Scientific Reports

https://doi.org/10.1038/s41598-026-39052-y

Article in Press

RFGLNet for adverse weather domain-generalized semantic segmentation with frequency low-rank enhancement

Received: 24 December 2025

Accepted: 2 February 2026

Published online: 11 February 2026

Cite this article as: Ye X., Shi X. & Li Y. RFGLNet for adverse weather domain-generalized semantic segmentation with frequency low-rank enhancement. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-39052-y

Xin Ye, Xiaoqi Shi & Yuxue Li

RFGLNet for Adverse Weather Domain-Generalized Semantic Segmentation with Frequency Low-Rank Enhancement

Xin Ye,[1] Xiaoqi Shi,[1] and Yuxue Li[2]

[1] Xi'an Technological University, Xi'an, China.
[2] Chaoyue Technology Co., Ltd., Shandong, China

Correspondence: shixiaoqi@st.xatu.edu.cn

Abstract

Semantic segmentation in adverse weather conditions presents significant challenges due to insufficient image brightness, excessive noise, and blurred object boundaries, which hinder the performance of traditional visual recognition methods. Domain generalization (DG) for semantic segmentation aims to leverage data from normal illumination domains to ensure robust model performance in unseen adverse weather domains—a critical requirement for autonomous driving robots. Recent advancements in parameter-efficient fine-tuning via frozen vision foundation models offer new avenues for DGs. However, conventional domain-generalized semantic segmentation methods often struggle with severe weather conditions, particularly in capturing object details and global structures. To overcome these limitations, we introduce RFGLNet, a domain-generalized semantic segmentation model designed for adverse weather scenarios. RFGLNet enhances segmentation accuracy by incorporating an SVD-Initialized Low-Rank Module (SLRM), a Fourier-Enhanced Channel Attention Module (FECA), and a Grouped Modeling Spatial Attention Module (GSAM). By leveraging frequency-domain information through Fourier transforms, RFGLNet improves global structural perception, facilitating a holistic understanding of complex scenarios. Additionally, the decompositional modeling spatial attention mechanism reduces cross-channel interference, enhancing local detail extraction. Using singular value decomposition for parameter fine-tuning ensures precise and rapid alignment with pretrained feature distributions. Our experiments show that RFGLNet achieves a mean intersection over union (mIoU) of 78.3% on the ACDC adverse weather test dataset, with only 4.32 M trainable parameters.

Keywords: semantic segmentation; autonomous driving robots; domain generalization; adverse weather environment; attention mechanism

1. Introduction

Semantic segmentation, a core task in computer vision, assigns semantic labels to each pixel to delineate object contours and positions, enabling advanced scene understanding critical for autonomous driving [6]. For autonomous robots, precise perception of roads, vehicles, and pedestrians under severe weather (nighttime, fog, rain, snow) is indispensable—but such images suffer from reduced brightness, excessive noise, contrast imbalance, and blurred edges [4], severely degrading segmentation performance. As annotated data for adverse weather is prohibitively costly, existing solutions face targeted flaws that this paper addresses with RFGLNet, a domain-generalized semantic segmentation framework tailored for adverse weather.

There are currently two major approaches. The first relies on independent preprocessing and segmentation pipelines (e.g., MPRNet [10], Restormer [11], AllWeather-Net [39], FDA-Net [37]), exhibiting three major shortcomings: First, their weather simulation and enhancement/denoising processes are overly simplified: MPRNet and Restormer excel at general image restoration but fail to model how complex real-world disturbances (e.g., dense rain streaks, dense fog occlusions) disrupt semantic structures. AllWeather-Net integrates a denoising module but does not bind it to the pixel-

level segmentation objective, resulting in insufficient semantic information preservation and limiting its applicability to mild scenarios like light fog or drizzle. FDA-Net focuses on frequency domain distribution alignment without specifically filtering weather noise or optimizing edge blurring. Second, the decoupled workflow of preprocessing and segmentation inevitably causes semantic information loss. This issue exists in UBCN [15] (rain removal + residual network) and LLE-Seg [17] (Retinex enhancement), resulting in irreversible semantic degradation. Third, additional image reconstruction significantly increases computational latency: Lu et al.'s dark-channel defogging [16] and Bi et al.'s style transfer-based synthetic dataset [14] (later employing RefineNet [40]) both suffer from severe computational redundancy, failing to meet autonomous driving's real-time requirements. The second category encompasses traditional domain generalization and adaptation techniques. While lightweight backbone networks like ResNet [3], VGG [7], DenseNet [8], and MobileNet [9] offer advantages in handling complex cross-domain semantic variations, they exhibit limited feature representation capabilities and are susceptible to weather noise contamination, resulting in low segmentation accuracy. Even parameter-efficient approaches like Rein [21] lack specialized optimization for weather noise and boundary ambiguity, while pseudo-label strategies (e.g., UITI [20] achieving only 58.3% mIoU on the GTA5→Cityscapes task) struggle with extreme lighting variations. On the other hand, semi-automated domain adaptation frameworks like those proposed by Gomaa et al. [49][50] heavily rely on target domain data and manual annotation optimization, making them difficult to implement when encountering unseen adverse weather conditions.

To address the core pain point of perception for autonomous driving robots in adverse weather conditions, this paper proposes RFGLNet, a domain-generalized semantic segmentation network model for inclement weather. Our approach involves freezing the backbone parameters of the pre-trained visual foundation model DINOv2 [40] to preserve its robust semantic extraction capabilities, while only fine-tuning low-rank parameters. This significantly reduces the scale of trainable parameters. We designed the Singular Value Decomposition-based Low-Rank Module (SLRM), which first constructs a simulated full-rank token matrix matching the feature dimensions of the backbone network. This matrix undergoes singular value decomposition (SVD) to yield three components: the left singular vector, singular values, and right singular vector transpose. Subsequently, the top k core singular values and their corresponding vector components are selected to initialize the low-rank matrices A and B. Matrix multiplication generates instance template tokens tailored to each layer's features in the backbone network. This module is embedded into each Transformer layer of the backbone network. After outputting raw features at each layer, an attention interaction between the raw features and the instance template Token calculates feature adjustment quantities. These adjustments are then residual-connected to the original features. Throughout this process, the backbone network's pre-trained parameters remain frozen, with only the values of the low-rank matrices A and B being optimized. The final output consists of multi-scale feature maps precisely fine-tuned layer by layer.

Second, we designed the Fourier-Enhanced Channel Attention (FECA) module: It first receives the multi-scale feature maps output by the low-rank modules, which have undergone layer-wise fine-tuning after initialization via singular value decomposition. These feature maps undergo a two-dimensional Fourier transform, mapping spatial domain features to the frequency domain. This separates low-frequency components representing core semantics from high-frequency components representing noise and redundancy. Subsequently, high-frequency components undergo weighted suppression while low-frequency components receive enhancement. An inverse Fourier transform then converts the processed frequency-domain features back into spatial feature maps. Finally, a channel attention mechanism generates channel weights to apply weighted enhancement to each feature map channel, producing a multi-scale feature map with enhanced channels and suppressed noise.

Furthermore, a spatial attention mechanism based on split-branch modeling is proposed: First, it receives fixed-channel-dimension feature maps obtained from the Mask2Former pixel decoder after cross-scale spatial fusion, resolution splitting, and feature projection. These feature maps are then uniformly grouped into channels, with the number of groups adapted to the resolution of the corresponding branch. Each group independently employs resolution-adaptive convolutional kernels to extract spatial information within the group, using small kernels for high resolutions and large kernels for low resolutions. Generate intra-group spatial weight maps to distinguish target details from noise regions. Perform global average pooling on all group weight maps to extract inter-group semantic correlations and generate cross-group weights. Fusion of intra-group spatial weights with cross-group

weights yields global spatial weights. Finally, element-wise multiplication of the global spatial weights with the original branch feature maps completes detail enhancement and noise suppression, outputting refined feature maps for each resolution branch. Through these designed modules, the model achieves cross-domain adaptation for adverse weather scenarios solely via source domain training, without relying on ACDC [27] target domain annotations or synthetic data.

Our main contributions are summarized as follows:

(1) Considering that the random initialization of traditional low-rank modules fails to adapt to the feature distribution of DINOv2 [42] pretrained models, we design an SVD-initialized low-rank module (SLRM), which is incorporated into each transformer layer of DINOv2 to achieve precise alignment and efficient fine-tuning of pretrained features.

(2) Considering that the original SE attention [31] relies solely on spatial-domain statistical information, we design a Fourier-enhanced channel attention module (FECA). Through frequency-domain analysis, this module filters high-frequency noise to enhance the anti-interference capability of global features in severe weather scenarios.

(3) Considering that traditional CBAM attention [30] is susceptible to cross-channel interference, leading to the loss of local details, we design a grouped modeling spatial attention module (GSAM). By leveraging channel grouping and local-to-global information fusion, this module ensures the accurate extraction of local details.

(4) We embed the two designed attention mechanism modules into the feature processing flow of the fine-tuning module and the multiscale feature branches of the decoder Mask2Former [25].

(5) The experimental results show that we achieve a mean intersection over union (mIoU) [46] score of 78.3% on the ACDC [27] test set. To evaluate the model's ability to learn semantic features of training scenarios, we also obtained an mIoU of 83.46% on the Cityscapes [26] validation set.

2. Related work

In this section, we introduce the following three key technologies: first, attention mechanisms, which discuss the application and limitations of SE [31] and CBAM [30]; second, low-rank modules and their initialization strategies, which analyze the challenges faced by existing methods during training; and finally, domain generalization, particularly how parameter fine-tuning methods can overcome challenges in adverse weather environments.

2.1. The Application of Existing Attention Mechanisms in Semantic Segmentation

In recent years, attention mechanisms have been widely applied in semantic segmentation tasks within the field of computer vision, with SE [31] and CBAM [30] being two of the most mainstream approaches.

The SE attention mechanism enhances feature representation capabilities by introducing channel weights to strengthen the model's focus on important channels. It can enhance effective features by learning dependencies between channels. However, its primary limitation lies in relying solely on spatio-temporal statistical information to capture features, thereby neglecting the frequency-domain structural information present in adverse weather images. FcaNet [43] points out that SE's reliance on the lowest-frequency components may cause it to overemphasize noisy regions in adverse weather, leading the model to overlook critical information and weaken feature extraction capabilities. As a representative frequency-domain attention method, FcaNet extracts frequency-domain features through discrete cosine transform (DCT). However, it only performs basic frequency-to-spatial mapping and focuses on cross-domain distribution alignment rather than precise filtering of weather interference.

The CBAM attention mechanism integrates a spatial attention module into SE to further enhance spatial feature representation capabilities. By applying attention-weighted processing to both the channel and spatial dimensions, this mechanism strengthens the model's ability to focus on critical regions. However, it faces similar challenges. Ding et al. [18] noted that CBAM's serial architecture may cause loss of global contextual information, particularly in severe weather scenarios, weakening the model's ability to extract features from details and boundaries. Furthermore, when processing

extreme weather images, more complex lighting variations or noise density may be overlooked, leading to insufficient adaptability and reduced robustness under such conditions.

To overcome these limitations, this paper proposes the FECA and GSAM modules: FECA integrates Fourier transform with channel attention mechanisms. On one hand, it employs Fourier transform (rather than discrete cosine transform) to more precisely separate low-frequency semantic information from high-frequency noise. On the other hand, it deeply integrates frequency-domain decomposition with channel attention, dynamically enhancing effective semantic channels and suppressing noisy channels through attention weights, rather than merely extracting frequency-domain features. This design addresses FcaNet's inadequate adaptation to adverse weather scenarios while overcoming the redistribution alignment and weak interference filtering issues of methods like FDA. GSAM employs a grouped spatial attention mechanism to reduce cross-channel interference, compensating for boundary detail loss caused by CBAM's sequence structure while achieving precise optimization of blurred image boundaries under harsh weather conditions.

2.2. Low-rank Modules and Initialization Strategies

Low-rank modules enhance the parameter efficiency of visual models by introducing low-rank matrices and learnable tokens, which are typically employed for domain generalization tasks. Rein [21] significantly reduced the number of training parameters through low-rank design, enabling models to maintain high performance while decreasing computational demands. However, the low-rank design also faces initialization and alignment challenges. Original low-rank modules typically employ random initializations, which may result in slow convergence during early training stages, particularly in cross-domain tasks where pretrained feature distributions cannot be precisely aligned. To address this, singular value decomposition (SVD) initialization has gained widespread adoption as an improved strategy. By extracting core patterns from pretrained features, SVD effectively accelerates training and enhances low-rank module performance, notably improving model adaptability in adverse weather scenarios.

Sun et al [36]. proposed the SVFit method, which decomposes pre-trained weight matrices via singular value decomposition (SVD). It extracts the first r core singular values and vectors to initialize low-rank matrices, demonstrating excellent pre-training feature retention and distribution alignment capabilities during parameter-efficient fine-tuning. However, this approach fails to address the cross-domain characteristics and deployment requirements of adverse weather semantic segmentation, exhibiting two core limitations that SLRM specifically overcomes: First, SVFit freezes the singular vector matrix obtained from SVD and adapts to tasks solely by scaling singular values. This locks the pre-trained core subspace, preventing capture of task-specific semantic structures like rain/fog occlusion and blurred contours. Furthermore, experiments lack coverage of complex cross-domain scenarios, limiting applicability. SLRM, however, performs SVD using simulated token matrices whose feature dimensions match the target model. It decomposes and constructs fully trainable low-rank matrices A and B, enabling precise alignment with adverse weather semantic features through subsequent training. This approach eliminates reliance on pre-trained weight distributions, addressing the deficiency in scene adaptability. Second, SVFit directly modifies the singular values of pre-trained weights. Switching between tasks requires repeatedly unloading, loading singular values, and reconstructing the full weight matrix, generating significant redundant computations and resulting in inefficient deployment. SLRM's A and B modules exist independently of pre-trained weights, stored separately after training. During inference, tasks are adapted solely through the product of pre-trained weights and A/B modules. Switching tasks requires no weight reconstruction—merely replacing the corresponding A/B modules eliminates redundant computations entirely, enabling highly efficient deployment.

2.3. Application of Domain Generalization in Adverse Weather Tasks

In recent years, domain generalization methods have been widely applied in adverse scenarios, aiming to maintain high performance in unseen severe weather target domains through training on source domain data. Images of severe weather often exhibit significant illumination variations and noise

interference, making the enhancement of model robustness in such scenarios a key research focus. To address this challenge, numerous parameter-efficient fine-tuning techniques based on visual foundation models [12], such as the Rein [21] framework, have been proposed in recent years. By fine-tuning large-scale visual foundation models and employing trainable learning sequences alongside low-rank parameter design, this approach substantially reduces the number of parameters requiring adjustment during training, thereby enhancing cross-domain generalization capabilities. Through this methodology, models achieve effective knowledge transfer between the source and target domains, demonstrating particularly robust performance in adverse weather scenario tasks. Additionally, pseudo label optimization and feature enhancement are key strategies for improving the cross-domain generalization ability. The UITI [20] framework proposed by An et al. achieves an mIoU of 58.3% on the GTA5→Cityscapes source-free domain adaptation task via image style-guided pseudo label selection and feature-level enhancement, fully validating the effectiveness of this combined strategy.

Notably, while domain adaptation has shown progress in adverse weather target detection, its inherent limitations make it less suitable for semantic segmentation tasks—especially in scenarios lacking target domain data. For example, Gomaa et al [49][50].'s semi-automatic domain adaptation frameworks rely heavily on target domain motion information and require manual refinement of clustered annotations, failing to achieve true source-free adaptation. Moreover, these methods prioritize object-level feature alignment for detection, ignoring the pixel-level boundary preservation and local feature integrity critical for semantic segmentation. Worse, their dependence on target domain data makes them impractical when severe weather target domains are unseen or unlabeled—an inherent flaw that highlights the superiority of domain generalization's source-only training paradigm for real-world adverse weather tasks.

The efficacy of domain generalization approaches in adverse weather tasks has garnered extensive attention. Researchers such as Wei [21] proposed that parameter-efficient fine-tuning via visual foundation models can effectively increase the segmentation accuracy across varying environmental conditions. Furthermore, Yang [13] et al. highlighted that cross-domain training enables models to better adapt to target domain environmental variations on the basis of source domain data. Although Rein [21] demonstrated commendable cross-domain adaptability, existing research continues to explore how to further refine model design [5] to achieve superior performance in adverse weather scenarios. For example, Yang [38] proposed multimodal learning that integrates spatial and temporal domain information, enabling more comprehensive capture of image structural features and thereby enhancing model robustness under severe weather conditions.

3. Methodology

In this section, we first present a detailed overview of the overall network architecture of RFGLNet, followed by a description of three core innovative modules designed for semantic segmentation tasks in adverse weather conditions: the SVD-initialized low-rank module (SLRM), Fourier-enhanced channel attention module (FECA), and grouped modeling spatial attention module (GSAM).

3.1. Network architecture

The core design philosophy of RFGLNet addresses the key limitations of the DINOv2 pretrained model in semantic segmentation tasks, including insufficient feature adaptability, weak anti-interference capability in complex scenarios, and limited accuracy in local detail extraction. To mitigate these issues, we propose three key components—the SLRM, FECA, and GSAM—and establish a comprehensive technical pipeline encompassing layerwise feature fine-tuning of the backbone network, frequency-domain feature denoising and enhancement, and detail refinement in multiresolution branches of the decoder. This integrated design aims to increase the model's segmentation accuracy and domain generalization capability in adverse weather scenarios. The overall architecture of the model is shown in Figure 1.

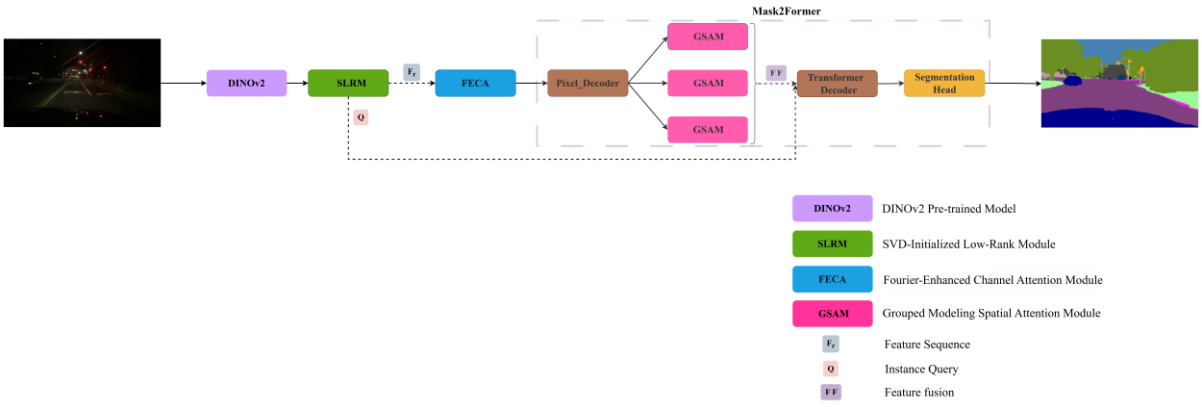


Figure 1. Overall architecture of our proposed network. Integrate three core modules (SLRM for parameter-efficient fine-tuning, FECA for frequency-domain denoising, GSAM for spatial detail refinement) into a unified pipeline, enabling domain generalization for adverse weather semantic segmentation without target domain data or independent preprocessing.

To fully leverage DINOv2's powerful pre-trained visual representation capabilities, the entire network takes batches of raw image tensors as input. It first passes through a parameter-frozen DINOv2 backbone network to output raw features from each Transformer layer. These feature sequences are then optimized via low-rank transformation using SVD-initialized SLRM (with minimal fine-tuning), generating instance query vectors. These are subsequently converted into multi-scale feature maps. These feature maps are fed into FECA for frequency-domain denoising and channel enhancement, yielding purified multi-scale feature maps. Mask2Former then receives these feature maps. A pixel decoder first fuses multi-scale information to generate mask features and memory features. GSAM subsequently performs spatial grouping refinement and detail augmentation on the fused features. Initial query features are constructed by combining these with instance query vectors generated layer-by-layer by SLRM. A Transformer decoder iteratively updates the query features, and a prediction head outputs category prediction tensors and mask prediction tensors. Finally, through category loss, mask loss, and Dice loss calculations, only the trainable parameters of the SLRM, FECA, GSAM, and the pixel decoder, decoder, and prediction head of Mask2Former are updated via backpropagation. Post-processing yields the final image segmentation results. Through this series of designs, the network enhances its generalization capabilities by fully leveraging pre-training while employing progressive processing—including layer-wise fine-tuning, feature denoising, and feature refinement.

3.2. SVD-Initialized Low-Rank Module

The SVD-initialized low-rank module (SLRM) aligns with pretrained feature distributions, reduces parameter counts, and simultaneously generates instance tokens and queries adapted to adverse weather conditions. Its overall architecture is illustrated in Figure 2.

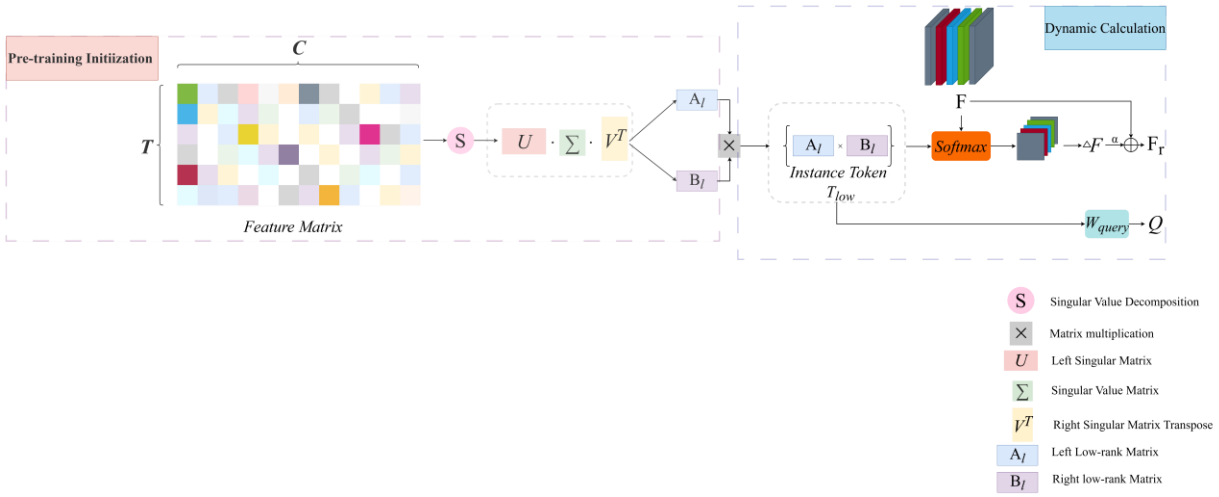


Figure 2. The architecture of our proposed SLRM. Embedding SLRM into DINOv2's Transformer layer enables efficient parameter fine-tuning by generating virtual tokens and performing SVD low-rank decomposition, optimizing only the low-rank matrix while dynamically computing feature correction quantities.

The SLRM workflow comprises two phases: pretraining initialization and dynamic computation during training. Prior to training commencement, it first simulates the generation of a full-rank token to create a virtual token T_{full} with the same dimension $(T \times C)$, where T=100 is the number of tokens, C=1024 is the number of channels, and $T_{full} \in \mathbb{R}^{T \times C}$ (where $\mathbb{R}$ denotes the set of real numbers) is initialized with a uniform distribution $U(-0.01, 0.01)$. Second, singular value decomposition is performed on T_{full}, decomposing it into three components: semantic patterns, importance weights, and channel correlations. T_{full} is computed as follows:

$$T_{full} = U \cdot \Sigma \cdot V^T \tag{1}$$

where $U \in \mathbb{R}^{T \times T}$ is the left singular vector, each column of which corresponds to a semantic pattern (e.g., semantic features of street lamps, vehicles), and where $\Sigma \in \mathbb{R}^{T \times C}$ is the singular value matrix, which is essentially a diagonal matrix (with only diagonal elements being nonzero). The diagonal elements are singular values, which measure the importance of the corresponding semantic patterns—i.e., the larger the singular value is, the stronger the contribution of the semantic pattern to the original token T_{full}; $V^T \in \mathbb{R}^{C \times C}$ is the right singular vector, and each element corresponds to the correlation strength between a specific semantic and a specific feature channel. For example, if the element value of the brightness channel in the right singular vector corresponding to road semantics is very large, the brightness channel is crucial for expressing road semantics. Second, the core singular values are selected according to the set singular value retention ratio γ, and the number of values is k. The formula for calculating k is as follows:

$$k = \lfloor R \cdot \gamma \rfloor \tag{2}$$

where $\gamma \in [0.1, 1.0]$ and where R is the low-rank dimension. We then extract the corresponding submatrices: $U_k = U[:, :k] \in \mathbb{R}^{T \times k}$, $\Sigma_k = \Sigma[:k, :k] \in \mathbb{R}^{k \times k}$, and $V_k^T = V^T[:k, :] \in \mathbb{R}^{k \times C}$. By filtering out low singular values, redundant semantics and other interferences are removed. The selected submatrices are subsequently expanded to the target dimension. The initial value of the left low-rank matrix is denoted as A_l^{init}, and its formula is as follows:

$$A_l^{init} = U_k \cdot \sqrt{\Sigma_k} \oplus 0_{T\times(R-k)} \tag{3}$$

Here, l denotes the initial value of the l-th left low-rank matrix, $\oplus$ represents horizontal matrix concatenation, and 0 denotes a zero matrix (with dimensions of T rows and $(R-k)$ columns here). The initial value of the right low-rank matrix is denoted as B_l^{init}, and its formula is as follows:

$$B_l^{init} = \sqrt{\Sigma_k} \cdot V_k^T \oplus 0_{(R-k)\times C} \tag{4}$$

where $\oplus$ represents vertical matrix concatenation, and the zero matrix here has dimensions of $(R-k)$ rows and C columns. Finally, A_l^{init} and B_l^{init} are assigned to $\{A_l, B_l\}$, which stores the low-rank matrices of each layer in the module. At this point, the SVD initialization of the low-rank matrices before training is completed.

In the dynamic computation phase during training, for the current transformer layer l, we use the trained and updated left low-rank matrix A_l (with dimensions $T\times R$) and right low-rank matrix B_l (with dimensions $R\times C$) to generate the instance token $T_{low} \in \mathbb{R}^{T\times C}$ via matrix multiplication, achieving semantic retention of the full-rank token while compressing the number of parameters. T_{low} is computed as follows:

$$T_{low} = A_l \cdot B_l \tag{5}$$

This token inherits the core semantics of the full-rank token, and its parameter count is only R/C of that of the full-rank structure. Next, we compute the feature correction term: a) Calculate the semantic similarity between the input feature sequence F and the low-rank token T_{low} and introduce the temperature coefficient $\sqrt{C}$ to balance the gradient scale. The formula for calculating the similarity matrix S is as follows:

$$S = Softmax\left(\frac{F \cdot T_{low}^T}{\sqrt{C}}\right) \tag{6}$$

where $S \in \mathbb{R}^{B\times L\times T}$ is the similarity matrix. b) Exclude the first background Token and retain the similarity submatrix $S' = S[:,:,1:T]$ corresponding to the remaining 99 instance Tokens. c) Optimize the semantic representation of the instance Token via a linear transformation $W_{token} \in \mathbb{R}^{C\times C}$ and a bias $b_{token} \in \mathbb{R}^{C}$ and then generate the feature correction term ΔF via combination with the similarity matrix. ΔF is computed as follows:

$$\Delta F = S' \cdot \left(T_{low}[1:T] \cdot W_{token} + b_{token}\right) \tag{7}$$

Next, a learnable scaling factor α is introduced, and combined with residual connections, the enhanced feature sequence F_r is obtained. The formula for calculating the feature sequence F_r is as follows:

$$F_r = F + \alpha \cdot \Delta F \tag{8}$$

Finally, the low-rank token T_{low} is converted into an instance query Q via the linear layer W_{query}. Q is computed as follows:

$$Q = T_{low} \cdot W_{query} \tag{9}$$

The final output channel-enhanced feature sequence F_r and instance query Q are used in the Fourier-enhanced channel attention module (FECA) and transformer decoder, respectively.

3.3. Fourier-Enhanced Channel Attention Module

The core reason why frequency domain processing is suitable for semantic segmentation in adverse weather conditions lies in its ability to precisely decouple semantic content from weather interference. SET [48] employs the Fast Fourier Transform to separate low-frequency semantic phase components from high-frequency weather amplitude components, demonstrating targeted interference suppression efficacy on the ACDC dataset. The amplitude hint aggregation mechanism proposed by Li et al [35]. also provides insights for frequency-domain interference mitigation. Building upon this foundation, this paper further optimizes the frequency-domain processing scheme by designing the Fourier-enhanced channel attention module (FECA). Our overall FECA architecture is illustrated in Figure 3.

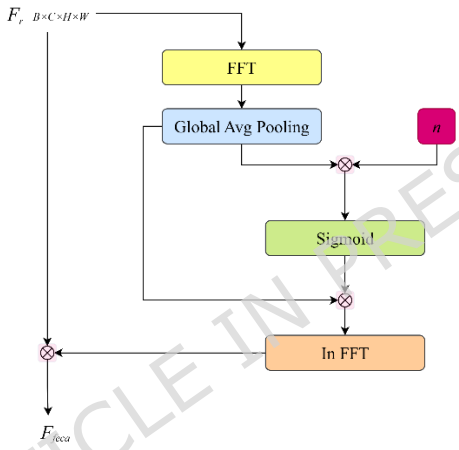


Figure 3. The architecture of our FECA. The FECA integrates Fourier frequency domain processing with channel attention to separate low-frequency semantic components from high-frequency noise components in features. It achieves noise reduction by suppressing high frequencies and enhancing low frequencies.

First, the input feature F_r is dimensionally transformed into multiscale feature maps, followed by the application of a 2D fast Fourier transform (2D FFT) to map spatial-domain features to the frequency domain, thereby yielding the frequency-domain feature F_{freq}. Since the energy distributions of semantic information (low-frequency, high-energy) and noise interference (high-frequency, abnormal energy) in adverse weather images differ significantly in the frequency domain and the frequency-domain result of the Fourier transform is complex—the magnitude spectrum can directly quantify the energy intensity of each frequency component and address the issue of insufficient noise resistance in traditional spatial-domain attention caused by the mixture of semantics and noise—it is necessary to extract the frequency-domain magnitude spectrum $\left|F_{freq}\right|$. The formula for calculating the frequency-domain magnitude spectrum $\left|F_{freq}\right|$ is as follows:

$$\left|F_{freq}\right| = \sqrt{\mathrm{Re}(F_{freq})^2 + \mathrm{Im}(F_{freq})^2} \tag{10}$$

where $\mathrm{Re}(\cdot)$ denotes the real part of a complex number and where $\mathrm{Im}(\cdot)$ denotes the imaginary part. Second, to generate frequency domain weights, a) first, global average pooling (GAP) is performed

on the magnitude spectrum along the spatial dimension to compress it into a channel-level frequency domain feature F_{freq_ch}. The formula for calculating F_{freq_ch} is as follows:

$$F_{freq\_ch} = GAP(|F_{freq}|) \tag{11}$$

Then, a learnable vector n initialized on the basis of the statistical distribution of adverse weather noise is introduced, and a noise suppression gate g is generated via the $Sigmoid$ function. The formula for calculating g is as follows:

$$g = Sigmoid(F_{freq\_ch} \cdot n) \tag{12}$$

Finally, F_{freq_ch} and g are multiplied to generate the frequency-domain attention weight α_{freq}. Subsequently, an inverse Fourier transform is performed on α_{freq} to obtain the spatial-domain channel attention weight $\alpha_{spatial}$. Eventually, $\alpha_{spatial}$ and the input feature F_r undergo elementwise multiplication, and the finally enhanced feature F_{feca} is obtained.

3.4. Grouped Modeling Spatial Attention Module

It is worth noting that standard spatial attention mechanisms like CBAM inherently possess two flaws when learning spatial weight maps through uniform processing of all feature channels, flaws that are further amplified in adverse weather scenarios. First is unavoidable cross-channel interference: in adverse conditions like fog or heavy rain, feature channels become unevenly contaminated. Certain channels carry critical semantic information such as road edges and pedestrian outlines, while others are dominated by weather noise (e.g., fog particles, raindrops). Standard spatial attention blends all channels during weight learning, allowing irrelevant noise channels to interfere with the spatial correlation modeling of semantic channels. This ultimately leads to blurred boundary detection. Second, insufficient sensitivity to local details: Standard spatial attention prioritizes global spatial patterns, but adverse weather often fragments local semantic features. Uniform all-channel processing fails to focus on subtle local variations. In contrast, GSAM's channel grouping mechanism addresses these issues through three core principles: 1) Isolating semantic and noise channels to minimize interference; 2) Implementing group-specific fine-grained attention to capture blurred local features; 3) Employing group convolutions to balance parameter efficiency and real-time performance—critical for autonomous driving scenarios. Our GSAM is shown in Figure 4.

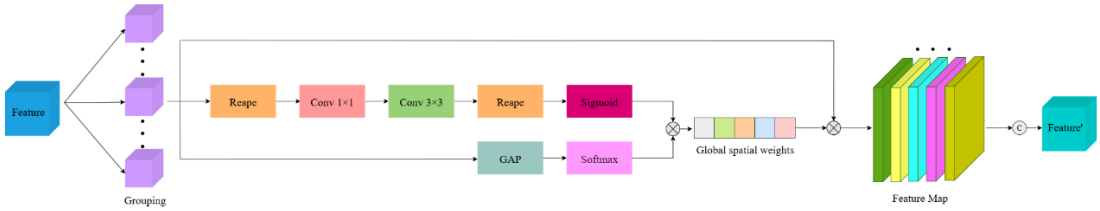


Figure 4. Overall architecture of our GSAM. The GSAM employs a channel grouping strategy to generate global spatial weights through lightweight convolution, global average pooling (GAP), and Softmax. By extracting intra-group spatial weights and fusing inter-group weights, it focuses on general semantic features for further refinement.

First, the input feature sequence F_{feca} is evenly divided into several groups G along the channel dimension to obtain the feature of each group F_g. The formula for calculating F_g is as follows:

$$F_g = Spilt\left(F_{feca}, \dim = 2, num\_split = G\right) \tag{13}$$

This grouping design reduces the number of channels per group from C to C/G, enabling each group to focus on processing specific local semantics. Second, spatial attention weights are independently generated for each group feature F_g: a) Reshape the intragroup features from the sequence dimension to the spatial dimension to obtain $F_{g,spatial}$; b) compress the channel dimension and extract spatial details via 1×1 convolution and 3×3 convolution to obtain $F_{g,conv}$; and c) reshape $F_{g,conv}$ back to the sequence dimension and generate intragroup spatial weights β_g via *Sigmoid* normalization. The formula for calculating β_g is as follows:

$$\beta_g = Sigmoid\left(\mathrm{Re}shape\left(F_{g,conv}, shape = (B, L, 1)\right)\right) \tag{14}$$

where B is the batch size. The closer β_g is to 1, the richer the detail features at that position are, which needs to be enhanced; if it is close to 0, it indicates that the semantics at that position are blurred, possibly in a noise region, and need to be suppressed. Next, to avoid the problem of information fragmentation that may be caused by grouping, it is necessary to fuse the semantic information of all groups and generate global cross-group weights λ. a) Perform global average pooling (GAP) on each group feature F_g along the spatial dimension to obtain intragroup global semantics $F_{g,global}$; b) map $F_{g,global}$ to the group weight dimension via linear transformation and then obtain cross-group weights λ_g via *Softmax* normalization. λ_g is computed as follows:

$$\lambda_g = Softmax(W_{cross} \cdot GAP(F_g)) \tag{15}$$

where W_{cross} is a learnable cross-group transformation matrix. Finally, the intragroup spatial weights and cross-group weights are multiplied to obtain the global spatial weights β_g^{fusion}; then, elementwise multiplication is performed between this weight and the original group feature F_g to complete intragroup spatial enhancement. Eventually, the enhanced features of all the groups are concatenated back to the original channel dimension to obtain the enhanced feature sequence F_{gsam}. F_{gsam} is computed as follows:

$$F_{gsam} = Concat(F_g \otimes \beta_g^{fusion} | g = 1, 2, ..., G) \tag{16}$$

where $\otimes$ denotes elementwise multiplication and where *Concat* denotes concatenation along the channel dimension.

4. Experiments and Results

In this section, we employed the widely adopted benchmark datasets Cityscapes [26] and ACDC [27]. Our experiments focus on domain generalization semantic segmentation tasks under adverse weather conditions. Consequently, model training is conducted on the Cityscapes dataset, whereas the validation phase employs both the Cityscapes dataset and the ACDC dataset. The Cityscapes validation set quantifies the model's learning efficacy regarding semantic features in training scenarios, whereas the ACDC validation set provides an initial assessment of its generalizability. The final evaluation is conducted on the ACDC dataset. To evaluate our proposed network model, we first introduce the relevant datasets, followed by a presentation of the experimental details. We subsequently outline the metrics employed to assess our approach. Thereafter, we conduct an ablation study on the proposed

modules. Finally, we compare our results on both datasets with those of other state-of-the-art semantic segmentation networks.

4.1. Datasets

4.1.1. Cityscapes

Cityscapes is a semantic segmentation dataset specifically designed for urban street scenes, featuring pixel-level annotations for daytime road environments. The entire dataset comprises street views from 50 distinct cities, containing 5,000 images at a resolution of 2048x1024. Of these, 2,975 images are designated for training, 500 for validation, and 1,525 for testing purposes. The Cityscapes dataset typically employs 19 commonly used categories for evaluating segmentation accuracy. In our experiments, this dataset serves as both the training dataset and the validation dataset. An example image from the Cityscapes dataset is shown in Figure 5.

Figure 5. Sample images and labels of the Cityscapes dataset.

4.1.2. ACDC

The ACDC dataset is a semantic segmentation dataset for autonomous driving and computer vision research under adverse conditions, comprising 4,006 driving scene images captured during four challenging weather scenarios: rain, snow, fog, and nighttime. Of these, 1,600 images serve for training, 400 for validation, and 2,006 for testing purposes, all at resolutions of 1920x1080. Each image is accompanied by high-quality, pixel-level semantic annotations, adhering directly to the Cityscapes annotation framework and encompassing 19 semantic categories. In our experiments, this dataset serves as both the validation and test datasets. An example image from the ACDC dataset is shown in Figure 6.

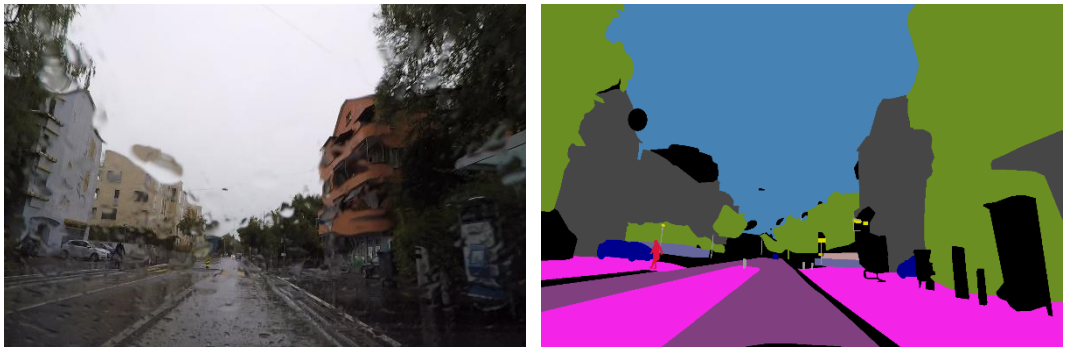


Figure 6. Sample images and labels of the ACDC dataset.

4.2. Training Details

All training and validation experiments are conducted via PyTorch 2.0.1 combined with MMSegmentation 1.2.0 to establish an end-to-end training workflow. These were implemented on the PyCharm development platform, accelerated with CUDA 11.7 alongside CuDNN 8.5.0, and executed on a single NVIDIA RTX 3080 Ti GPU.

We employ the AdamW [44] optimizer to train our network model on the Cityscapes dataset. The optimizer featured a momentum of (0.9, 0.999), a batch size of 2, and a weight decay of 0.05. Additionally, we employed a linear learning rate scheduler to warm-start the learning rate during the initial training phase.

During training, similar to [24][45], we employ the cosineAnnealingParamScheduler with a minimum learning rate of 1e--5. The learning rate decay formula for this scheduler is as follows:

$$lr_{\mathrm{t}} = lr_{\min} + 0.5 \times \left(lr_{\mathrm{init}} - lr_{\min} \right) \times \left(1 + \cos\left(\pi \times \frac{\mathrm{t}}{T} \right) \right) \tag{17}$$

where lr_{t} is the learning rate at step t, $lr_{\min}$ is the minimum learning rate corresponding to eta_min in the configuration, and lr_{init} is the initial learning rate. Assuming that the effective iteration range of the scheduler is $[begin, end]$, the total number of iterations is $T = end - begin$.

4.2.1. Loss Function

We specifically employ a multi-task weighted loss function. The total loss is the weighted sum of classification loss (L_{cls}), mask loss (L_{mask}), and Dice loss (L_{dice}), with the formula as follows:

$$L_{\mathrm{total}} = L_{\mathrm{cls}} \times 2.0 + L_{\mathrm{mask}} \times 5.0 + L_{\mathrm{dice}} \times 5.0 \tag{18}$$

L_{cls} employs cross-entropy loss without using the Sigmoid activation function, calculating the loss using the mean method. To suppress the dominant interference of fog particles and sky in adverse weather images, we applied differentiated weighting to the categories: all 19 target semantic categories were assigned a weight of 1.0, while the background category's weight was adjusted to 0.1,with a loss weight of 2.0, L_{cls} focuses on optimizing the accuracy of pixel-level category prediction, ensuring the model can correctly distinguish between core target categories and background noise across domains.

L_{mask} adopts cross-entropy loss combined with a sigmoid activation function to predict binary masks, calculating the loss using the mean, with weights set to 5.0. This loss is designed to address the blurred target boundaries caused by adverse weather. By applying Sigmoid activation for binary mask prediction, it directly optimizes the spatial alignment between the predicted mask and the ground truth, prioritizing the refinement of boundary details critical for semantic segmentation.

L_{dice} employs the basic Dice loss, incorporating a Sigmoid activation function to enhance feature discrimination. By introducing a numerical stability coefficient: ε, it prevents zero denominators during computation. The loss weight is also set to 5.0. This loss is highly robust to class imbalance, which is particularly severe for small targets in adverse weather scenarios.

4.2.2. Data Augmentation Strategy

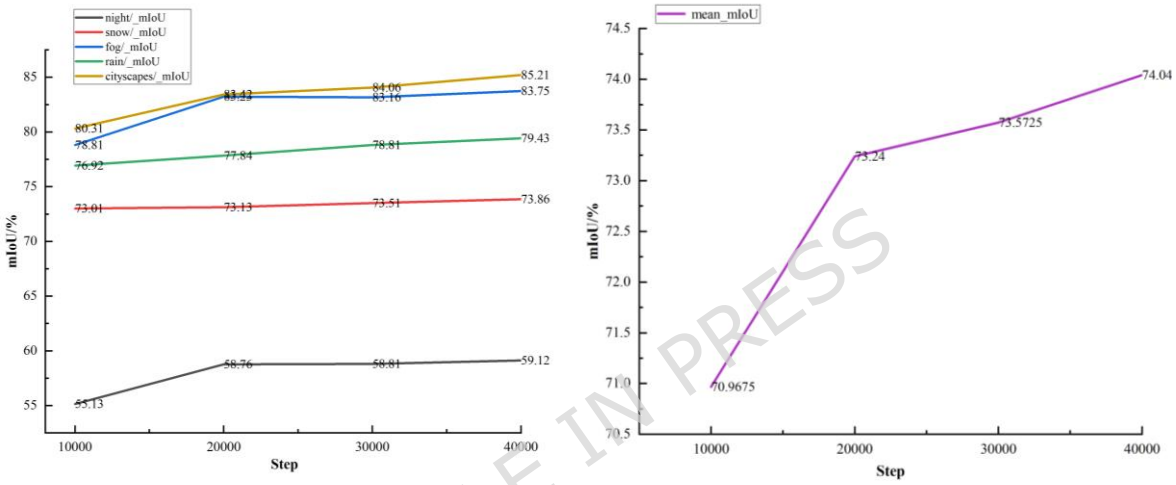
We employ the data augmentation strategies encapsulated in MMSegmentation to enhance the model's domain generalization capability in adverse weather scenarios. Specifically, this includes: randomly scaling the shortest side of images to a range of (512–2048) pixels to accommodate scale differences across targets at varying distances; randomly cropping to a (512×512) pixel specification while imposing a constraint that no single class exceeds 75% coverage to avoid extreme background samples; Applying horizontal flipping with a 50% probability to mitigate overfitting risks related to target orientation; and applying photometric distortions by randomly adjusting brightness to (0.5–1.5)

times, contrast to (0.5–1.5) times, saturation to (0.5–1.5) times, and hue to (-0.1–0.1) times. This simulates lighting fluctuations and color distortions in scenarios like rain, snow, fog, and nighttime, guiding the model to learn core semantic features independent of environmental conditions.

4.2.3. Convergence Behavior

The model convergence is validated on the validation sets of Cityscapes (source domain) and ACDC (target domain), with the mIoU curves shown in Figure 7. As shown in Figure 7(a), the mIoU for each scene steadily increases with training steps and stabilizes after 30,000 steps: Cityscapes achieves a stable mIoU of 85.21%, benefiting from consistent training and validation data distributions; Rain and Fog scenes reach 79.43% and 83.75% respectively, validating the effectiveness of FECA frequency-domain denoising; while Rain and Night scenes achieved 73.86% and 59.12%, respectively.

Figure 7(b) displays the average mIoU evolution across the four weather scenes: rising from an initial 70.9675% to 74.04% at 40,000 steps. The model exhibits no overfitting.



(a) (b)

Figure 7. Validation set mIoU convergence curves of RFGLNet. (a) presents the mIoU variation trends of five individual scenarios, including Cityscapes (source domain) and four adverse weather scenarios (Night, Snow, Fog, Rain); (b) shows the average mIoU trend of the four adverse weather scenarios.

All curves are derived from the combined validation set of Cityscapes and ACDC datasets, reflecting the model's fitting ability on the source domain and generalization ability on adverse weather target domains during training.

4.3. Evaluation Metrics

The validation phase of our experiment is conducted on the Cityscapes and ACDC datasets, and the final testing phase is completed on the ACDC dataset. Here, no data augmentation strategies are adopted; instead, only image loading and fixed-size resizing are performed. Both our validation evaluator and test evaluator take the mean intersection over union (mIoU) [46] as the core fine-grained scene metric to evaluate the accuracy and generalization ability of the proposed network model.

For the calculation of the mIoU, first, the IoU of each category is computed, and then the average of the IoUs of all categories is taken. The formula for the IoU is as follows:

$$IoU = \frac{TP}{TP + FP + FN} \tag{19}$$

where TP denotes a true positive, FP denotes a false positive, and FN denotes a false negative. The formula for the mIoU is as follows:

$$mIoU = \frac{1}{k}\sum_{i=1}^{k} IoU_i \tag{20}$$

where k denotes the total number of categories and where IoU_i denotes the IoU value of the i-th category.

4.4. Ablation Study

To demonstrate the effectiveness of the components in our RFGLNet, we utilize DINOv2 as the backbone network and conduct a series of ablation experiments on the ACDC dataset to evaluate the proposed SLRM, FECA, and GSAM.

Table 1. Ablation experiment results on the ACDC dataset. SLRM: SVD-initialized low-rank module, FECA: Fourier-enhanced channel attention module, GSAM: grouped modeling spatial attention module.

Models	SLRM	FECA	GSAM	mIoU(%)	Trainable Parameters(M)
DINOv2+Mask2Former				67.3	304.2
+SLRM	√			75.2	2.882
+SLRM+FECA	√	√		76.1	4.212
+SLRM+GSAM	√		√	76.6	2.99
RFGLNet	√	√	√	78.3	4.32

Table 2. Comparisons of the SLRM with other PEFT modules. LoRA: low-rank adaptation; Prefix-T: prefix tuning.

PEFT Module	ACDC
	mIoU(%)
LoRA [33]	77.1
Prefix-T [34]	76.5
SLRM	78.3

Table 3. Comparisons of FECA with other attention modules. SE-Block: squeeze-and-excitation block; ECA: efficient channel attention.

Attention Module	ACDC
	mIoU(%)
SE-Block [31]	76.6
ECA [32]	76.1
FECA	78.3

Table 4. Comparisons of the GSAM with other attention modules. CBAM: Convolutional block attention module, MS-Attention: Multiscale attention.

Attention Module	ACDC
	mIoU(%)
CBAM [30]	77.2
MS-Attention [51]	76.3
GSAM	78.3

SLRM: The SLRM focuses on filtering redundant features of the source domain through singular value decomposition and performing parameter-efficient fine-tuning on core semantic features, thereby providing streamlined cross-domain features for subsequent modules. To verify the effectiveness of the SLRM, relevant tests were conducted. As shown in Table 1, when only the SLRM is added, the mIoU value increases from 67.3% of the baseline model value to 75.2%, and the number of trainable parameters is reduced to 2.882 M.

Furthermore, we compared SLRM with other parameter fine-tuning methods (LoRA and Adapter) while keeping other parts of RFGLNet unchanged. LoRA [31], proposed by Hu et al., inserts low-rank matrix pairs into key layers of the pretrained model and fine-tunes only a small number of low-rank parameters, significantly reducing the number of training parameters. Prefix-T [32], proposed by Li et al., guides the pretrained model to adapt to downstream tasks by inserting trainable prefix parameters before the input sequence while freezing the main parameters of the model to reduce training costs. The

experimental results are presented in Table 2, where our proposed SLRM achieves the best performance on the ACDC test set.

FECA: FECA focuses on calibrating channel features to adapt to variations caused by adverse weather conditions. As demonstrated in Table 1, when FECA is integrated with SLRM, the mIoU value increases by 76.6% relative to the baseline, with trainable parameters reaching 4.212 million.

To demonstrate the superiority of FECA, we compared it with alternative attention modules (ECA and SE-Block). Specifically, SE-Block [31], proposed by Hu et al., enhances key feature representations through global average pooling and channel weight learning. ECA [30], proposed by Wang et al., replaces the SE-Block's fully connected layer with 1D convolutions, enabling cross-channel interactions by avoiding dimensionality reduction. The experimental results are presented in Table 3, where FECA outperforms the other two attention modules on the ACDC test set. Traditional spatial denoising methods have demonstrated excellent performance in adverse weather detection—for example, the median filter (MF) proposed by Alzanin [2] achieves an accuracy of 98.83% in noise suppression tasks. Drawing on its core logic of "first purification followed by enhancement," the FECA module in our study separates the high-frequency details and low-frequency structures of features via Fourier transform, dynamically calibrating channel features in both the frequency and spatial domains to enhance the model's performance.

GSAM: The GSAM aims to optimize the capture of spatial relationships among multiscale objects. It is evident from Table 1 that after finally adding the GSAM, the mIoU value of the full model reaches 78.3%, with the number of trainable parameters being 4.32 M.

We compared the GSAM with other modules (CBAM and MS attention). Briefly, CBAM [30], proposed by Woo et al., fuses dual branches of channel attention and spatial attention to achieve dual refinement of input features, enhancing the model's representation capability. MS-Attention [51], proposed by Li et al., independently calculates attention weights through multiscale branches and dynamically fuses features of different resolutions, improving the model's accuracy and adaptability. The experimental results are shown in Table 4, where our GSAM outperforms the two attention mechanisms in terms of the mIoU. This module groups features at the semantic scale and relies on mechanisms of refined intragroup spatial correlation and dynamic fusion of intergroup features. This process not only effectively suppresses cross-scale feature redundancy and highlights key features of multiscale semantics but also improves computational efficiency and model performance.

4.5. Performance evaluation

4.5.1. Performance evaluation of ACDC

In this subsection, we present a comparative analysis of various domain generalization frameworks, including methods such as HRNet [47], DAFormer [24], Refign [19], HRDA [29], RefineNet [40], Rein [21], MGCDA [28], and HGformer [18]. Table 5 summarizes the performance of each method on the basis of their intersection over union (IoU) evaluation results across different categories. The best result is highlighted in bold.

Table 5 Per-class IoU(%) performance on the ACDC dataset.

Methods / Category	IoU/%								
	HRNet	DAFormer	Refign	HRDA	RefineNet	Rein	MGCDA	HGFormer	Ours
road	55.7	58.4	89.5	88.3	66.3	**94.5**	73.4	85.4	**94.5**
sidewalk	10.9	51.3	63.4	57.9	28.9	78.2	28.7	67.9	**81.1**
building	55.4	84.0	84.3	88.1	67.6	**92.0**	69.9	74.9	91.9
wall	7.7	42.7	43.6	55.2	19.2	61.9	19.3	47.6	**66.3**
fence	15.9	35.1	34.3	36.7	25.9	55.0	26.3	**57.1**	57
pole	21.7	50.7	52.3	56.3	36.7	**64.8**	36.8	63.0	62.2
traffic light	37.8	30.0	63.2	62.9	50.0	73.7	53.0	68.6	**75.9**
traffic sign	42.5	57.0	61.4	65.3	47.5	72.7	53.3	**82.5**	76.9
vegetation	67.4	74.8	86.9	74.2	69.4	88.3	75.4	85.2	**88.7**
terrain	13.3	52.8	58.5	57.7	28.8	67.4	32.0	49.3	**71.9**

Methods / Category	IoU/%								
	HRNet	DAFormer	Refign	HRDA	RefineNet	Rein	MGCDA	HGFormer	Ours
sky	59.0	51.3	**95.7**	85.9	83.0	95.4	84.6	90.6	95.2
person	38.7	58.3	63.1	68.8	42.1	77.1	51.0	62.8	**79.9**
rider	14.0	32.6	39.3	45.7	17.7	**60.2**	26.1	59.2	59.8
car	68.4	82.7	84.1	88.5	72.6	92.6	77.6	80.9	**93.9**
truck	23.8	58.3	65.7	76.4	30.9	**84.1**	43.2	46.4	82.7
bus	48.0	54.9	71.3	82.4	31.6	**86.9**	45.9	50.1	80.0
train	48.3	82.4	85.4	87.7	48.9	**92.5**	53.9	67.5	89.9
motorcycle	18.0	44.1	47.9	52.7	26.1	67.5	32.7	64.6	**69.8**
bicycle	23.6	50.7	52.8	60.4	36.7	68.6	41.5	72.6	**70.9**
mIoU/%	35.3	55.4	65.5	68.0	62.8	77.5	48.7	67.2	**78.3**

As shown in Table 5, our RFGLNet achieves outstanding overall performance with a mean intersection over union (mIoU) of 78.3%, surpassing not only the second-place Rein (77.5%) but also significantly outperforming methods like HRDA (68.0%) and HGFormer (67.2%), while maintaining a stable advantage across multi-class tasks. Specifically, categories like walls and fences show substantial improvements. This is because walls and fences often exhibit low contrast and blurred edges under adverse weather conditions. FECA's frequency-domain denoising effectively enhances their edge features, while GSAM's grouped spatial attention mechanism precisely captures their fragmented local details. In contrast, inherently high-contrast categories like sky and road show only moderate gains, as their global semantic features are already well-captured by the DINOv2 pre-trained model.

Furthermore, to visually demonstrate the segmentation results of different methods, we selected 4 common images to present multiple visual comparison examples of DAFormer, Refign, Rein, and our proposed method. These visual comparison results can be clearly observed in Figure 8.

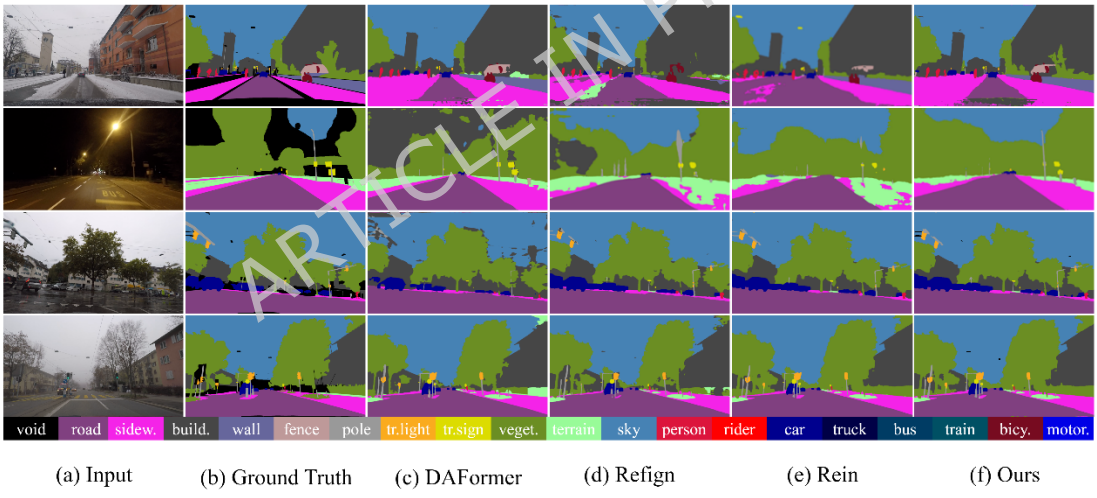


Figure 8. The visual comparison results on the ACDC dataset. From left to right: Input, ground truth, DAFormer, Refign, Rein and our method.

However, the comprehensive performance evaluation of a model should focus not only on overall segmentation accuracy but also on the scale of trainable parameters and their robustness in specific subscenarios. To this end, Table 6 focuses on comparing various methods in terms of their backbone, trainable parameters, and performance across the four subscenarios of ACDC.

Table 6 Comprehensive comparison results with several domain generalization methods on the ACDC testing set. The metric consists of the backbone, trainable parameter, four subsets (Fog, Night, Rain, Snow), and overall mIoU.

Methods	Backbone	Trainable Parameters(M)	IoU/%				
			Fog	Night	Rain	Snow	All
RefineNet [40]	ResNet-101	44.6	46.4	29.0	52.6	43.3	43.7

Methods	Backbone	Trainable Parameters(M)	IoU/%				
			Fog	Night	Rain	Snow	All
Mask2former [25]	Swin-L	216	69.1	53.1	68.3	65.2	65.0
SegFormer [22]	MiT-B5	84.7	63.2	47.8	66.4	63.7	62.0
Rein [21]	DINOv2	2.99	76.4	70.6	79.4	79.5	77.6
RFGLNet(Ours)	DINOv2	4.32	79.9	70.1	81.4	81.9	78.3

The comparison results in Table 6 indicate that, compared with the other methods, our RFGLNet achieves the best performance under three adverse weather conditions (fog, rain, and snow). As a representative method for cross-domain segmentation in unstructured scenarios, IndiVNet [1] achieves efficient adaptation to both structured and unstructured environments with 43.7 million parameters. In contrast, our RFGLNet achieves an mIoU of 78.3% when only 4.32 M trainable parameters are used—fully demonstrating that our approach strikes a better balance between runtime efficiency and accuracy.

4.5.2. Model Efficiency Testing

To verify the applicability for autonomous driving (latency-sensitive), we tested deployment metrics on two hardware platforms: NVIDIA RTX 3080 Ti. Metrics include time and memory consumption during the training phase, as well as throughput and memory consumption during the inference phase.

Table 7. Verification of the efficiency of different real-time semantic segmentation methods on the ACDC dataset using RTX 3080 Ti.

Methods	Training		Inference	
	Time	Memory	Throughput (imgs/s)	Memory
Refign [19]	8.6h	9.8G	30.4	4.1G
FADA [41]	10.5h	12.8G	20.7	6.2G
SET [48]	9.8h	12.5G	21.3	5.8G
SoMA [23]	9.5h	12.7G	56.4	4.4G
Rein [21]	10.3h	10.8G	33.6	4.7G
RFGLNet(ours)	10h	11.2G	57.1	4.4G

As shown in Table 7, under the validation scenario using the RTX 3080 Ti hardware platform and the ACDC dataset, RFGLNet demonstrated efficiency advantages tailored for autonomous driving deployment needs during both training and inference phases: During training, its 10.1-hour training time falls within a reasonable range, while its 11.2GB training memory consumption is significantly lower than methods like FADA and SET, reducing hardware resource requirements for training. During the inference phase—a core concern for autonomous driving—RFGLNet achieved the highest throughput among all tested methods at 57.1 images per second. Simultaneously, its inference memory consumption remained low at 4.4 GB, balancing high real-time performance with low resource demands for deployment. These results validate RFGLNet's effectiveness in lightweight design and computational efficiency optimization.

5. Conclusion

In this paper, we propose a domain-generalized semantic segmentation network for adverse weather conditions called RFGLNet, which achieves a better balance between accuracy and parameter efficiency. Our proposed RFGLNet employs the frozen visual foundation model DINOv2 as the backbone, retaining its robust feature extraction capability, and further optimizes and enhances features through the synergy of three modules. Specifically, we design the SVD-initialized low-rank module (SLRM) to achieve precise alignment of feature distributions and compression of the parameter count. The Fourier-enhanced channel attention module (FECA) is used to enhance features, strengthening their anti-interference capability. The grouped modeling spatial attention module (GSAM) is incorporated

into the decoder, enabling the decoder to achieve precise spatial attention allocation on features of different resolutions, with the final step fusing the enhanced features of all groups.

Extensive experiments are conducted on the Cityscapes → ACDC adverse weather domain generalization task, verifying that our method achieves an mIoU of 78.3% on the ACDC test set with only 4.32 M trainable parameters. For future work, we will consider incorporating more adverse weather data and improving/optimizing our method to adapt to more complex real-world scenarios.

Conflicts of authorship contribution statement

Introduction: Xin Ye and Xiaoqi Shi; methodology: Xin Ye and Xiaoqi Shi; software: Xiaoqi Shi; validation: Yuxue Li; writing-original draft preparation: Xin Ye and Xiaoqi Shi; writing-review and editing: Yuxue Li. All the authors have read and agreed to the published version of the manuscript.

Data availability

The datasets used and/or analysed during the current study are available in the [Cityscapes official repository], [https://www.cityscapes-dataset.com/] and in the [ACDC official repository], [https://acdc.vision.ee.ethz.ch/].

Declaration of competing interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Funding declaration

No funding was received for this paper.

REFERENCES.

[1] Chakraborty, P., Bandyopadhyay, A., Bhattacharyya, S. et al. IndiVNet A region adaptive semantic image segmentation for autonomous driving in unstructured environments. Sci Rep (2025). https://doi.org/10.1038/s41598-025-32305-2.

[2] Alzanin, S. Explainable artificial intelligence with temporal convolutional networks for adverse weather condition detection in driverless vehicles. Sci Rep 15, 19475 (2025). https://doi.org/10.1038/s41598-025-05136-4.

[3] He, K., Zhang, X., Ren, S., & Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90.

[4] Wang, Y., Chen, X., & Zhang, L. Deep degradation prior for low-quality image classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11046-11055 (2020). https://doi.org/10.1109/CVPR42600.2020.01106.

[5] Li, J., Zhang, H., Liu, Y., & Tang, X. UniMix: Toward domain adaptive and generalizable LiDAR semantic segmentation in adverse weather. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14781-14791 (2024). https://doi.org/10.1109/CVPR52733.2024.01400.

[6] Schalfuss, J., Müller, T., & Geiger, A. Distracting downpour: Adversarial weather attacks for motion estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 10072-10082 (2023). https://doi.org/10.1109/ICCV51070.2023.00927.

[7] Simonyan, K., & Zisserman, A. Very deep convolutional networks for large-scale image recognition [C]. 3rd International Conference on Learning Representations (ICLR), (2015).

[8] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. Densely connected convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2261-2269 (2017). https://doi.org/10.1109/CVPR.2017.243.

[9] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. MobileNets: Efficient convolutional neural networks for mobile vision applications. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), (2017). https://doi.org/10.48550/arXiv.1704.04861.

[10] Zamir, S. W., Arora, A., Khan, S. H., Hayat, M., Khan, F. S., & Yang, M. H. Multistage progressive image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14816-14826 (2021). https://doi.org/10.1109/CVPR46437.2021.01458.

[11] Zamir, S. W., Afifi, M., Khan, S. H., Hayat, M., Khan, F. S., Yang, M. H., & Shao, L. Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5718-5729 (2022). https://doi.org/10.1109/CVPR52688.2022.00564.

[12] Peng, D., Lei, Y. J., Hayat, M., Guo, Y. L., & Li, W. (2022). Semantic-aware domain generalized segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2584-2595 (2022). https://doi.org/10.1109/CVPR52688.2022.00262.

[13] Yang, X., Yan, W., Yuan, Y., Mi, M. B., & Tan, R. T. Semantic segmentation in multiple adverse weather conditions with domain knowledge retention. arXiv preprint arXiv:2401.07459, (2024). https://doi.org/10.48550/arXiv.2401.07459.

[14] Bi, L., Zhang, W., Zhang, X., & Li, C. A Nighttime Driving-Scene Segmentation Method Based on Light-Enhanced Network. World Electr Veh J n15, 490 (2024). https://doi.org/10.3390/wevj15110490.

[15] Li, Y., Chang, Y., Yu, C., & Yan, L. Close the loop: A unified bottom-up and top-down paradigm for joint image deraining and segmentation. In Poceedings of the AAAI Conference on Artificial Intelligence, 36(2), 1438-1446 (2022). https://doi.org/10.1609/aaai.v36i2.20033.

[16] Lu, Z., Wang, H. B., Wang, M. Y., & Wang, Z. W. Improved dark channel priori single image defogging technique using image segmentation and joint filtering.Science Progress, 107(1), 1-31 (2024). https://doi.org/10.1177/00368504231221407.

[17] Guo, X., Liu, Y., Xue, W., Zhang, Z., & Zhuang, Y. Low-Light Enhancement and Global-Local Feature Interaction for RGB-T Semantic Segmentation. IEEE Transactions on Instrumentation and Measurement, vol. 74, 1-13 (2025). https://doi.org/10.1109/TIM.2025.3545511.

[18] Ding, J., Xue, N., Xia, G. S., Schiele, B., & Dai, D. X. HGFormer: Hierarchical grouping transformer for domain generalized semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 15413-15423 (2023). https://doi.org/10.1109/CVPR52729.2023.01479.

[19] Bruggemann, D., Sakaridis, C., Truong, P., & Van Gool, L. Refign: Align and refine for adaptation of semantic segmentation to adverse conditions. IEEE Workshop on Applications of Computer Vision (WACV), 3173-3183 (2023). https://doi.org/10.1109/WACV56688.2023.00319.

[20] An, J., He, Z., Guo, J. et al. Unpaired image to image translation for source free domain adaptation in semantic segmentation. Sci Rep 15, 23318 (2025). https://doi.org/10.1038/s41598-025-05648-z.

[21] Wei, Z. X., Chen, L., Jin, Y., Ma, X. X., Liu, T. L., Ling, P. Y., Wang, B., Chen, H. A., & Zheng, J. J. Stronger, Fewer, & Superior: Harnessing Vision Foundation Models for Domain Generalized Semantic Segmentation. Conference on Computer Vision and Pattern Recognition (CVPR), 28619-28630 (2023). https://doi.org/10.1109/CVPR52733.2024.02704.

[22] Xie, E., Wang, W. H., Yu, Z. D., Anandkumar, A., Alvarez, J. M., & Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. In Proceedings of the European Conference on Computer Vision and Pattern Recognition, arXiv:2105.15203 (2021). https://doi.org/10.48550/arXiv.2105.15203.

[23] S. Yun, S. Chae, D. Lee and Y. Ro, "SoMA: Singular Value Decomposed Minor Components Adaptation for Domain Generalizable Representation Learning,"2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 25602-25612, (2025). https://doi.org/10.1109/CVPR52734.2025.02384.

[24] Hoyer, L., Dai, D. X., & Van Gool, L. DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition (CVPR), 9914-9925 (2022). https://doi.org/10.1109/CVPR52688.2022.00969.

[25] Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1280-1289 (2022). https://doi.org/10.1109/CVPR52688.2022.00135.

[26] Cordts, M., Enzweiler, M., Omran, M., Benenson, R., Ramos, S., Franke, U., Rehfeld, T., Roth, S., & Schiele, B. The Cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3213–3223 (2016). https://doi.org/10.1109/CVPR.2016.350.

[27] Sakaridis, C., Wang, H., Li, K., Zurbrügg, R., Jadon, A., Abbeloos, W., Olmeda Reino, D., Van Gool, L., & Dai, D. X. ACDC: The adverse conditions dataset with correspondences for robust semantic driving scene perception. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 10745-10755 (2021). https://doi.org/10.1109/ICCV48922.2021.01059.

[28] Sakaridis, C., Dai, D. X., & Van Gool, L. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 3139-3153 (2020). https://doi.org/10.1109/TPAMI.2020.3045882.

[29] Hoyer, L., Dai, D. X., & Van Gool, L. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. Computer Vision–ECCV 2022, 372–391 (2022). https://doi.org/10.1007/978-3-031-20056-4_22.

[30] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), 3–19 (2018). https://doi.org/10.1007/978-3-030-01234-2_1.

[31] Hu, J., Shen, L., & Sun, G. Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(8), 2011-2023 (2018). https://doi.org/10.1109/TPAMI.2019.2913372.

[32] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 11531-11539 (2020). https://doi.org/10.1109/CVPR42600.2020.01155.

[33] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. LoRA: Low-rank adaptation of large language models. The Tenth International Conference on Learning Representations (ICLR) arXiv preprint arXiv:2106.09685 (2022). https://doi.org/10.48550/arXiv.2106.09685

[34] Li, X., & Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 4582–4597 (2021). https://doi.org/10.18653/v1/2021.acl-long.353.

[35] Li Z Y, Lu J H, Deng J C, et al. SAS: Segment Any 3D Scene with Integrated 2D Priors [EB/OL]. arXiv preprint arXiv:2503.08512 (2025). https://doi.org/10.48550/arXiv.2503.08512.

[36] Sun, C. W., Wei, J. W., Wu, Y. J., Shi, Y. M., He, S. Y., Ma, Z. Y., Xie, N., & Yang, Y. (2024). SVFit: Parameter-efficient fine-tuning of large pretrained models using singular values. arXiv preprint arXiv:2409.05926. https://doi.org/10.48550/arXiv.2409.05926.

[37] Yang, Y. C., & Soatto, S. FDA: Fourier domain adaptation for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4084-4094 (2020). https://doi.org/10.1109/CVPR42600.2020.00414.

[38] Yang, G. FreqCross: A Multi-Modal Frequency-Spatial Fusion Network for Robust Detection of Stable Diffusion 3.5 Generated Images. Computer Vision and Pattern Recognition. arXiv preprint arXiv:2507.02995. https://doi.org/10.48550/arXiv.2507.02995

[39] Qian, C., Rezaei, M., Anwar, S., Li, W. J., Hussain, T., Azarmi, M., & Wang, W. AllWeather-Net: Unified image enhancement for autonomous driving under adverse weather and low-light conditions. In Proceedings of the 2024 International Conference on Pattern Recognition (ICPR), Lecture Notes in Computer Science, vol. 15330, 151-166 (2024). https://doi.org/10.1007/978-3-031-78113-1_11.

[40] Lin, G., Milan, A., Shen, C., & Reid, I. RefineNet: Multipath refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5168-5177 (2017). https://doi.org/10.1109/CVPR.2017.549.

[41]Bi, Q., Yi, J.J., Zheng, H., Zhan, H.L., Huang, Y.W., Ji, W., Li, Y.X., Zheng, Y.F. Learning Frequency-Adapted Vision Foundation Model for Domain Generalized Semantic Segmentation. InNeurIPS (2024).

[42] Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., & Bojanowski, P. DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023). https://doi.org/10.48550/arXiv.2304.07193.

[43] Qin, Z., Zhang, P., Wu, F., & Li, X. FcaNet: Frequency Channel Attention Networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 763-772 (2021). https://doi.org/10.1109/ICCV48922.2021.00082.

[44] Loshchilov, I., & Hutter, F. Decoupled weight decay regularization. In Proceedings of the 7th International Conference on Learning Representations (ICLR). (2019).

[45] Loshchilov, I., & Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In Proceedings of the International Conference on Learning Representations (ICLR). (2017).

[46] Long, J., Shelhamer, E., & Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3431-3440 (2015). https://doi.org/10.1109/CVPR.2015.7298965.

[47] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., & Liu, W. Deep high-resolution representation learning for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(10), 3349-3364 (2021). https://doi.org/10.1109/TPAMI.2020.2983686.

[48]Yi, J.J., Bi, Q., Zheng, H., Zhan, H.L., Ji, W., Huang, Y.W., Li, Y.X., Zheng, Y.F. Learning Spectral-Decomposited Tokens for Domain Generalized Semantic Segmentation. In ACMMM, 8159-8168 (2024). https://doi.org/10.1145/3664647.3680906.

[49] A. Gomaa. Advanced Domain Adaptation Technique for Object Detection Leveraging Semi-Automated Dataset Construction and Enhanced YOLOv8[C]//Proceedings of the 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES 2024). IEEE, 211-214, (2024). https://doi.org/10.1109/NILES63360.2024.10753164.

[50]A. Gomaa, A. Abdalrazik. Novel Deep Learning Domain Adaptation Approach for Object Detection Using Semi-Self Building Dataset and Modified YOLOv4[J]. World Electric Vehicle Journal, 15(6), 255 (2024). https://doi.org/10.3390/wevj15060255.

[51]Wang, Y., Li, Y. S., Wang, G., & Liu, X. G. Multi-scale Attention Network for Single Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 5950-5960, (2024), https://doi.org/10.1109/CVPRW63382.2024.00602.