Frequency-domain guided swin transformer and global-local feature integration for remote sensing images semantic segmentation.

ZHANG, H., XIE, G., LI, L., XIE, X. and REN, J.

2025

Frequency-Domain Guided Swin Transformer and Global-Local Feature Integration for Remote Sensing Images Semantic Segmentation

Haoxue Zhang, Gang Xie, *Member, IEEE,* Linjuan Li, Xinlin Xie, and Jinchang Ren, *Senior Member, IEEE,*

***Abstract*—Convolutional Neural Networks (CNNs), transformers, and the hybrid methods have been significant application in remote sensing. However, existing methods are limited in effectively modeling frequency domain information, which affects their ability to capture detailed information. Therefore, we propose a frequency-domain guided feature coupled mechanism and a global-local feature integration method (FGNet) for semantic segmentation. Specifically, a frequency-domain guided Swin transformer (FGSwin) is designed by introducing dilation group convolution, Fast Fourier Transform (FFT) and learnable weights to enhance the expression capability of frequency-domain and space-domain, local and global features, simultaneously. In addition, a global-local feature integration module (GLFI) is proposed for aggregating features to further enhance the discrimination of each category. Comprehensive experimental results demonstrate that, compared to existing methods, the proposed method achieves superior performance in terms of mean intersection over union (mIoU), reaching 71.46% and 74.04% on the ISPRS Potsdam and Vaihingen, two widely used datasets.**

***Index Terms*—Remote sensing semantic segmentation, global-local features, frequency-domain guided Swin transformer, feature integration**

I. Introduction

With the development of remote sensing technology, the availability of high-resolution images has increased, and the research on remote sensing semantic segmentation methods has been intensified. The purpose of semantic segmentation is to assign labels for each pixel to obtain the interpretation of the whole image, which has important applications in precision agriculture [1], urban development [2], natural resource utilization [3], and disaster assessment [4].

Semantic segmentation of remote sensing images consists of traditional learning-based methods and deep learning-based methods, the former is accomplished by manually designing features and segmentation criteria, whose generalization ability and robustness need to be strengthened; the latter automatically captures the image information with a high level of intelligence and has received extensive attention in recent years. Among the methods based on deep learning, those of convolutional neural networks (CNNs), transformers and their hybrid architectures occupy the mainstream.

CNNs extract local image features through convolution operations, enabling multiscale feature representation. Fully convolutional networks (FCN) [5] introduced per-pixel classification, followed by U-net [6] with a symmetric encoder-decoder structure. PSPNet [7] and FPN [8] expanded receptive fields with spatial pyramid pooling, while other methods [9], such as orientation attention network [10] and stair fusion network [11], incorporated attention mechanisms. However, CNNs rely on local operations, limiting their ability to capture global information, which is critical for semantic segmentation. To address the limitations of CNNs in modeling global context, transformers utilize a multi-head self-attention (MSA) mechanism to extract global information, overcoming the constraints of local receptive fields in convolutional networks. This capability enhances image interpretation by differentiating objects of the same category from others [12], [13]. The Rss-former [14] employs an adaptive transformer fusion module to suppress background noise and enhance foreground saliency. The boundary-aware multiscale network [15] introduces a scale attention module to construct long-range dependencies. Mixed-mask transformer [16] uses a hierarchical encoder for multiscale learning. Despite their effectiveness in capturing global information, transformers face challenges in computational complexity. The Swin transformer [17] mitigates this issue through non-overlapping local windows, ensuring linear scalability with image size. Nonetheless, further optimization of information extraction mechanisms is necessary, highlighting the potential of hybrid architectures.

Recognizing the strengths of CNNs in capturing local information and the capability of transformers to model global features, hybrid architectures have been developed to integrate the benefits of both approaches. UNetFormer [18] and CNNs

This work was supported in part by the Fundamental Research Funds for the Key Research and Development Plan of Shanxi Province 202202010101005, Industry-University-Research Innovation Fund for Chinese Universities 2021ZYA11005, and in part by the Guangdong Province Key Construction Discipline Scientific Research Ability Promotion Project (2022ZDJS015, 2021ZDJS025), Special Projects in Key Fields of Ordinary Universities of Guangdong Province under Grant 2021ZDZX1087, and the Guangzhou Science and Technology Plan Project under Grants (2024B03J1361, 2023B03J1327). *(Corresponding author: Gang Xie.)*

Haoxue Zhang, Gang Xie, Linjuan Li, and Xinlin Xie are with the Shanxi Key Laboratory of Advanced Control and Industrial Intelligence and the School of Electronic and Information Engineering, Taiyuan University of Science and Technology, Taiyuan, 030024, China (e-mail: zhanghaoxue95@stu.tyust.edu.cn; xiegang@tyust.edu.cn; linjuanli@tyust.edu.cn; xiexinlin@tyust.edu.cn).

Jinchang Ren is with the School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China. He is also with the School of Computing, Engineering and Technology, Robert Gordon University, Aberdeen, U.K. (e-mail: jinchang.ren@ieee.org).

and multiscale transformer fusion network (CMTFNet) [19] build a hybrid architecture, which use CNNs as an encoder and transformer as a decoder. Conversely, class-guidance network [20] uses a transformer as an encoder to mine the class-specific perceptual information for each semantic class, while CNNs as a decoder to enhance the information. Moreover, SRCBTFusion-Net [21] is designed through a cascade dual-coding structure and a multi-scale up-sampling integration module in which resolution information can be maintained. But the cascade structure usually increases the number of parameters, which requires the use of lightweight backbone network of pre-trained for feature extraction. Therefore, dual-branch networks, due to their efficiency, have been gradually proposed, such as STUNet [22], a global and edge enhanced transformer [23] and CSTUNet [24].

While hybrid architectures combining CNNs and transformers effectively integrate spatial domain information, they predominantly emphasize spatial representations, often neglecting the potential of frequency domain modeling. In contrast, robust representation of frequency domain information significantly enhances semantic segmentation performance, as high-frequency characteristics—particularly lines and edges—play a crucial role in delineating category boundaries and improving overall accuracy. A straightforward approach is to incorporate the frequency domain into convolution or attention mechanisms. For instance, frequency adaptive dilated convolution [25] dynamically adjusts dilation rates based on local frequency components, whereas frequency domain feature-guided network [26] employs a frequency enhancement attention module to identify and strengthen frequency details. While these methods have made some progress in feature extraction, there is still scope for improvement in the effective utilisation of frequency characteristics, which may restrict the performance of the models in complex scenarios. To make more effective use of the frequency domain features, multi-scale frequency attention gating network [27] and SFFNet [28] focus on frequency decomposition, which isolates image components across various frequency bands, particularly around edges. Additionally, the frequency-driven edge network [29] sharpens boundary definitions by employing a two-dimensional discrete wavelet transform to suppress low-frequency noise while accentuating salient edge details within spatial features. SSCNet [30] proposes spectral attention to capture the spectral context in the frequency domain, which maps the feature map into the frequency domain and calculates edge loss for high-frequency components. Despite the effectiveness of these methods in addressing complex targets, they still require refinement in their integration with spatial information to enhance both robustness and precision. A growing number of approaches now aim to explore more sophisticated frequency-space fusion strategies. MsanlfNet [31] combines multi-scale attention with a non-local filter to process spatial and frequency characteristics at multiple scales. MIFNet [32] builds on this by incorporating local, global, and frequency data within a unified module. The dual-domain fusion network [33], based on wavelet frequency decomposition and fuzzy spatial constraints, further enhances segmentation by integrating spatial and frequency information. wavelet feature enhancement [34] performs a multi-scale, lossless decomposition of the input image, which helps preserve high-frequency details. Spatial-frequency network [35] embeds contextual feature dependencies in both spatial and frequency domains. Nevertheless, current approaches typically incorporate frequency-domain transformations within convolutional or attention mechanisms or integrate frequency-domain information as a supplementary module. However, these strategies exhibit inherent limitations, as any information lost during the feature extraction process cannot be effectively recovered in later stages. Thus, there is a pressing need for further research to focus on optimizing the interplay between frequency-domain and spatial-domain processing during feature extraction.

Therefore, based on the hybrid architecture, we propose a frequency-domain information enhancement mechanism, which is integrated into the feature extraction process to improve the combined representation of global-local and frequency-space domain information. Additionally, we design a global-local information integration method to further enhance feature effectiveness. Our main contributions are summarized as follows:

- Utilizing learnable weights and shallow feature extraction methods, we propose a frequency-domain information enhancement mechanism (FGSwin) and couple it into Swin transformers, which can model frequency-domain and space-domain information and is different with exist frequency information utilizing approaches. By improving the feature description capability and coupling it into the Swin transformer for better long-range dependent information modeling.
- We develop a global-local feature integration module (GLFI) by utilizing a small convolutional kernel and a more easily convergent activation function. This design enhances class discrimination while maintaining a low parameter count in the network.
- A hybrid architecture for semantic segmentation of remote sensing images based on FGSwin, residual CNNs, and GLFI is designed. This architecture achieves performance of 71.46% and 74.04% on widely used Potsdam and Vaihingen datasets, which demonstrates competitive results compared to the state-of-the-art (SOTA) methods.

The remainder of the article is organized as follows. Section II provides the motivation and overview of the FGNet. In Section III, the results of specific experiments are reported. In Section V, the conclusions are summarized.

II. Methodology

A. Motivation and overall Architecture of the Network

Semantic segmentation is a process of recognizing homogeneity and heterogeneity information in the image, usually, homogeneity manifests itself in low-frequency, and heterogeneity can be distinguished by high-frequency information. Frequency domain analysis converts an image from physical space to frequency space with clear meaning. It cannot only obviously distinguish between low-frequency components and

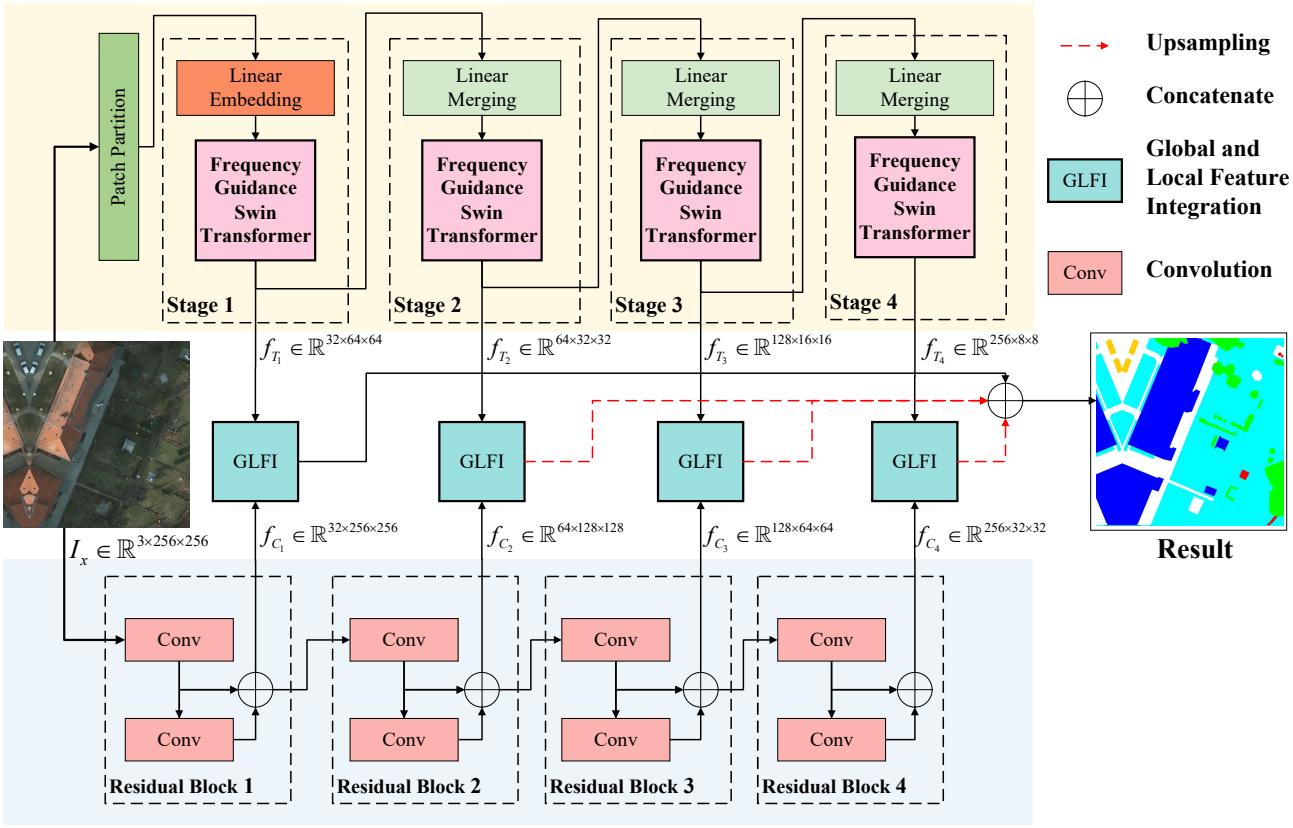

Fig. 1. The overall of propose FGNet.

high-frequency components but also reduce processing complexity. Therefore, the methods modeling frequency characteristics can improve the performance of semantic segmentation theoretically [27], [36], [37].

However, the existing CNNs-based and transformer-based methods only capture local and global information in the physical space and are limited in analyzing frequency features. The existing frequency analysis methods usually represent the frequency information by constructing the fusion module or attention in combination with the network [31], [29], [37] and the validation of more effective coupling ways and applications in semantic segmentation need to be further investigated. Thus, we propose a frequency-domain guided feature enhancement mechanism for semantic segmentation and couples it into Swin transformer. Assisting with the global-local feature integration module, the effective modeling of global-local and frequency domain information can be achieved. The specific network structure is shown in Fig.1. FGNet uses FGSwin transformer and residual CNNs as dual-branch encoder to fully extract the features of the image, which contains four stage blocks. Also, the corresponding features of each stage are fed into GLFI to integrate local and global context expressed by designed FGSwin and CNNs.

B. *Frequency-domain guided Swin Transformer*

Since different information manifests as different frequency components in the frequency domain, the discrimination of some information, including lines and edges, is improved compared to the physical space. Thus, considering enhancing different frequency components to different degrees through learnable weights, self-attention to different frequency components can be achieved, which in turn guides the network to optimize in the direction more conducive to category discrimination. To obtain different frequency components, it is necessary to transform between the physical space and the frequency domain through the Fast Fourier transform (FFT) and inverse transform (IFFT) [31].

A single completion of frequency domain transformation and enhancement is not enough, an important issue is how to couple the above processes into the deep network structure. Considering different networks have different extraction capabilities, if some information has been neglected in the extraction process, coupling it into the feature fusion does not compensate for the information. Therefore, it becomes one of the feasible ways to couple it into the feature extraction network to guide the process and realize the comprehensive analysis of frequency domain and physical space information. Swin transformer [17] designs a sliding window mechanism to represent multiscale information by restricting self-attentive

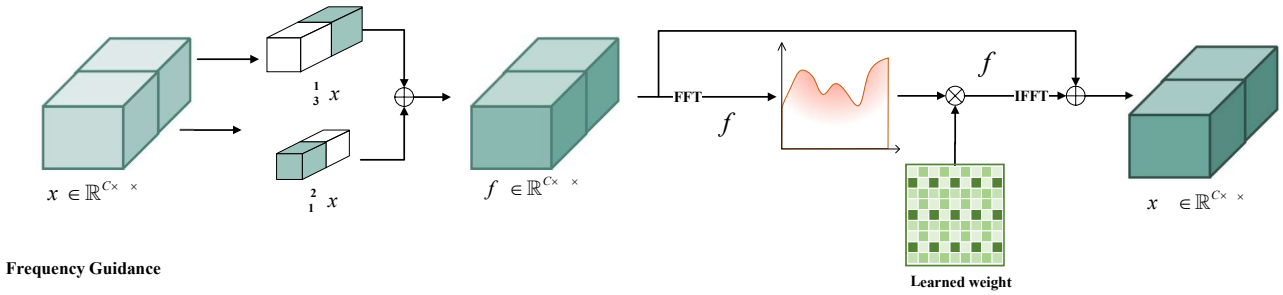

Fig. 2. The details of frequency-domain guidance in FGSwin.

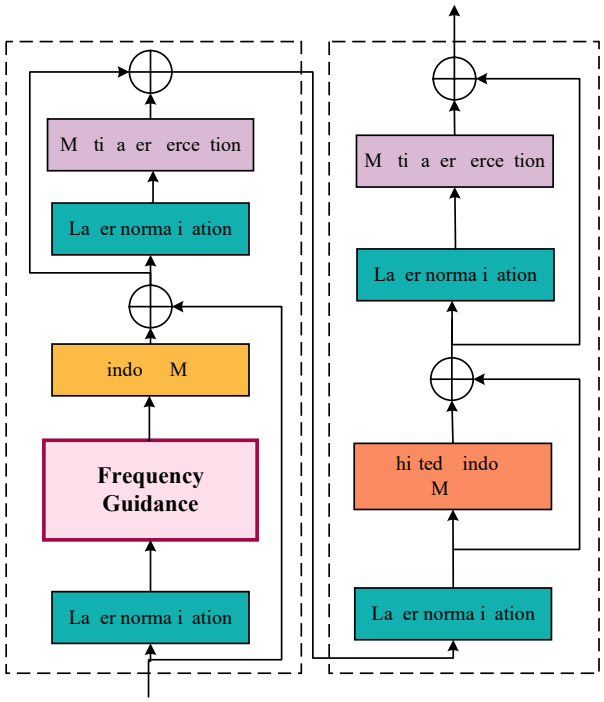

Fig. 3. Architecture of the FGSwin.

computation to non-overlapping local windows with a low computational complexity, which has been used for semantic segmentation in the last two years and has obtained better result. Hence, a frequency-domain guided Swin transformer is designed, as shown in Fig.2 and Fig.3, at the same time, group convolutional structures with different dilation parameters are proposed to obtain the shallow features simultaneously. This processing way can increase the receptive field and reduce the parameters.

Given an image $\mathbf{I}_x$, partitioning it as $\mathbf{I}_{x_i}$ where i is the index of the image patch. Then performing shallow feature extraction using group convolution, which can be represented as

$$\mathbf{x}_i = \mathrm{LN}(\mathrm{F}(\mathrm{P}(\mathbf{I}_{\mathrm{x_i}}))) \tag{1}$$

$$\mathbf{f}_{\mathrm{G}} = \mathrm{conv}_1(\mathrm{CAT}(\mathrm{conv}_3^1(\mathbf{x}_{\mathrm{i}}), \mathrm{conv}_1^2(\mathbf{x}_{\mathrm{i}}))) \tag{2}$$

where $\mathrm{P}(\cdot)$, $\mathrm{F}(\cdot)$ and $\mathrm{LN}(\cdot)$ represents the patch embedding, flatten and layer normalization. $\mathbf{x}_i$ is the output of the linear embedding and also the input of the frequency guidance. $\mathrm{conv}_{\mathrm{k}}^{\mathrm{d}}(\cdot)$ represents the convolutional operation with the dilation parameter is d and kernel size is k, while $\mathrm{conv}_{\mathrm{k}}(\cdot)$ refers to the convolution operation with kernel size k, $\mathrm{CAT}(\cdot)$ represents the concatenate operations. $\mathbf{f}_{\mathrm{G}}$ is the output feature of the group convolution and the input feature of FFT. This processing way can reduce the parameter quantity of each convolution layer, thereby preventing over-fitting. Furthermore, the use of small scale convolution kernels and different dilation parameters provide a scheme for developing lightweight models and improving receptive fields [38].

To express the information in the frequency domain, FFT and IFFT are employed to transform shallow features in the frequency domain. Furthermore, learnable weights are introduced in the frequency domain to facilitate adaptive attention to different frequency components, thereby optimizing and enhancing features. The specific calculation formula of the processing is as follows:

$$\mathbf{f}_{\mathrm{SF}} = \mathrm{FFT}(\mathbf{f}_{\mathrm{G}}) \tag{3}$$

$$\mathbf{f}_{\mathrm{FW}} = \mathbf{f}_{\mathrm{SF}} \odot \mathbf{W}_{\mathrm{L}} \tag{4}$$

$$\mathbf{f}_{\mathrm{FS}} = \mathrm{IFFT}(\mathbf{f}_{\mathrm{FW}}) + \mathbf{f}_{\mathrm{G}} \tag{5}$$

where $\mathrm{FFT}(\cdot)$ and $\mathrm{IFFT}(\cdot)$ are FFT and IFFT operation, $\mathbf{W}_{\mathrm{L}}$ represented the learnable weight parameter. $\mathbf{f}_{\mathrm{SF}}$ and $\mathbf{f}_{\mathrm{FS}}$ denote the frequency domain features obtained by FFT and the frequency domain features obtained by IFFT, respectively. $\mathbf{f}_{\mathrm{FW}}$ is frequency-domain feature with the learned weight and $\odot$ is Hadamard product of the two tensors. It can be demonstrated that the frequency domain enhancement mechanism is not located at the front end of feature coding nor within the fusion module, but rather is coupled within the feature extraction process. Shallow features are often rich in line and edges, and enhancing the frequency domain of shallow features can improve the effectiveness of the input features. Consequently, the guiding mechanism is incorporated into the attention model within the Swin transformer [17] structure to obtain the final features. Specifically, in the FGSwin, the output of frequency-domain enhanced feature f_F undergoes layer normalization followed by the windows MSA as follows:

$$\hat{\mathbf{x}}^l = \mathbf{f}_{\mathrm{FS}} + \mathrm{W\text{-}MSA}(\mathrm{LN}(\mathbf{f}_{\mathrm{FS}})) \tag{6}$$

where $\hat{x}^l$ is the output at layer l, $\mathrm{W\text{-}MSA}(\cdot)$ represents the windows MSA, and $\hat{\mathbf{x}}^l$ denotes the output of the windows MSA for block l. This process helps reduce the computational

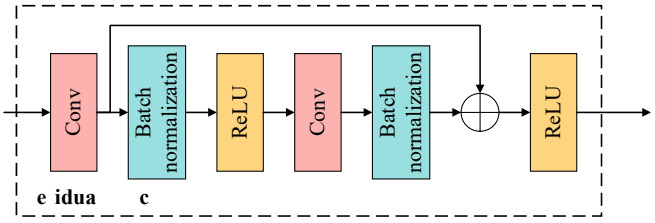

Fig. 4. Structure of the CNNs branch.

complexity of traditional self-attention by limiting the scope of attention to smaller regions. Subsequently, the output $\hat{x}^l$ is processed by a multi-layer perception after another layer of normalization.

$$\mathbf{x}^l = \hat{\mathbf{x}}^l + \mathrm{MLP}(\mathrm{LN}(\hat{\mathbf{x}}^l)) \tag{7}$$

where $\mathrm{MLP}(\cdot)$ is the multi-layer perception, and $\mathbf{x}^l$ is the output of multi-layer perception for block l. This step further refines the representation by incorporating non-linear transformations while preserving the residual connection for stability. In addition, the output feature is further processed by shift windows MSA, which can be represented as

$$\hat{\mathbf{x}}^{l+1} = \mathbf{x}^l + \text{SW-WSA}(\mathrm{LN}(\mathbf{x}^l)) \tag{8}$$

where $\text{SW-MSA}(\cdot)$ is shift windows MSA, which introduces a shift in the window partitioning process. The output $\hat{x}^(l+1)$ is fed into another multi-layer perception following normalization, as depicted as

$$\mathbf{x}^{l+1} = \hat{\mathbf{x}}^{l+1} + \mathrm{MLP}(LN(\hat{\mathbf{x}}^{l+1})) \tag{9}$$

Finally, the features extracted by FGSwin transformer are fed into the last layer normalization calculated by

$$\mathbf{f}_\mathrm{T} = \mathrm{LN}(\mathbf{x}^{\mathrm{l}+1}) \tag{10}$$

where $\mathbf{f}_\mathrm{T}$ represents the final output features of transformer branch. The design of the modules through a multi-scale strategy allows for the acquisition of four distinct groups of feature maps, thereby facilitating the effective expression of global-local, frequency domain, and spatial domain information.

C. Local Feature representation by residual CNNs

FGSwin enables the modeling of long-distance dependence information in frequency domain and the spatial domain. Although local convolution is introduced in shallow features encoding, the ability to represent local information of high-level features needs to be enhanced. Therefore, a residual CNNs is designed to make up for the above problems. The specific residual block structure is shown in Fig. 4.

Given an input $\mathbf{f}_x$, the residual CNNs block can be expressed as follows

$$\mathbf{f}_\mathrm{res} = \mathrm{BN}(\mathrm{conv}_1(\mathrm{ReLU}(\mathrm{BN}(\mathrm{conv}_3(\mathbf{f}_x))))) \tag{11}$$

$$\mathbf{f}_\mathrm{C} = \mathrm{ReLU}((\mathrm{conv}(\mathbf{f}_\mathrm{x}) + \mathbf{f}_\mathrm{res})) \tag{12}$$

where $\mathrm{BN}(\cdot)$ denotes batch normalization and $\mathrm{ReLU}(\cdot)$ represents the activation function of rectified linear unit, $\mathbf{f}_\mathrm{res}$ is the residual feature, and $\mathbf{f}_C$ is the final feature of the residual CNNs branch. The design of the modules through a multi-scale strategy allows for the acquisition of four distinct groups of feature maps, thereby facilitating the effective expression of local information.

D. Global and Local Feature Integration and Loss Function

The global and local information of an image can be fully expressed through the designed FGSwin and CNNs. To achieve effective feature fusion, which is different from previous fusion methods using direct stacking or simple connection, GLFI, a lightweight and intuitive feature fusion method, has been designed as Fig.5.

Assuming that the first stage output features of the FGSwin and residual CNNs are $\mathbf{f}_{\mathrm{T}_1}$ and $\mathbf{f}_{\mathrm{C}_1}$, respectively, feature fusion can be represented as

$$\mathbf{f}_{\mathrm{u}_1} = \mathrm{CAT}(\mathbf{f}_{\mathrm{C}_1} \odot \mathrm{U}(\mathrm{conv}_1(\mathbf{f}_{\mathrm{T}_1})), \mathbf{f}_{\mathrm{C}_1}) \tag{13}$$

$$\mathbf{f}_1 = \mathrm{SiLU}(\mathrm{BN}(\mathrm{conv}_3(\mathrm{f}_{\mathrm{u}_1}))) \tag{14}$$

where $\mathbf{f}_{\mathrm{u}_1}$ is the first concatenated feature, $\mathrm{U}(\cdot)$ represents upsampling, $\mathbf{f}_1$ is the first stage fused feature of GLFI. $\mathrm{SiLU}(\cdot)$ is sigmoid linear unit. This processing way can improve the performance of the model, accelerate the training process and improve the generalization ability. Similarly, other three fused features $\mathbf{f}_2$, $\mathbf{f}_3$ and $\mathbf{f}_4$ can be obtained. The final four features are concatenate as

$$\mathbf{f} = \mathrm{CAT}(\mathrm{U}(\mathbf{f}_1), \mathrm{U}(\mathbf{f}_2), \mathrm{U}(\mathbf{f}_3), \mathbf{f}_4) \tag{15}$$

where $\mathbf{f}$ is the final feature map.

To achieve the optimal configuration of the entire network structure and obtain the segmentation result, the cross-entropy loss function is applied. At this point, we use the pixel-wise representation of the final feature map $\mathbf{f}$, which is the corresponding output from the network, as $\hat{\mathbf{y}}_m^{(n)}$. Thus, the loss function calculated by

$$\mathcal{L}_{ce} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M}\mathbf{y}_m^{(n)} log\hat{\mathbf{y}}_m^{(n)} \tag{16}$$

where N and M represent the number of samples and the number of classes, respectively. $\mathbf{y}_m^{(n)}$ is one-hot vectors of the true labels.

III. Experiments

A. Datasets

To verify the effectiveness of the FGNet, experiments are conducted on two widely used datasets: Potsdam [39] and Vaihingen datasets [40].

1) Potsdam datasets: The datasets contain 38 high-resolution true orthophoto images of the size of 6000×6000, with a ground sampling distance (GSD) of 0.05m and four channels (red: R, green: G, blue: B, infrared: IR), which has been annotated with six categories: Car, Tree, Low Vegetation/grass (Low. Veg.), Building, Impervious Surfaces (Imp. Surf.), and Clutter. The IR-R-G false color images are utilized for the subsequent experiments. Meanwhile, each image is divided into 256×256 patches and randomly allocated to the training set and test set, with 16,082 images and 4,020 images, respectively.

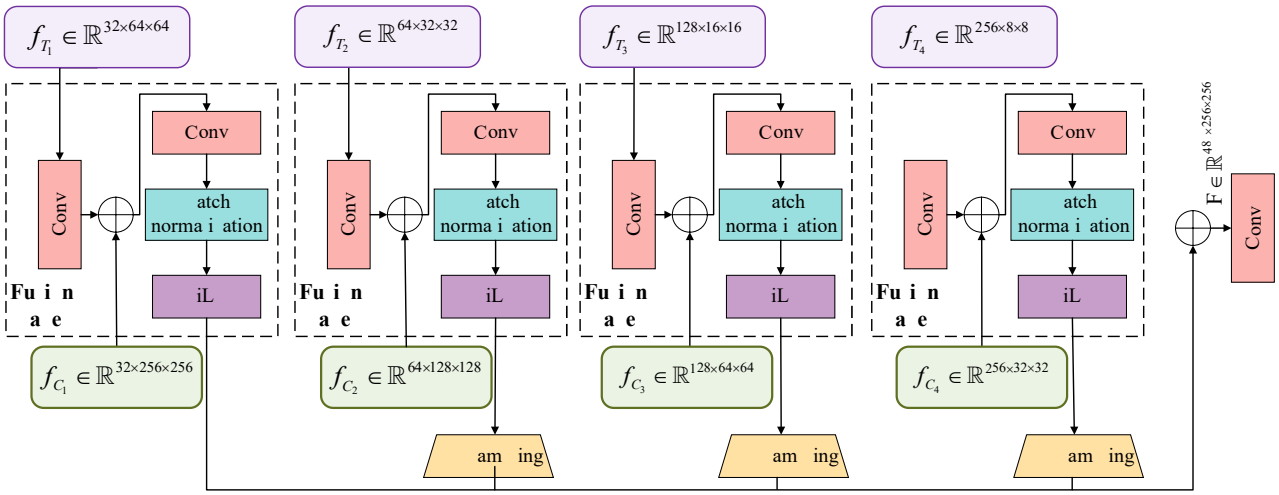

Fig. 5. Strucure of the GLFI module.

2) Vaihingen datasets: The second Vaihingen dataset comprises 33 images, with a GSD of 0.09m. The largest image in the dataset measures 3816×2550 pixels, while the smallest measures 2555×1388 pixels. Collectively, these images cover an area of 1.38 square kilometers in Vaihingen, Germany. Each image comprises three bands: IR, R, and G. The dataset has been annotated according to the six categories of the Potsdam. The dataset has been also divided into 256×256 patches and randomly split into 3344 train images and 210 test images, respectively.

B. Evaluation Metrics and Implement Details

Three metrics are employed to quality the proposed FGNet, including Overall Accuracy (OA), F1-score and mean intersection over union (mIoU). Among these, OA and F1-score are given as

$$\mathrm{OA} = \frac{TP + TN}{TP + FP + TN + FN} \tag{17}$$

and

$$\mathrm{F_1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{18}$$

where OA and F1-score are based on the confusion matrix relying on precision and recall, which calculated by

$$\mathrm{Precision} = \frac{TP}{TP + FP}, \mathrm{Recall} = \frac{TP}{TP + FN} \tag{19}$$

where TP, FP, TN and FN denote true positives, false positives, true negatives, and false negatives, respectively. Consequently, for formal consistency, mIoU can be equated with

$$\mathrm{mIoU} = \frac{1}{N} \frac{TP}{FN + FP + TP} \tag{20}$$

where N is the number of classes.

The mIoU is the average value of intersection over union for all classes, which provides an overall assessment of segmentation accuracy by measuring the degree of overlap between the predicted segmentation mask and the ground truth mask. The OA is defined as the ratio of the number of correctly predicted pixels to the total number of pixels. Precision and Recall are two crucial metrics used to evaluate the performance of models. Precision refers to the proportion of true positive predictions out of all positive predictions made by the model, reflecting ability of the model to avoid false positives. In contrast, Recall represents the proportion of true positive predictions out of all actual positive instances, emphasizing the ability to capture all relevant instances and avoid false negatives. A high Precision indicates fewer false alarms, while a high Recall suggests that the model is successful at identifying most of the relevant instances. However, there is often a trade-off between Precision and Recall: improving one can sometimes result in a decrease in the other. This trade-off is crucial when balancing model performance for different use cases, and the F1-score provides a harmonic mean of both to offer a balanced measure when both metrics are important. It can be demonstrated that the greater the values of the aforementioned indicators, the more accurate the segmentation result.

The experimental environment used a server equipped with an 13th Gen Intel(R) Core (TM) i9-13900K CPU and an NVIDIA GeForce RTX 4090 24G GPU running Linux version Ubuntu 20.04.4LTS. All comparison methods utilize their public code and are conducted within the same test framework and environment. The optimal parameters are selected as well. Adam optimizer is employed with an initial learning rate of 0.001 and a weight decay of $2.5e^{-4}$. To manage the learning rate, we employ the cosine annealing strategy with a warmup and restart, whose hyperparameters are the number of iterations for the first restart and the factor increases of the restart, which are 15 and 2, respectively. Additionally, we set a batch size of 8 and 300 epochs.

C. The comparison methods

A comprehensive array of benchmark methods was selected for quantitative comparison, including FCN [41], U-net [42], SwinB-UperNet [17], SegFormer [13], UNetFormer [18], SR-

CBTFusion [21], STUnet [22], CMTFNet [19], MsanlfNet [31], SFFNet [28] and MIFNet [32].

1) FCN: As the one of the most classic CNNs method, FCN uses the fully connected layer to replace with a convolution layer, enable pixel-wise classification.

2) U-net: U-net represents a further SOTA method of the most classic CNNs, which is structured as a symmetric encoder-decoder comprising dual paths for feature extraction and information propagation.

3) SwinB-UperNet: It is an outperforming transformer architecture combining FPN and Swin transformer to represent global prior information and reduce computational burden.

4) SegFormer: SegFormer is a state-of-the-art transformer architecture model, which comprises a hierarchically structured Transformer encoder that outputs multiscale features.

5) UNetFormer: UNetFormer is a hybrid model combining CNNs and transformer, using ResNet as the encoder and a transformer-based decoder to capture both global and local information effectively.

6) SRCBTFusion: It adopts a cascade CNNs-transformer encoder-decoder structure with semantic information, and edge segmentation to improve the feature selection and feature aggregation.

7) STUNet: STUNet combines the Swin transformer with CNNs in a dual encoder structure, meanwhile, it can enhance the feature representation and accuracy through spatial interaction, feature compression, and relational aggregation.

8) CMTFNet: CMTFNet, as a CNNs-transformer hybrid model, integrates local information from CNNs with multiscale global information from transformer. Moreover, it uses a multiscale MSA and the attention fusion module to enhance feature extraction.

9) MsanlfNet: This method employs the combination of feature fusion module to achieve frequency-domain enhancement, thereby attaining a superior segmentation effect. The method is selected for analysis to ascertain the advantages of double branches and to verify the advantages of coupling frequency guidance into the processing of feature extraction.

10) SFFNet: SFFNet introduces the Haar wavelet-based wavelet transform feature decomposer , which separates and integrates both low- and high-frequency components with spatial features.

11) MIFNet: MIFNet builds on this by incorporating local, global, and frequency data within a unified module, which adaptively associates these dimensions.

D. Comparison with SOTA Methods on Potsdam Datasets

The specific experimental results on the Potsdam Dataset are shown in Table I and Fig. 6. It can be observed that methods based on CNNs or transformers alone achieve relatively low accuracy, primarily because relying on either a CNN or transformer feature extraction framework independently is insufficient for comprehensive information extraction at both global and local levels. For example, FCN achieves 67.49% mIoU and U-net achieves 68.71% mIoU. They perform well in capturing local spatial features through its convolutional layers, yet it struggles to effectively capture global context, which is essential for accurately segmenting complex remote sensing images. SwinB-UperNet and Segformer achieve 65.88% and 64.67% mIoU, which introduce transformer-based mechanisms to capture global information, but lack the ability to model fine-grained frequency details that are critical for delineating precise boundaries and small-scale structures. The integration of CNNs and transformer methods, such as UNetFormer and SRCBTFusion, has led to an enhancement in the performance of semantic segmentation. This is evidenced by the improved mIoU of 68.78% and 69.52%, notably within the category of "Imp. Surf" and "building". It is reasonably concluded that this comprehensive hybrid architecture is capable of enhancing the distinction between classes and consequently improving the segmentation effect. To facilitate a more detailed comparison of the impact of different hybrid methods, STUNet and CMTFNet are replicated and achieve corresponding improvements. While the aforementioned four methods (UNetFormer, SRCBTFusion, STUnet, and CMTFNet) demonstrate enhanced segmentation outcomes of hybrid architecture, the segmentation results of STUNet and CMTFNet employing the two-branch stream are more optimal, reaching 69.67% and 71.21% mIoU, respectively. In particular, CMTFNet demonstrates the most effective segmentation performance in the categories of "Building", "Low. Veg.", and "Car", with respective mIoU of 86.52%, 70.07%, and 75.07%. To illustrate the advantages of frequency-domain guidance, a comparison between MsanlfNet, SFFNet, MIFNet and the proposed FGNet is presented. The results demonstrate that the mean Intersection over Union (mIoU) for the three methods compared are 70.95%, 65.47%, and 70.36%, respectively, all of which are inferior to the performance of the proposed method. This superiority can be attributed to the novel integration of frequency domain information within our approach. By simultaneously modeling both spatial and frequency domain features, our method significantly enhances the feature representation capabilities for remote sensing images. In contrast, the coupling mechanisms employed by the other methods fall short, particularly in effectively distinguishing between various feature types. This underscores the advantages of our method in leveraging frequency domain information and highlights its superior effectiveness compared to alternative coupling strategies.

Moreover, the proposed FGNet has been shown to perform the best segmentation result through a qualitative analysis of multiple images. As illustrated in the Fig. 6, the yellow box represents the most significant improvement of the proposed method. It is obvious that FGNet exhibits a distinct boundary, high class consistency, and an effective segmentation result.

E. Comparison with SOTA Methods on Vaihingen Datasets

The experimental results on Vaihingen Dataset are shown in Table II and Fig. 7. Similar to the findings on Potsdam Dataset, it is observed that methods alone based on CNNs or transformers exhibit relatively low accuracy. UNetFormer and SRCBTFusion, enhance segmentation performance with the cascade hybrid architecture, achieving mIoU of 71.76% and 72.98%, respectively. Furthermore, the two-branch hybrid

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON POTSDAM DATASET. THE BOLD FONT INDICATES THE BEST DATA, WHILE THE UNDERLINED DENOTES THE SECOND-BEST DATA.

Methods	IoU/%						mIoU/%	OA/%	F1/%	Precision/%	Recall/%
	Imp. Suf.	Building	Low. Veg.	Tree	Car	Clutter					
FCN [41]	75.88	84.22	65.48	64.84	72.13	42.37	67.49	83.70	79.81	80.01	79.88
U-net [42]	76.73	84.81	67.52	66.89	73.50	42.81	68.71	84.49	80.68	82.00	79.74
SwinB-UperNet [17]	73.87	81.48	65.64	62.25	69.36	42.70	65.88	82.57	78.75	79.39	78.44
SegFormer [13]	73.47	80.88	63.69	60.86	69.04	40.07	64.67	81.74	77.75	78.01	77.66
UNetFormer [18]	77.42	84.98	67.05	64.55	73.81	44.90	68.78	84.44	80.80	81.92	80.04
SRCBTFusion [21]	78.11	85.96	68.73	65.62	71.73	46.98	69.52	85.24	81.59	81.50	81.79
STUNet [22]	77.64	86.14	68.10	66.63	72.47	47.03	69.67	84.24	80.67	80.84	80.66
CMTFNet [19]	78.74	86.52	70.07	**68.07**	75.07	48.78	71.21	85.95	82.60	**83.97**	81.51
MsanlfNet [31]	78.94	**87.34**	69.50	67.28	74.19	48.44	70.95	84.86	82.37	83.26	81.61
SFFNet [28]	74.45	81.49	64.61	59.24	69.88	43.16	65.47	82.26	78.44	79.77	77.81
MIFNet [32]	78.66	86.60	69.42	66.22	73.85	47.42	70.36	85.58	81.97	83.03	81.19
Ours (FGNet)	**80.09**	85.93	**70.16**	67.81	**75.64**	**49.14**	**71.46**	**86.10**	**82.78**	83.63	**82.09**

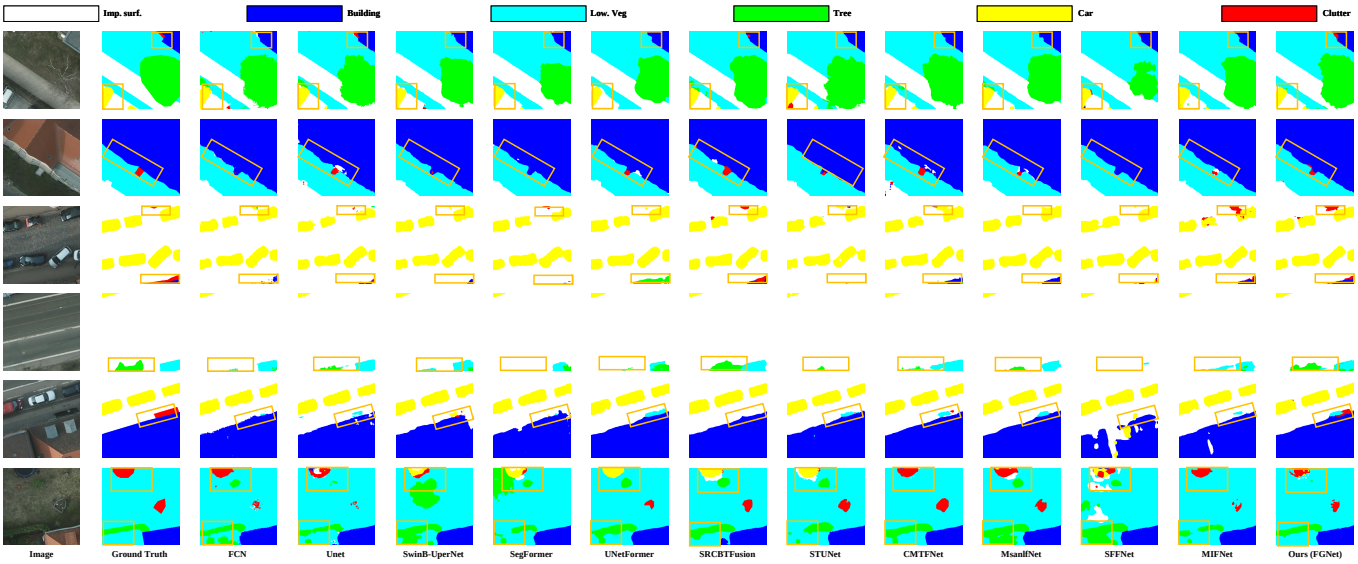

Fig. 6. The Visualization results of Potsdam Datasets.

architectures CMTFNet and STUNet also show better results than the cascade architecture, and obtain more optimal segmentation results, with mIoU reaching 70.45% and 73.21%. Specifically, CMTFNet demonstrates superior segmentation performence in the categories of "Imp. Surf", "Low. Veg.", and "Tree" are improved to 77.78%, 59.48%, and 73.21%, respectively. Similarly, MsanlfNet, SFFNet and MIFNet, which incorporate frequency-domain guidance within CNNs methods enhances segmentation performance, with achieving an mIoU of 73.21%, 70.12%, and 72.99%, respectively. The proposed FGNet achieves the highest segmentation performance again. The frequency guidance mechanism and the feature fusion module optimize feature quality, leading to superior segmentation results. It has also been demonstrated that the utilisation of the frequency domain can enhance the results of semantic segmentation. Moreover, the proposed FGNet achieves the highest segmentation performance again. The frequency guidance mechanism and the feature fusion module optimize feature quality, leading to superior segmentation results. The comparative analysis between MsanlfNet and FGNet confirms that the proposed coupling approach is more effective, highlighting the efficacy of the designed coupling method in improving feature effectiveness and class separability.

Similarly, through qualitative analysis as shown in Fig. 7, the proposed method FGNet also achieves the best segmentation performance on Vaihingen Dataset. The yellow boxes highlight the most significant improvements. The FGNet demonstrates superior segmentation results with distinct boundaries and high class consistency.

A further examination of the model efficiency reveals that the proposed method exhibits the lowest for parameters (params) and the floating point of operations (FLOPs), as is shown in Table III. The rationale behind this lies in the fact that the effective parameters of the CNNs branch structure are minimal, and the incorporation of group convolution further reduces the parameters of the FGSwin. Additionally, GLFI is characterized by a minimalist approach, with small parameters and a simple structure. The aforementioned design elements contribute to the overall lightweight construction of the FGNet.

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON VAIHINGEN DATASET. THE BOLD INDICATES THE BEST DATA, WHILE THE UNDERLINED DENOTES THE SECOND-BEST DATA.

Methods	IoU/%						mIoU/%	OA/%	F1/%	Precision/%	Recall/%
	Imp. Suf.	Building	Low. Veg.	Tree	Car	Clutter					
FCN [41]	77.74	84.23	58.99	71.87	49.91	84.54	71.21	85.04	82.49	84.13	81.12
U-net [42]	75.43	83.65	55.70	71.62	57.11	83.33	71.14	85.10	82.74	84.68	81.45
SwinB-UperNet [17]	77.21	82.65	59.44	**73.74**	50.22	86.74	71.66	85.16	82.81	84.65	81.28
SegFormer [13]	77.20	82.81	58.30	72.25	48.04	89.21	71.29	84.76	82.41	84.87	80.58
UnetFormer [18]	77.23	82.48	57.87	70.01	51.12	**91.82**	71.76	84.30	82.77	85.07	81.00
SRCBFusion [21]	77.24	83.99	58.27	72.53	55.80	90.07	72.98	85.15	83.76	85.06	82.70
STUNet [22]	75.11	81.64	56.39	70.36	49.72	89.51	70.45	83.58	81.88	84.29	79.96
CMTFNet [19]	77.78	83.97	59.48	72.52	52.17	89.60	72.59	85.31	83.42	85.22	82.04
MsanlfNet [31]	**79.63**	**85.36**	**60.72**	73.21	55.10	85.24	73.21	**86.06**	83.99	84.74	83.44
SFFNet [28]	76.50	82.50	57.75	71.05	44.48	88.43	70.12	84.22	81.46	83.11	80.17
MIFNet [32]	78.67	84.33	59.35	73.21	56.94	84.52	72.99	85.17	83.75	85.33	82.55
Ours (FGNet)	77.67	84.35	59.58	73.51	**57.39**	91.73	**74.04**	85.62	**84.49**	**85.65**	**83.51**

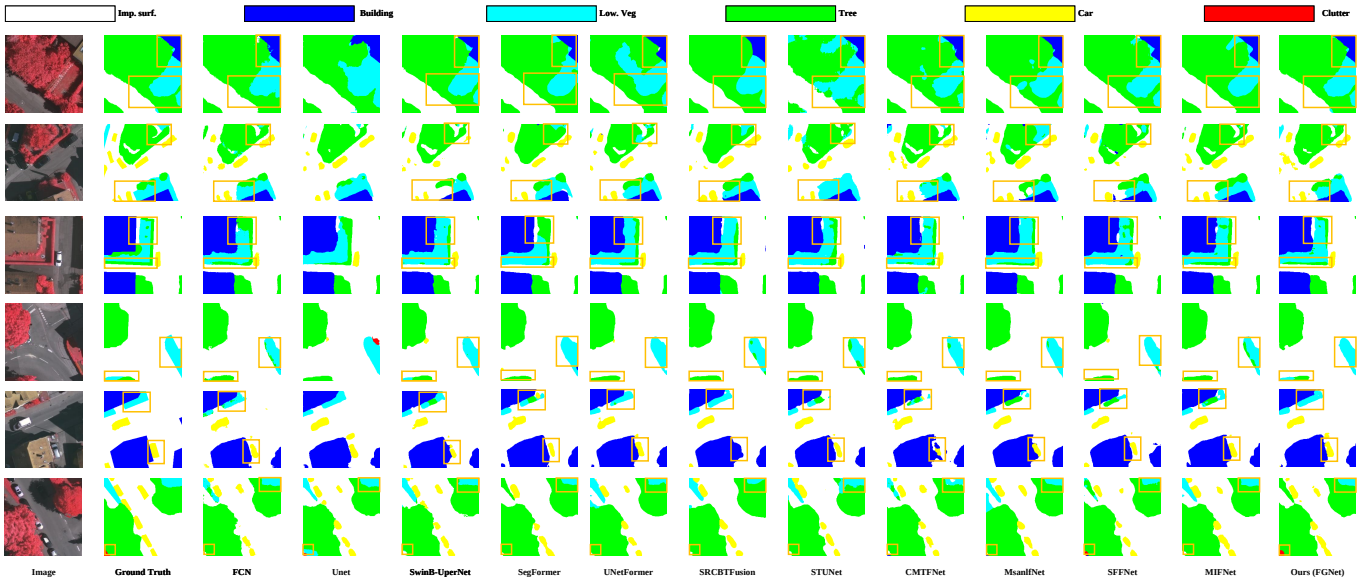

Fig. 7. The Visualization results of the Vaihingen Datasets.

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS OF PARAMS AND FLOPS

Methods	Parameters/M	FLOPS/G
FCN [41]	18.64	430.12
U-net [42]	13.40	249.12
SwinB-UperNet [17]	33.24	430.12
SegFormer [13]	26.52	32.24
UNetFormer [18]	11.68	23.48
SRCBTFusion [21]	74.41	256.63
STUNet [22]	325.35	157.50
CMTFNet [19]	30.07	68.56
MsanlfNet [31]	32.24	55.62
SFFNet [28]	103.35	34.16
MIFNet [32]	70.84	25.35
Ours (FGNet)	**8.08**	**8.02**

F. Ablation Study

Take Vaihingen Dataset as example, in this part, ablation study of the proposed method is constructed. The specifically experimental results are shown in Table IV and Fig. 8. In these experiments, "Swin + CNNs" refers to a dual-branch structure where the upper branch utilizes the Swin transformer as the feature extraction model, and the lower branch employs the designed residual CNNs model. The segmentation results are obtained by directly using the up-sampling and stacking approach. The "FGSwin + CNNs" method couples the frequency domain guidance mechanism to the feature extraction of the Swin converter under the same dual-branch architecture. This configuration demonstrates the impact of the devised frequency-domain guidance mechanism. The "FGSwin + CNNs + GLFI" denotes aforementioned cube with GLFI module, which is subjected to further analysis to ascertain its efficiency.

Through the quantitative results in Table IV and the segmentation results in Fig. 8, it can be observed that the method using only "Swin transformer + CNNs" achieved relatively low segmentation performance, with mIoU, OA and F1-score of 70.41%, 87.50% and 69.69%, respectively. The proposed frequency domain guidance mechanism, coupling strategy,

TABLE IV
ABLATION EXPERIMENTS OF THE PROPOSED MODULE ON THE VAIHINGEN DATASET. THE BOLD INDICATES THE BEST DATA.

Methods	IoU/%						mIoU/%	OA/%	F1/%
	Imp. Suf.	Building	Low. Veg.	Tree	Car	Clutter			
Swin + CNNs	76.80	83.24	59.42	**73.97**	56.09	72.97	70.41	85.07	82.26
FGSwin + CNNs	**77.81**	83.74	58.18	72.39	55.72	89.87	72.95	85.11	83.74
FGSwin + CNNs + GLFI (FGNet)	77.67	**84.35**	**59.58**	73.51	**57.39**	**91.73**	**74.04**	**85.62**	**84.49**

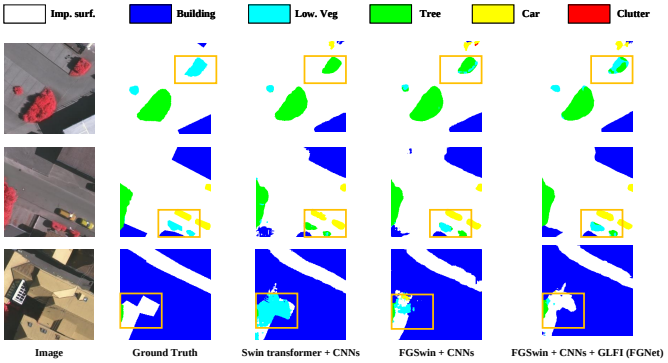

Fig. 8. The Visualization results of ablation study.

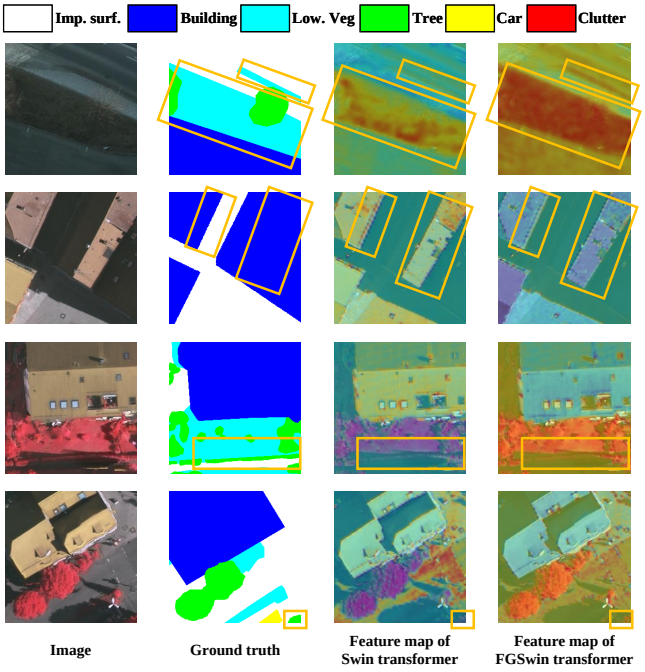

Fig. 9. The visualization of mid-feature maps of Swin transformer and FGSwin transformer.

and GLFI module demonstrate significant advantages and effectiveness. Specifically, "FGSwin + CNNs" obtain 72.95% mIoU, 94.04% OA and 76.23% F1-score, which has an obvious increasing of about 2.5% mIoU , 6.5% OA and 6.6% F1-score. We contend that the process of feature extraction is enhanced by the application of frequency-domain guidance. Furthermore, the "FGSwin + CNNs + GLFI" approach results in an additional increase of over 1% in mIoU, reaching a total of 74.04%. The GLFI module is also capable of enhancing classification separability through the utilization of a small-scale convolutional kernel and SiLU. Also, through qualitative analysis as shown in Fig. 8, the yellow boxes highlight the most significant improvements. The proposed FGNet exhibits superior segmentation results with higher class consistency and discernibility. These collectively contributed to a notable enhancement in performance, as evidenced by the quantitative and qualitative visual analysis.

IV. DISCUSSION

To further analyze the performance of our proposed method, we visualize the feature maps of the FGSwin transformer alongside the baseline Swin transformer, as shown in Fig. 9. The visualization enables an in-depth exploration of the mechanisms that enhance segmentation accuracy, especially in contexts involving edge and the analysis of homogeneity and heterogeneity within images.

The results demonstrate that the frequency-domain guided FGSwin transformer significantly enhances boundary sharpness and preserves intricate details. In contrast, the Swin transformer feature maps exhibit a lower level of clarity, suggesting challenges in capturing fine-grained details without explicit frequency-based enhancements. This limitation may result in inaccuracies, especially in areas with complex textures or abrupt transitions between classes. For example, in the first row, the FGSwin transformer demonstrates enhanced intra-class consistency within the central 'Low Veg.' category, effectively maintaining uniformity across this region. These improvements underscore the ability of proposed FGSwin to leverage frequency-domain information for superior edge preservation and more cohesive feature representation across complex textures. It enhances the ability of our network to capture and integrate frequency-domain information during the feature extraction process. This enhancement facilitates improved handling of both global and local details, which demonstrates the advantages of our frequency-guided approach in complex remote sensing scenarios.

V. CONCLUSION

This paper proposes a lightweight semantic segmentation framework based on a dual-branch hybrid architecture. By incorporating a frequency-domain enhancement mechanism into the feature extraction framework, it is possible to ensure comprehensive expression of global-local, frequency-domain,

and spatial-domain information. Furthermore, a residual convolution network and a feature fusion module utilizing small-scale convolution kernels have been devised, which serve to enhance the capacity for semantic feature extraction and improve the classification separability and segmentation effect. Two publicly available datasets, Potsdam and Vaihingen, are employed in the experiments, and a comparison of the current SOTA semantic segmentation methods demonstrates that the proposed method exhibits excellent segmentation performance.

REFERENCES

[1] Z. Cai, Q. Hu, X. Zhang, J. Yang, H. Wei, J. Wang, Y. Zeng, G. Yin, W. Li, L. You, B. Xu, and Z. Shi, "Improving agricultural field parcel delineation with a dual branch spatiotemporal fusion network by integrating multimodal satellite data," *ISPRS J. Photogramm. Remote Sens.*, vol. 205, pp. 34–49, 2023.

[2] X. Huang, W. Wang, J. Li, L. Wang, and X. Xie, "A stepwise refining image-level weakly supervised semantic segmentation method for detecting exposed surface for buildings (esb) from very high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–17, 2024.

[3] E. Zhang, H. Zong, X. Li, M. Feng, and J. Ren, "Icsf: Integrating inter-modal and cross-modal learning framework for self-supervised heterogeneous change detection," *IEEE Trans. Geosci. Remote Sens.*, 2024.

[4] Y. He, J. Wang, Y. Zhang, and C. Liao, "An efficient urban flood mapping framework towards disaster response driven by weakly supervised semantic segmentation with decoupled training samples," *ISPRS J. Photogramm. Remote Sens.*, vol. 207, pp. 338–358, 2024.

[5] M. Pastorino, G. Moser, S. B. Serpico, and J. Zerubia, "Semantic segmentation of remote-sensing images through fully convolutional neural networks and hierarchical probabilistic graphical models," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[6] L. Liu, Z. Tong, Z. Cai, H. Wu, R. Zhang, A. Le Bris, and A.-M. Olteanu-Raimond, "Hieru-net: A hierarchical semantic segmentation method for land cover mapping," *IEEE Trans. Geosci. Remote Sens.*, 2024.

[7] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 6230–6239.

[8] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 6392–6401.

[9] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 4095–4104.

[10] J. Wang, Z. Feng, Y. Jiang, S. Yang, and H. Meng, "Orientation attention network for semantic segmentation of remote sensing images," *Knowl. Based Syst.*, vol. 267, p. 110415, 2023.

[11] J. Liu, W. Hua, W. Zhang, F. Liu, and L. Xiao, "Stair fusion network with context-refined attention for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–17, 2024.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Int. Conf. Learn. Represent.*, 2021.

[13] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[14] R. Xu, C. Wang, J. Zhang, S. Xu, W. Meng, and X. Zhang, "Rssformer: Foreground saliency enhancement for remote sensing land-cover segmentation," *IEEE Trans. Image Process.*, vol. 32, pp. 1052–1064, 2023.

[15] C. You, L. Jiao, X. Liu, L. Li, F. Liu, W. Ma, and S. Yang, "Boundary-aware multiscale learning perception for remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

[16] Z. Xu, J. Geng, and W. Jiang, "Mmt: Mixed-mask transformer for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

[17] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 9992–10 002.

[18] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 190, pp. 196–214, 2022.

[19] H. Wu, P. Huang, M. Zhang, W. Tang, and X. Yu, "Cmtfnet: Cnn and multiscale transformer fusion network for remote-sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023.

[20] S. Du and M. Liu, "Class-guidance network based on the pyramid vision transformer for efficient semantic segmentation of high-resolution remote sensing images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 5578–5589, 2023.

[21] J. Chen, J. Yi, A. Chen, and H. Lin, "Srcbtfusion-net: An efficient fusion architecture via stacked residual convolution blocks and transformer for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023.

[22] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding unet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[23] H. Wang, X. Li, L. Huo, and C. Hu, "Global and edge enhanced transformer for semantic segmentation of remote sensing," *Appl. Intell.*, pp. 1–16, 2024.

[24] L. Fan, Y. Zhou, H. Liu, Y. Li, and D. Cao, "Combining swin transformer with unet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–11, 2023.

[25] L. Chen, L. Gu, D. Zheng, and Y. Fu, "Frequency-adaptive dilated convolution for semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 3414–3425.

[26] X. Li, F. Xu, H. Gao, F. Liu, and X. Lyu, "A frequency domain feature-guided network for semantic segmentation of remote sensing images," *IEEE Signal Process. Lett.*, 2024.

[27] J. Liu, D. Zhang, L. He, X. Yu, and W. Han, "Mfagnet: Multi-scale frequency attention gating network for land cover classification," *Int. J. Remote Sens.*, vol. 44, no. 21, pp. 6670–6697, 2023.

[28] Y. Yang, G. Yuan, and J. Li, "Sffnet: A wavelet-based spatial and frequency domain fusion network for remote sensing segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–17, 2024.

[29] C. Yang and Z. Zhang, "Pfd-net: Pyramid fourier deformable network for medical image segmentation," *Comput. Biol. Med.*, vol. 172, p. 108302, 2024.

[30] X. Li, F. Xu, X. Yong, D. Chen, R. Xia, B. Ye, H. Gao, Z. Chen, and X. Lyu, "Sscnet: A spectrum-space collaborative network for semantic segmentation of remote sensing images," *Remote Sens.*, vol. 15, no. 23, p. 5610, 2023.

[31] L. Bai, X. Lin, Z. Ye, D. Xue, C. Yao, and M. Hui, "Msanlfnet: Semantic segmentation network with multiscale attention and nonlocal filters for high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[32] J. Fan, J. Li, Y. Liu, and F. Zhang, "Frequency-aware robust multidimensional information fusion framework for remote sensing image segmentation," *Eng. Appl. Artif. Intell.*, vol. 129, p. 107638, 2024.

[33] G. Wei, J. Xu, W. Yan, Q. Chong, H. Xing, and M. Ni, "Dual-domain fusion network based on wavelet frequency decomposition and fuzzy spatial constraint for remote sensing image segmentation," *Remote Sens.*, vol. 16, no. 19, p. 3594, 2024.

[34] Y. Li, Z. Liu, J. Yang, and H. Zhang, "Wavelet transform feature enhancement for semantic segmentation of remote sensing images," *Remote Sens.*, vol. 15, no. 24, p. 5644, 2023.

[35] T. Zhang and R. P. Dick, "Spatial-frequency network for segmentation of remote sensing images," in *Proc. Int. Conf. Image Process (ICIP)*. IEEE, 2023, pp. 3553–3557.

[36] X. Zhou, F. Liang, L. Chen, H. Liu, Q. Song, G. Vivone, and J. Chanussot, "Mesam: Multiscale enhanced segment anything model for optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024.

[37] J. Li, S. Zhang, Y. Sun, Q. Han, Y. Sun, and Y. Wang, "Frequency-driven edge guidance network for semantic segmentation of remote sensing images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 9677–9693, 2024.

[38] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016.

[39] L. Fang, P. Zhou, X. Liu, P. Ghamisi, and S. Chen, "Context enhancing representation for semantic segmentation in remote sensing images," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 35, no. 3, pp. 4138–4152, 2024.

[40] X. Li, F. Xu, F. Liu, Y. Tong, X. Lyu, and J. Zhou, "Semantic segmentation of remote sensing images by interactive representation refinement and geometric prior-guided inference," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–18, 2024.

[41] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3431–3440.

[42] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.