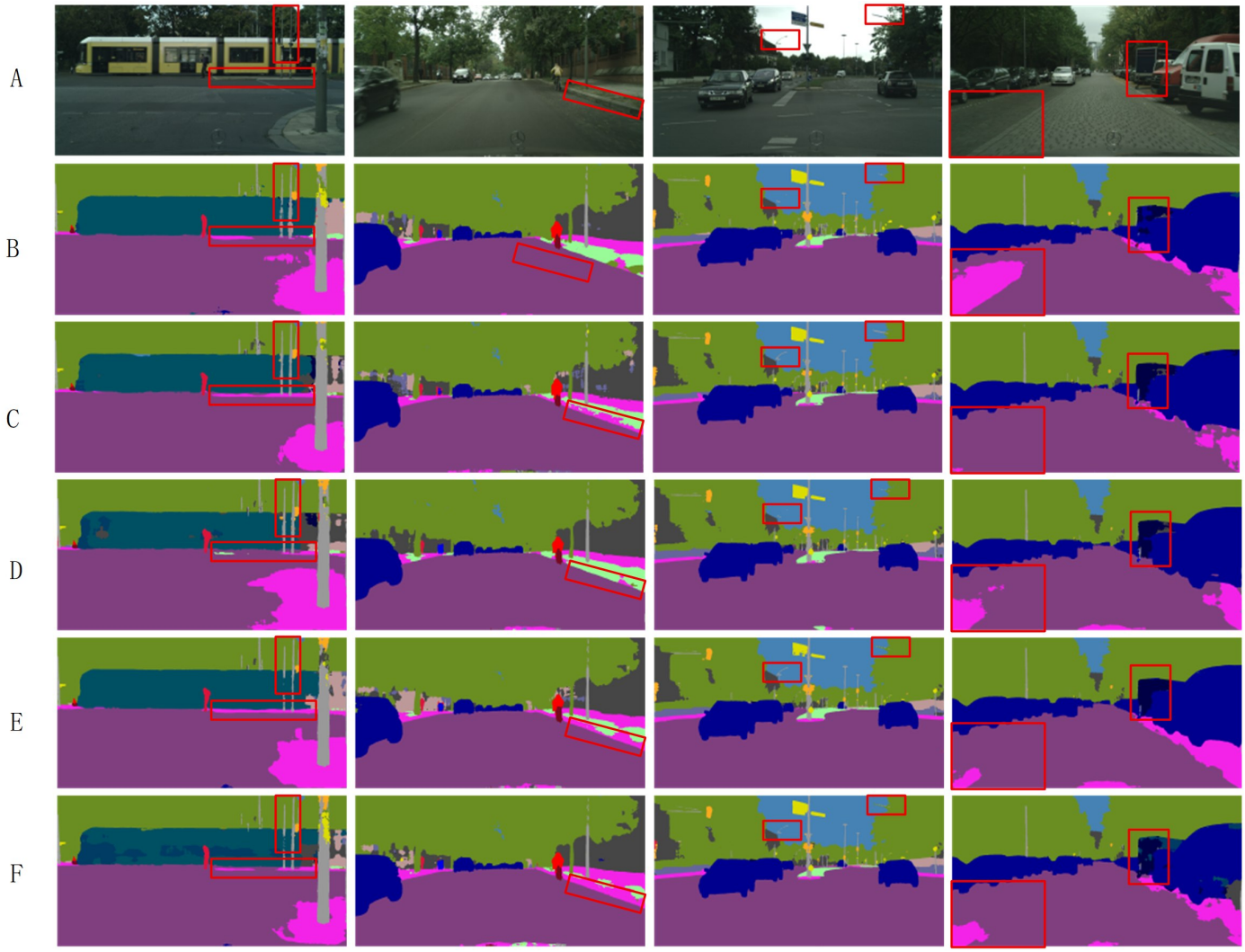

Fig. 5. Qualitative comparison of semantic segmentation results produced by different methods on the Cityscapes dataset.

18.86 GFLOPs. Despite a moderate increase in computational complexity compared with DeepLabV3+, the proposed model still achieves an inference speed of approximately 213 FPS, substantially outperforming FCN-8s, U-Net, and PSPNet. These results demonstrate that FSCS successfully balances segmentation accuracy and inference efficiency, making it well suited for real-time traffic scene understanding applications.

TABLE IV. PERFORMANCE COMPARISON ON THE CAMVID DATASET

Algorithm	Params (M)	GFLOPs	mIoU	Mean Acc	FPS
FCN-8s	134.46	182.12	0.5613	0.6914	38.33
U-Net	26.36	196.78	0.5910	0.7012	68.29
PSPNet	51.44	162.39	0.6017	0.7132	90.46
DeepLabV3+	5.22	14.92	0.6177	0.7184	235.48
Ours (FSCS)	4.96	18.86	0.6342	0.7325	213.45

3) Comprehensive analysis: The experimental results on both the Cityscapes and CamVid datasets demonstrate that the proposed FSCS framework achieves stable and competitive segmentation performance under different image resolutions and scene complexities. The performance gains can be primarily attributed to the complementary effects of the three proposed modules.

Specifically, SDAM performs amplitude–phase decoupling and adaptive spectral modulation in the frequency domain, effectively enhancing the representation capability for small objects and fine-grained structures. HSDM models spatial dependencies hierarchically through local structural consistency constraints and global semantic propagation, thereby preserving spatial continuity while capturing long-range contextual relationships. Furthermore, SAMF introduces structural priors into the multi-scale fusion process, dynamically integrating features with different receptive fields to improve the segmentation quality of object boundaries and slender structures.

The complete FSCS framework consistently outperforms all individual module variants, indicating that the three modules provide complementary rather than redundant contributions. Although the proposed method incurs a slight increase in computational cost compared with the baseline, it achieves substantial improvements in segmentation accuracy while maintaining high inference efficiency. These findings verify the effectiveness of the proposed frequency-spatial collaborative optimization strategy for real-time semantic segmentation in

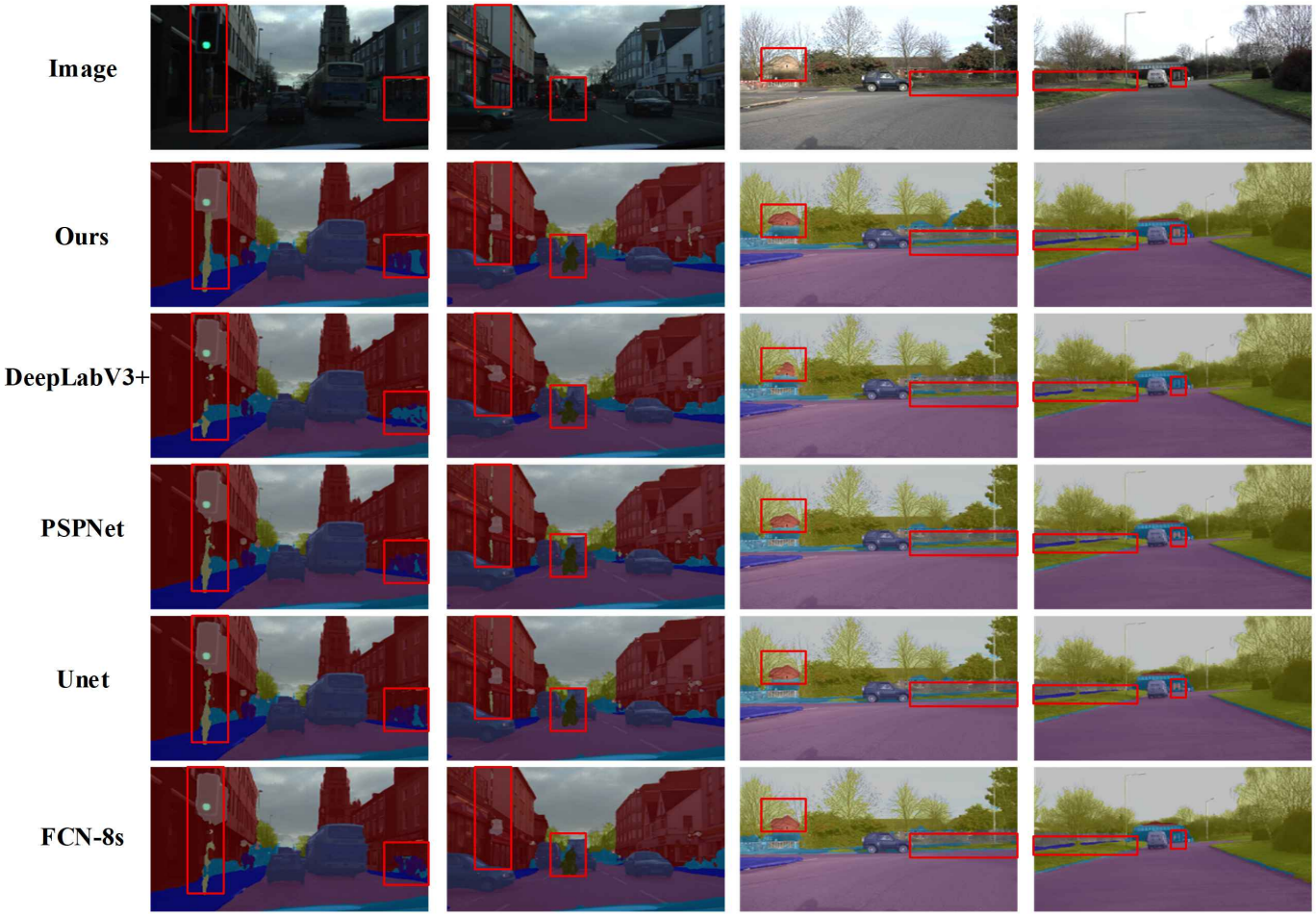

Fig. 6. Qualitative comparison of semantic segmentation results produced by different methods on the CamVid dataset.

complex traffic environments.

4) Qualitative comparison: Fig. 5 and 6 present qualitative comparisons of semantic segmentation results obtained by different methods on the Cityscapes and CamVid datasets, respectively.

As illustrated in Fig. 5, the baseline model exhibits noticeable omissions and blurred boundaries when segmenting fine-grained objects and distant targets. U-Net achieves satisfactory performance on large-scale structures but struggles to accurately delineate small objects and elongated regions. FCN-8s demonstrates the weakest performance in terms of small-object recognition and boundary localization, while also exhibiting slight confusion in certain large-scale regions. PSPNet improves the segmentation quality of fine structures to some extent; however, misclassification and omission of distant objects remain evident.

In contrast, the proposed FSCS framework produces the most accurate segmentation results within the highlighted regions. Specifically, FSCS successfully identifies small objects, slender structures, and distant vehicles while preserving the continuity of large-scale semantic regions such as roads and buildings. These observations further confirm that the collaborative optimization of frequency-domain enhancement, hierarchical spatial modeling, and structure-guided multi-scale fusion effectively improves both local discrimination capability and global semantic consistency.

Similarly, the visualization results on the CamVid dataset, shown in Fig. 6, further validate the robustness and generalization capability of the proposed method under different scene characteristics and dataset scales.

V. Conclusion

In this study, a Frequency-Spatial Collaborative Segmentation framework (FSCS) is proposed to improve the performance of DeepLabV3+ for traffic scene semantic segmentation. To address the limitations of conventional convolution-based segmentation methods in modeling fine-grained structures, long-range dependencies, and multi-scale contextual information, three dedicated modules are introduced.

First, the Spectral-Decoupled Adaptive Modulation (SDAM) module explicitly performs amplitude–phase decoupling in the frequency domain and adopts adaptive spectral modulation to enhance informative frequency components while preserving structural consistency, thereby improving the representation capability for small objects and boundary regions. Second, the Hierarchical Spatial Dependency Modeling (HSDM) module progressively captures spatial dependencies from local structural consistency to global

semantic propagation, enabling effective long-range context modeling while alleviating the structural ambiguity caused by direct global interactions. Finally, the Structure-guided Adaptive Multi-scale Fusion (SAMF) module introduces structural priors derived from the original input image to dynamically guide the fusion of multi-scale features, thereby enhancing the segmentation quality of slender structures and complex boundaries.

Experimental results on the Cityscapes and CamVid datasets demonstrate that the proposed FSCS framework consistently outperforms several representative semantic segmentation methods while maintaining favorable real-time performance and computational efficiency. The ablation studies further verify that the three proposed modules contribute complementary improvements and jointly enhance segmentation performance in complex traffic environments.

In future work, more efficient frequency-domain representation strategies and lightweight spatial modeling mechanisms will be investigated to further reduce computational complexity while improving the segmentation accuracy of small-scale and cross-scale objects. In addition, extending the proposed framework to more challenging scenarios, such as adverse weather conditions and domain-generalized traffic scene understanding, remains an important direction for future research.

REFERENCES

[1] Y. Wang *et al.*, "An automated learning method of semantic segmentation for train autonomous driving environment understanding," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 4, pp. 6913–6922, 2024.

[2] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[3] C. Song *et al.*, "Road scene semantic segmentation based on mpnet," *Electronics*, vol. 14, no. 13, p. 2565, 2025.

[4] L. Shan, X. Li, and W. Wang, "Decouple the high-frequency and low-frequency information of images for semantic segmentation," in *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 1805–1809.

[5] T. Zhang *et al.*, "Adaptive semantic segmentation of traffic scenes via frequency domain analysis," in *2025 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2025.

[6] C. Ge *et al.*, "Sf-net: Spatial-frequency feature synthesis for semantic segmentation of high-resolution remote sensing imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2026.

[7] Z. Yang *et al.*, "Ffr: Frequency feature rectification for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[8] J. Huang *et al.*, "Fsdr: Frequency space domain randomization for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6891–6902.

[9] H. Sima, M. Gao, and L. Liu, "A real-time semantic segmentation network leveraging spatial and contextual features for enhanced scene understanding," *Intelligent Systems with Applications*, vol. 27, p. 200542, 2025.

[10] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 325–341.

[11] H. Wang *et al.*, "Context-aware method for small object segmentation in road scenes," in *2022 International Conference on Advanced Robotics and Mechatronics (ICARM)*. IEEE, 2022.

[12] X. Wu *et al.*, "Wavelet attention fusion for semi-supervised ultrasound segmentation," *Biomedical Signal Processing and Control*, vol. 117, p. 109566, 2026.

[13] Y. Zhang *et al.*, "Segcft: Context-aware fourier transform for efficient semantic segmentation," *Neurocomputing*, vol. 596, p. 127946, 2024.

[14] H. Niu *et al.*, "Exploring semantic consistency and style diversity for domain generalized semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 6, 2025.

[15] Y. Bai, Z. Huang, L. Shen *et al.*, "Multi-scale feature aggregation by cross-scale pixel-to-region relation operation for semantic segmentation," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5889–5896, 2021.

[16] T. Li, X. Tang, H. Wang *et al.*, "Occupancy network-based 3d semantic segmentation method for autonomous driving images," *Laser & Optoelectronics Progress*, vol. 63, no. 4, pp. 192–203, 2026.

[17] M. Cordts, M. Omran, S. Ramos *et al.*, "The cityscapes dataset for semantic urban scene understanding," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3213–3223, 2016.

[18] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.