

GFFNet: Global Feature Fusion Network for Semantic Segmentation of Large-scale Remote Sensing Images

Yong Cao, Chunlei Huo, *Member, IEEE*, Shiming Xiang, *Member, IEEE*, Chunhong Pan, *Member, IEEE*

Abstract—Semantic segmentation plays a pivotal role in interpreting high-resolution remote sensing images (RSIs), where contextual information is essential for achieving accurate segmentation. Despite the common practice of partitioning large RSIs into smaller patches for deep model input, existing methods often rely on adaptations from natural image semantic segmentation techniques, limiting their contextual scope to individual images. To address this limitation and harness a broader range of contextual information from original large-scale RSIs, this study introduces a global feature fusion network (GFFNet). GFFNet employs a novel approach by incorporating a group Transformer structure alternated with group convolution, forming a lightweight global context learning branch. This design facilitates the extraction of global contextual features from the large-scale RSIs. Additionally, we propose a cross feature fusion (CFF) module that seamlessly integrates local features obtained from the convolutional network with the global contextual features. GFFNet serves as a versatile plugin for existing remote sensing image semantic segmentation models, particularly beneficial when the target dataset involves cropping. This integration enhances the model's performance, especially in terms of segmenting large-scale objects. Experimental results on the ISPRS and GID-15 datasets validate the effectiveness of GFFNet in improving segmentation capabilities for large-scale objects in remote sensing images.

Index Terms—Global context learning; cross feature fusion; group Transformer; semantic segmentation.

I. INTRODUCTION

SEMANTIC segmentation of remote sensing images (RSIs) involves labeling each pixel in the image according to its corresponding class, enabling the recognition of land cover types and their spatial distribution. Semantic segmentation of RSIs holds extensive application value in various fields, including urban planning [1], road extraction [2], and land cover classification [3]. With the rapid growth of high-resolution (HR) RSI data, effectively processing these large-scale, high-dimensional, and complex data while enhancing the accuracy and efficiency of semantic segmentation has emerged as a significant challenge in current research [4].

Convolutional neural networks (CNNs) are currently the mainstream method for semantic segmentation of RSIs. They

can automatically learn image features and perform feature fusion at multiple scales to improve classification performance. However, original CNNs have several limitations, such as difficulty in capturing large-scale contextual information, neglecting long-range dependencies, and sensitivity to rotation and scale changes. To address these issues, subsequent research has proposed numerous improvement strategies, including utilizing contextual information [5], employing dilated convolution [6]–[8], and incorporating attention mechanisms [9], [10] to enhance the ability of CNN-based models to effectively capture contextual information.

While CNNs have made breakthroughs in the field of semantic segmentation of RSIs, they are limited in their ability to capture global contextual information. Although improvements such as the use of contextual information, atrous convolution, and attention mechanisms have helped alleviate this limitation to some extent, they have not fundamentally resolved it, as CNNs are inherently locality-based models. In contrast, Transformers [11] have shown remarkable success in capturing global contextual information in natural language processing (NLP). Researchers have started applying these models to computer vision, achieving significant progress in image recognition, detection, and segmentation. Recently, visual foundation models like segment anything model (SAM) [12] opens new horizons for intelligent image analysis and interpretation. Trained on over a billion masks, SAM showcases remarkable performance in zero-shot segmentation. But SAM's potential is somewhat constrained in specialized fields like remote sensing, primarily due to its heavy reliance on prior manual guidance. Addressing this, RSPrompter [13] presents a novel, automated instance segmentation approach tailored for remote sensing images, leveraging the capabilities of the SAM model. However, despite the success in various computer vision tasks, these Transformer-based models have not yet demonstrated a clear advantage over traditional CNN-based models in the specific area of semantic segmentation of RSIs. This is attributed to the fact that Transformer-based models lack the inherent inductive bias of CNNs, such as translation equivariance. Consequently, training these models effectively requires a larger amount of annotated data. Unfortunately, there exists a considerable gap in the scale of labeled remote sensing image datasets when compared to natural image datasets. Furthermore, RSIs possess unique characteristics that distinguish them from natural images, such as the presence of semantically complex objects like buildings, water bodies, vegetation, and roads. The boundaries between these objects

Yong Cao, Chunlei Huo (Corresponding Author), Shiming Xiang, Chunhong Pan are with National Key Laboratory for Multi-modal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences and School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: {yong.cao, clhuo, smxiang, chpan}@nlpr.ia.ac.cn). This research was supported by the National Key Research and Development Program of China under Grant No. 2018AAA0100400, the National Natural Science Foundation of China under Grant No. 62071466, No. 62076242 and No. 62306310.

are often intricate, necessitating the fusion of both local and global contextual information to achieve improved segmentation results.

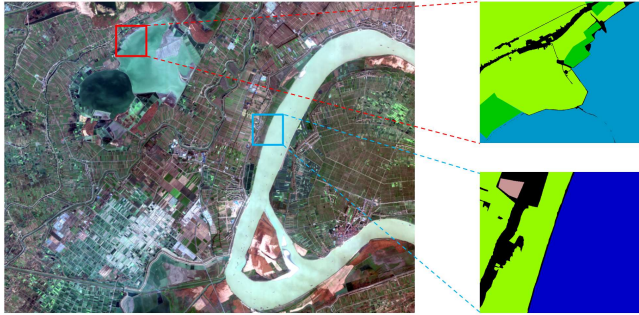


Fig. 1. This is a large-scale RSI, in which the red and blue boxes are the small patches actually sent to the network during the training process. It can be seen that it is difficult to determine whether the water belongs to a river or a lake in the absence of larger contextual features.

Moreover, unlike natural images, RSIs are typically very large in size. For instance, a single image from the GID-15 dataset [14] can reach dimensions of 6800×7200 pixels. Such images often need to be divided into smaller patches, such as 480×480 pixels, to be compatible with GPUs during training. However, this cropping can lead to the inclusion of the same object in multiple small images. Unfortunately, existing context feature extraction methods, whether based on CNNs or Transformers, typically focus on individual cropped patches, inevitably resulting in information loss, as depicted in Figure 1.

To address this issue and extract comprehensive context information, we propose a global feature fusion network (GFFNet) that employs a novel group Transformer to extract global context information from the entire large image. Given the limited volume of annotated data, a CNN-based model is used as a local feature extractor, capitalizing on its unique inductive bias. Furthermore, we introduce a cross feature fusion (CFF) module to dynamically fuse the two types of features. Our main innovations can be summarized as follows:

(1) Introduction of a group Transformer structure. By grouping feature maps, this structure effectively reduces the computational complexity of the Transformer block while retaining positional information. It maintains the global modeling capability of the original Transformer.

(2) Development of a lightweight global feature fusion network by alternating group convolution with the proposed group Transformer. This network can seamlessly integrate into existing remote sensing image semantic segmentation models, enabling the extraction of rich global semantic information from the original RSIs.

(3) Design of a CFF module that effectively merges local spatial features with global context features. This empowers the model to capture contextual information from the entire large image, thereby achieving genuine global context modeling.

(4) Extensive experiments on the ISPRS [15] and GID-15 datasets showcase the effectiveness of the proposed GFFNet and CFF module.

II. RELATED WORK

Contextual information, encompassing spatial, semantic, and global context, plays a crucial role in achieving accurate segmentation and is highly significant in semantic segmentation. Deep learning-based models for semantic segmentation often prioritize the extraction of comprehensive contextual information. In this section, we review notable contributions that have explored context information extraction in semantic segmentation.

CNN-based context extraction: While the original CNN can aggregate contextual information within the size of the convolution kernel in each layer, such as 3×3 or 5×5 , stacking layers alone proves insufficient for semantic segmentation to obtain an expanded contextual scope. The DeepLab series [6]–[8] introduced dilated convolution, enabling the incorporation of a wider contextual area. Dilated convolution involves introducing different hole sizes within the convolution kernel, producing feature maps with varied receptive field sizes. This innovation leads to enhanced accuracy for semantic segmentation. Additionally, the DeepLab series introduced global pooling operations to extract contextual information from the entire image, further enhancing segmentation performance. The pyramid scene parsing network (PSPNet) [5] introduced feature pyramid pooling, generating multiple feature maps through pooling operations of different scales applied to the input image. These feature maps are convolved with distinct pooling kernels, upsampled, and concatenated to form the result of feature pyramid pooling. This approach effectively employs feature maps of diverse scales to capture varying contextual ranges and, by weighting these scale feature maps, optimizes the model's utilization of different scale information. The dual attention network (DANet) [9] employs a dual attention mechanism for improved contextual perception. DANet integrates two attention mechanisms on the feature map: channel attention, which emphasizes essential features by weighting different channels, and spatial attention, which enhances spatial information perception by weighting features from distinct positions. Through the amalgamation of channel and spatial attention, DANet enhances contextual information utilization and consequently, the semantic segmentation performance.

Transformer-based context extraction: Comparatively, Transformers excel in handling long-range dependencies, unlike CNNs that execute local convolution operations. CNN's limited receptive field prevents modeling of extensive global dependencies between pixels due to the local nature of convolution and weight sharing. Conversely, Transformers utilize self-attention mechanisms for learning pixel dependencies. This mechanism allows the model to compute attention weights for each element in a sequence, such as image pixels, to comprehend their interactions with other elements. The self-attention principle of Transformers permits interactions between all positions, fostering global dependency modeling among pixels. The inception of the vision Transformer (ViT) [16] marked the application of a pure Transformer to visual domains. Segmentation Transformer (SETR) [17] extended this concept to semantic segmentation, treating it as a sequence-to-

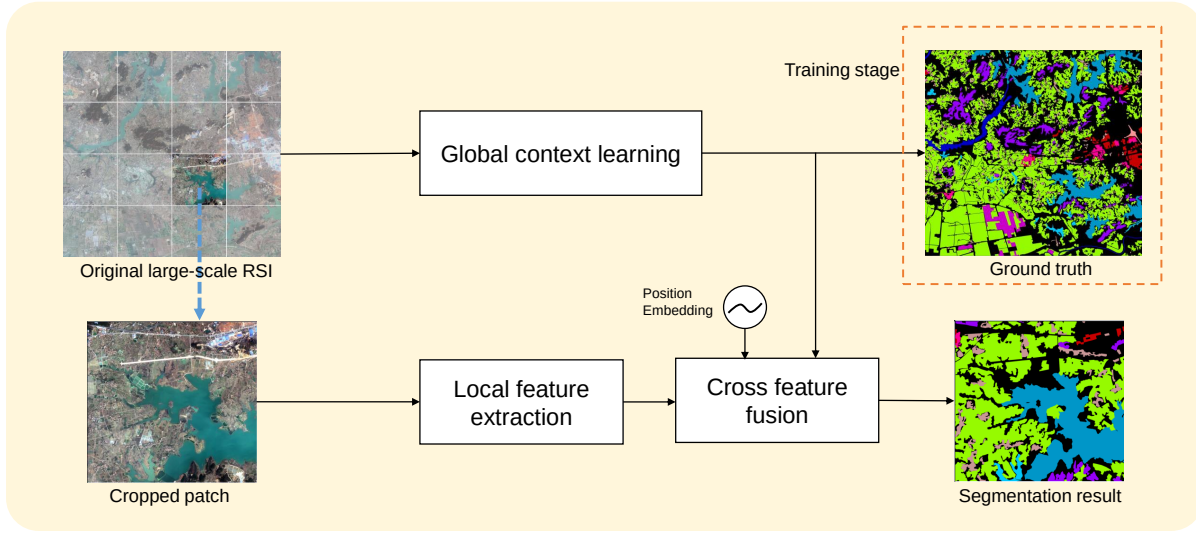


Fig. 2. Overview of GFFNet: first passing the original large image and the cropped small patch through the global contextual learning branch and local feature extraction branch, respectively, then the cross feature fusion module is used to combine them and obtain the final segmentation result.

sequence prediction task. SETR encodes images as sequences of patches using a pure Transformer and subsequently decodes output embeddings to generate segmentation masks. Segmenter [18] expands on ViT, integrating point-wise linear or mask Transformer decoders. TransUNet [19] introduces an innovative architecture combining a Transformer encoder and a UNet-like decoder for medical image segmentation. TransUNet employs a pre-trained ViT model as the encoder and adapts it for segmentation through additional linear projection and patch expansion layers.

Context extraction in RSIs: Compared to natural images, RSIs contain a wider range of diverse and complex targets, which places a greater demand on the need for contextual information. In the realm of semantic segmentation for RSIs, researchers have extensively explored methods to harness contextual information more effectively. ScasNet [20] introduces a self-cascaded network architecture that sequentially aggregates global-to-local contexts, enhancing labeling coherence for fine-structured and man-made objects. HRCNet [21] employs a lightweight dual attention module to extract global context information during feature extraction. It incorporates a feature enhancement feature pyramid structure to fuse contextual information of varying scales. Context aggregation networks [22] leverage context fuse, attention mix, and residual convolutional modules to manage multi-scale object recognition and accurate object localization. Adaptive effective receptive field convolution [23] proposes a method to control the sampling location of convolution adaptively, automating effective receptive field adjustments without incurring substantial parameter or computational cost increases. The emergence of Transformers has prompted considerable research on combining them with CNNs for RSIs semantic segmentation [24]–[26]. UNetFormer [27] proposes a Transformer-based decoder to capture the global context and construct an UNet-like Transformer for real-time urban scene segmentation. RSSFormer [28] presents an adaptive Transformer fusion module that can adaptively

suppress background noise and enhance object saliency when fusing multi-scale features. However, these efforts primarily focus on small cropped images, which inevitably lead to information loss. Conversely, our proposed method genuinely incorporates global contextual information, presenting a more comprehensive solution.

III. METHODOLOGY

In this section, we introduce our proposed GFFNet, outlined in Figure 2. GFFNet comprises a global context learning branch utilizing a lightweight Transformer to extract contextual information from the original, uncropped image before cropping. Additionally, a local feature extraction branch composed of CNNs is employed to extract spatial features from the cropped small patches. Finally, we integrate global contextual information with local spatial features through a CFF module. Detailed explanations of each component are provided below.

A. Global context learning

As highlighted earlier, contextual information holds paramount importance for image semantic segmentation, particularly in the realm of RSIs. For a given training set: $\mathcal{D} = \{I_1, I_2 \dots I_l\}$, $I_l \in \mathbb{R}^{c \times h_l \times w_l}$ where c, h_l, w_l representing spectral bands and spatial dimensions of each image. The dimensions h_l and w_l are typically smaller (within 1000) for natural images, allowing for direct GPU training. Thus, any two images I_m and I_n within dataset $\mathcal{D}$ are contextually independent. However, this assumption doesn't hold true for remote sensing images. This is because remote sensing images are typically collected and annotated based on regions, which leads to significantly larger dimensions of the original dataset images (e.g., in the GID-15 dataset, $h_l = 7200, w_l = 6800$). These large-sized images cannot be directly used for training and must be pre-cropped to obtain a smaller training dataset $\mathcal{D}_p = \{I_1^p, I_2^p \dots I_l^p\}$, that can be used for training purposes. It is evident that the small patches within $\mathcal{D}_p$ don't possess mutual

independence in terms of their contexts. The cropping process might divide the same object into multiple image blocks, while complex objects may demand a broader contextual information range for accurate land cover determination. However, existing context extraction methods, whether based on convolution (like dilated convolution or pyramid pooling) or Transformers, have been primarily developed for natural images and overlook inter-image contextual dependencies. Hence, it's crucial to devise dedicated network architectures tailored for the cropping-training process of remote sensing images. Such architectures should encompass a wider contextual information range to enhance semantic segmentation performance of RSIs.

Group Transformer The primary challenge in modeling and extracting contextual features from the entire large-scale RSI lies in the computational complexity of the model. The dimensions of original RSIs are often over 10 times larger than those of natural images. Applying conventional CNN or Transformer structures would result in a significant increase in computational requirements and the number of model parameters. This elevated computational complexity poses substantial challenges for effective model training and deployment. MobileNet [29] is a notable advancement in CNNs, with its core innovation being the introduction of depth-wise separable convolution—a two-step process. First, it utilizes depth-wise convolution with a $group = 1$ setting to extract spatial features. Then, point-wise convolution with a kernel size of 1 combines features from different channels. This design significantly reduces the computational complexity of the model while retaining its capacity to capture meaningful features. Nonetheless, due to the inherent characteristics of CNNs, the receptive field of such structures is constrained by the convolutional kernel size, limiting their ability to capture global contextual information. Drawing inspiration from group convolution, we incorporate this idea into the Transformer architecture, resulting in the creation of a group Transformer structure. This novel approach effectively reduces computational complexity while preserving the Transformer's proficiency in extracting global contextual information.

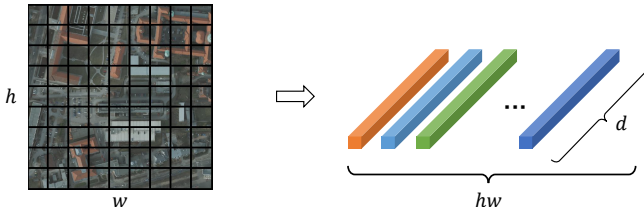


Fig. 3. Traditional visual feature serialization method, it treats each vector on the length and width of the feature map as an independent feature and feeds them into the Transformer.

As illustrated in Figure 3, the conventional Transformer structure usually involves flattening the input feature map, with dimensions $h \times w \times d$, into a sequence of length hw . This sequence is then processed through self-attention operations. However, this approach incurs a computational complexity of $\mathcal{O}((hw)^2d)$. For large-sized feature maps, this leads to substantial computational demands. Moreover, this structure disrupts the spatial organization of the original feature map,

causing loss of image-specific inductive biases and often necessitating more data for effective model training.

The proposed approach adeptly maintains a global receptive field, even while performing self-attention operations within each group. This is accomplished by subsequently integrating information from distinct groups through convolutional operations. Moreover, the substantial reduction in sequence length significantly decreases the computational complexity of self-attention to $\mathcal{O}((\frac{hw}{k^2})^2d) \times k^2 = \mathcal{O}(\frac{hw}{k^2}d)$. Crucially, this innovative structure conserves the spatial arrangement of the feature map and image-specific inductive biases, thereby mitigating the need for an extensive training dataset. Striking a balance between computational efficiency and the retention of crucial image context renders this approach highly effective for remote sensing image analysis.

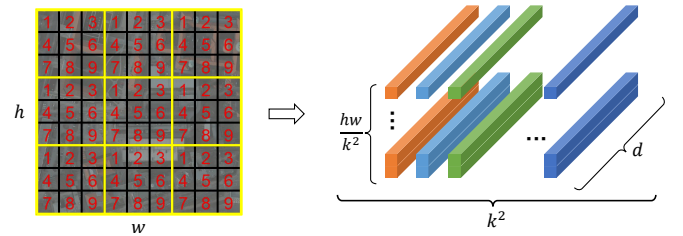


Fig. 4. Our proposed group Transformer, it divides the feature map into non-overlapping windows. Within each window, features at the same position are grouped together, and self-attention is only applied within each group. Information exchange between different groups is achieved through subsequent convolutions. This approach reduces computational complexity while retaining the global feature extraction capability of the Transformer.

To address this challenge, we introduce the concept of group convolution to the Transformer, as shown in Figure 4. For a feature map with height h and width w , we first partition it into $\frac{h}{k} \times \frac{w}{k}$ small blocks of size $k \times k$, based on the convolutional kernel size. Subsequently, each feature vector within these small blocks is assigned a unique index $\{1, 2, \dots, k^2\}$. This process results in the creation of $k \times k$ groups, where feature vectors sharing the same index from different small blocks are grouped together. Within each group, self-attention operations are carried out independently. The resultant feature vectors are then reinstated to their original positions within the $h \times w$ feature map.

Network Architecture The network architecture of our proposed global context learning branch is depicted in Figure 5. Initially, both the original large image and its corresponding ground truth are down-sampled to $\frac{1}{4}$ of their original size. Subsequently, the down-sampled images undergo a convolutional layer with a 3×3 kernel and a stride of 2. The resulting feature maps are then processed through consecutive blocks comprising MobileNet (MV) and our novel group Transformer components. The MV block takes inspiration from the design of MobileNet V2 [30], while the group Transformer encompasses a multi-head self-attention mechanism and a convolutional layer. Unlike traditional Transformers, our approach employs the lightweight group Transformer, introduced earlier, to extract grouped global features. The subsequent convolutional layer fuses features from diverse groups, leading to the acquisition of authentic global features.

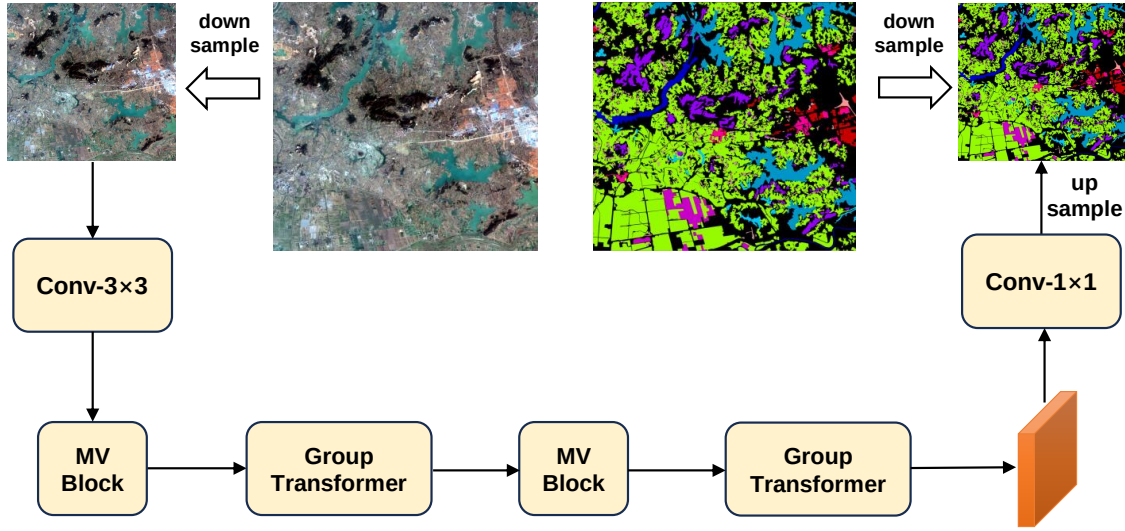


Fig. 5. The network structure of the global contextual learning branch, it consists of alternating mobileNet blocks and group Transformers, forming a lightweight feature extraction network. This network learns and captures the global contextual information of the large image.

Finally, the obtained global feature maps undergo a 1×1 convolution, followed by upsampling, and are compared to the ground truth using cross-entropy loss during training. It should be noted that, during the testing phase, the global feature extraction branch only needs to obtain the global features and does not require the final segmentation result of the original large RSIs.

B. Local feature extraction

Local Feature Extraction Branch: In the context of ViT's emergence, Transformers have displayed remarkable adaptability in the realm of computer vision, leading to numerous semantic segmentation models built exclusively upon Transformer architectures. It's important to note that these advancements heavily rely on extensive labeled datasets for training. In the domain of remote sensing image semantic segmentation, where data availability is often limited, Transformers have not exhibited significant advantages. Conversely, due to their ability to preserve image-specific inductive biases, CNNs demonstrate superior generalization capabilities. Thus, in our approach, we opt to utilize CNNs as the local feature extraction branch of GFFNet, recognizing their efficacy in addressing challenges posed by remote sensing data.

HRNet [31] stands as a prominently recognized backbone network within the CNN domain. Originally conceived for keypoint detection, HRNet has since excelled across various dense prediction tasks. Diverging from other semantic segmentation models—some of which either learn low-resolution representations from low-resolution feature maps (e.g., FCN [32], PSPNet) or recover high-resolution representations through skip connections (e.g., UNet [33], SegNet [34]), HRNet employs a unique approach. It retains a set of high-resolution representations while integrating low-resolution branches for multi-resolution fusion. This distinctive strategy empowers the final feature maps with both high-resolution fine-grained details and low-resolution semantic information. Given the

intricacies of land cover types in remote sensing images, we have strategically selected HRNet as the local feature extraction branch. It's vital to emphasize that our global context learning branch seamlessly integrates with any existing semantic segmentation model—a pivotal aspect we will elaborate on extensively within the experimental section.

C. Cross feature fusion

The cross-attention mechanism serves as an extension of the self-attention mechanism. Unlike the self-attention mechanism, which primarily captures correlations between different positions within the same input sequence, the cross-attention mechanism bolsters modeling accuracy by introducing an additional input sequence, thereby amalgamating information from two distinct sources. To achieve optimal fusion of global contextual information and local spatial features, we have conceived a cross feature fusion (CFF) module founded upon the cross-attention mechanism of the Transformer decoder, as depicted in Figure 6. Utilizing the global and local feature extraction branches, we derive the global context feature $F_g \in \mathbb{R}^{h_g \times w_g \times d_g}$ and the local feature $F_l \in \mathbb{R}^{h_l \times w_l \times d_l}$, respectively. Subsequently, both feature maps are serialized with a window size of $m \times m$. These serialized representations then undergo local feature embedding and global context embedding layers, followed by layer normalization. This yields $F'_g \in \mathbb{R}^{N_g \times d_t}$ and $F'_l \in \mathbb{R}^{N_l \times d_t}$, where N_g and N_l represent the lengths of sequence, and d_t is the dimension of the vectors. Subsequently, we utilize F'_l to create the query matrix and F'_g to generate the key and value matrices:

$$Q = W_l^q F'_l, \quad K = W_g^k F'_g, \quad V = W_g^v F'_g \quad (1)$$

These matrices are then subjected to L layers of multi-head attention, similar to the self-attention process:

$$F_f = \text{softmax} \left(\frac{Q_l K_g^T}{\sqrt{d_t}} \right) \quad (2)$$

The resulting vectors are reshaped to match the dimensions of F_l and concatenated together. Finally, the concatenated features undergo a 1×1 convolution followed by upsampling, resulting in the ultimate prediction result.

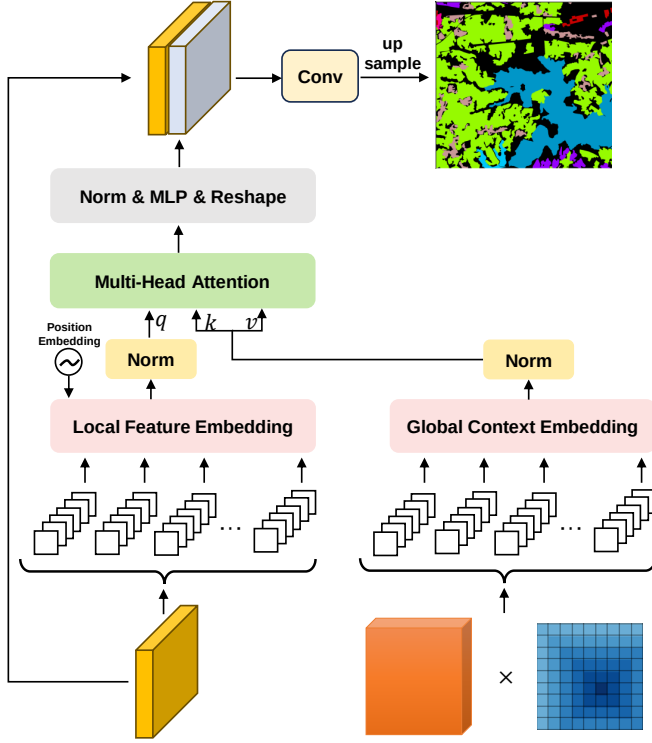


Fig. 6. Cross feature fusion module. It encodes the feature maps obtained from the global contextual learning branch and the local feature extraction branch into vectors containing positional information. Then, a cross-attention mechanism is applied for feature fusion. The resulting feature maps are restored to their original size and concatenated with the local features to obtain the final feature map. Finally, segmentation results are obtained through convolution and upsampling.

It's important to highlight that while learnable position encodings have been introduced to incorporate positional information into local features, the effective availability of training data for large images when dealing with global contextual features is relatively limited. We have observed that using learnable position encodings for global features often fails to capture meaningful positional information. In light of this, we have devised a position mask based on a two-dimensional Gaussian distribution:

$$f(X) = \frac{1}{(2\pi)^{1/2} |\Sigma^{-1}|} \exp \left[-\frac{1}{2} (X - \mu)^T \Sigma^{-1} (X - \mu) \right] \quad (3)$$

Here, X signifies the coordinates of the center position of small patches within the larger image. The purpose of this position mask is to accentuate the significance of different regions within the global context. Proximity to the image block results in higher weights, as these regions bear greater semantic relevance for segmenting the image block.

This cross feature fusion module, as an integral part of GFFNet, adeptly amalgamates global and local information. Consequently, the model's capacity to capture contextual in-

tricacies and spatial details for precise remote sensing image segmentation is significantly enhanced.

IV. EXPERIMENTS

We conducted extensive experiments to validate the effectiveness of the proposed GFFNet. This section provides detailed information about the experiments, including an introduction to the datasets and evaluation metrics, network parameter settings, and hyperparameter selection. We also present comparisons with results from other models, numerous ablation experiments, and showcase some segmentation results.

A. Datasets and settings

The ISPRS 2D semantic labeling dataset is one of the most widely used benchmarks in the field of semantic segmentation of RSIs. It consists of two subsets: Vaihingen and Potsdam. The dataset includes very high-resolution true ortho photo (TOP) tiles and corresponding digital surface models (DSMs). Each subset contains six land cover classes: Impervious surfaces, Building, Low vegetation, Tree, Car, and Clutter. The Vaihingen dataset comprises 33 patches, each with an approximate size of 2000×2500 pixels. On the other hand, the Potsdam dataset includes 38 patches, all of which have dimensions of 6000×6000 pixels. Notably, the Vaihingen dataset provides near-infrared (NIR), red, and green bands, while the Potsdam dataset provides near-infrared (NIR), red, green, and blue bands. To ensure fair comparison, we utilized only the NIR-R-G channels from both datasets and excluded the DSM data in our experiments.

The GID-15 dataset, also known as Gaofen Image Dataset with 15 categories, is a recent and extensive land-cover dataset notable for its large coverage, wide distribution, and high spatial resolution. The dataset includes 150 pixel-level annotated GF-2 images, each labeled with 15 categories, including industrial land (IDL), urban residential (UR), rural residential (RR), traffic land (TL), paddy field (PF), irrigated land (IL), dry cropland (DC), garden land (GL), arbor forest (AF), shrub land (SL), natural meadow (NM), artificial meadow (AM), river (RV), lake (LK), and pond (PN). All images in the GID-15 dataset are of dimensions 6800×7200 pixels with RGB three channels.

For the Potsdam dataset, we utilized 30 images for training, 4 images for validation, and an additional 4 images for testing. Regarding the Vaihingen dataset, we selected 25 images for training, 4 images for validation, and another 4 images for testing. In the case of the GID-15 dataset, we employed 90 images for training, allocated 10 images for validation, and reserved a separate set of 10 images for testing. This choice stems from the fact that, at present, only 110 annotated images from the training and validation sets have been made publicly available, and the official test server is currently inaccessible. For all datasets, our approach involved cropping small patches with dimensions of 600×600 pixels. The larger images were processed without alteration to preserve their original sizes.

Our experimental code is built upon mmsegmentation, a specialized deep learning framework for semantic segmentation developed using PyTorch. The experiments were conducted on a Tesla V100 server equipped with four GPUs, each having 32GB of memory. To ensure a fair comparison, we set the batch size for all models to 16 and the total epochs to 80. We employed the poly learning rate decay strategy, expressed by the formula:

$$lr = base_lr \times \left(1 - \frac{iter}{max_iter}\right)^{power} \quad (4)$$

In this equation, $base_lr$ is set to 0.01, and $power$ is set to 0.9. We utilized the SGD optimization algorithm with a momentum of 0.9 and a weight decay of 0.0005.

We employed several widely-used evaluation metrics in the field of semantic segmentation to gauge the performance of our model: overall accuracy (OA), mean intersection over union (mIoU, the average of all categories' IoU) and the harmonic mean of the precision and recall (F1-score). The calculation for these metrics are outlined in Equation (5)-(7).

$$OA = \frac{TP + TN}{TP + FP + FN + TN} \quad (5)$$

$$IoU = \frac{TP}{TP + FN + FP} \quad (6)$$

$$\begin{cases} Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN} \\ F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \end{cases} \quad (7)$$

In the formulas, TP, FP, TN, and FN correspond to the terms in the confusion matrix: True Positive, False Positive, True Negative, and False Negative, respectively.

B. Comparative Experiments

To validate the efficacy of our proposed GFFNet, we meticulously selected several models that hold significance in the field of semantic segmentation of RSIs. These models encompass a wide range of architectural approaches, spanning from CNN-based designs to Transformer-based frameworks. All methods share the same experimental settings with the proposed GFFNet. The chosen models are as follows:

- **FCN**: The fully convolutional network (FCN) marked a pivotal advancement in semantic segmentation. FCN pioneered the utilization of an end-to-end convolutional neural network for pixel-level image classification. Unlike models primarily designed for image classification, FCN innovatively incorporated transpose convolutions for efficient upsampling and skip connections. This combination enabled FCN to effectively capture and preserve spatial information across various scales.
- **PSPNet**: The pyramid scene parsing network (PSPNet) addresses the challenge of capturing diverse contextual scales within an image for enhanced segmentation accuracy. PSPNet's key innovation lies in its incorporation of a pyramid pooling module. This module adeptly aggregates features from multiple spatial resolutions, enabling the

network to consider both fine-grained local details and broader global context.

- **DeepLabV3+**: Building upon the foundation of previous DeepLab models, DeepLabV3+ introduces significant innovations. Notably, it adopts an expanded spatial pyramid pooling module, capturing contextual information across various scales while preserving intricate spatial details. Another noteworthy enhancement is the inclusion of a "decoder" module, which combines low-resolution, high-level features with the upsampled outputs of the Atrous module.
- **DANet**: The dual attention network (DANet) tackles the challenge of capturing long-range dependencies and contextual information by seamlessly integrating two attention mechanisms: spatial attention and channel attention. The spatial attention mechanism focuses on modeling relationships among different spatial positions within feature maps, while the channel attention mechanism deciphers interdependencies among feature channels.
- **SETR**: As a pioneering semantic segmentation study based on the Transformer architecture, SETR builds upon the ViT framework. It extends ViT by incorporating a decoder component, enabling it to perform semantic segmentation tasks and produce outcomes comparable to traditional CNN-based models.
- **SegFormer**: By employing a hierarchical Transformer encoder and a lightweight MLP decoder, SegFormer excels at extracting and merging multiscale features. Unlike the ViT paradigm that generates single-resolution feature maps, SegFormer generates multi-level features reminiscent of those produced by conventional CNNs. This allows it to capture high-resolution semantic details and low-resolution spatial characteristics simultaneously.
- **HRCNet**: Tailored to the specific demands of semantic segmentation in remote sensing images, HRCNet strategically integrates HRNet for spatial information extraction. It enhances this capability by incorporating a lightweight dual-attention module, facilitating comprehensive global context comprehension. Additionally, the network utilizes a feature-enhanced pyramid to effectively merge information across various scales.
- **LANet**: Addressing the limitations of traditional attention mechanisms, LANet [35] employs patch-level scene information to improve the semantic segmentation of RSIs. It features a patch attention module for improved context encoding via patchwise local descriptors, and an attention embedding module to enrich low-level layers with high-level attention, effectively integrating both for superior feature representation in RSIs.
- **RSSFormer**: The remote sensing segmentation framework (RSSFormer) is designed to tackle the challenges of land-cover segmentation in RSIs. It enhances foreground saliency modeling through three components: adaptive Transformer fusion module, Detail-aware attention layer and foreground saliency guided loss. RSSFormer achieves a good balance between the computational overhead and the accuracy.

TABLE I

COMPARATIVE RESULTS ON VAIHINGEN DATASET, INCLUDING THE IOU FOR EACH CLASS, AS WELL AS THE OVERALL EVALUATION METRICS FOR ALL CLASSES, MEASURED IN PERCENTAGE (%).

Method	Imp. surf.	Buildings	Low veg.	Trees	Cars	OA	mIoU	F1
FCN-8s [32]	83.1	90.2	69.4	77.3	76.9	88.5	79.4	87.1
PSPNet [5]	85.2	91.3	70.8	79.8	76.2	89.9	80.7	88.7
DeepLabV3+ [6]	84.8	90.7	71.3	78.4	77.1	89.6	80.5	88.4
DANet [9]	86.3	91.2	71.6	79.9	76.7	90.3	81.1	89.2
SETR [17]	84.4	89.2	68.9	77.2	73.5	87.4	78.6	86.3
SegFormer [36]	86.6	91.7	72.1	80.2	77.3	91.1	81.6	89.7
HRCNet [21]	85.9	91.1	71.2	80.5	77.4	90.7	81.2	89.3
LANet [35]	85.5	90.6	70.9	79.7	75.8	90.5	80.5	88.6
RSSFormer [28]	86.7	91.8	71.1	80.4	77.5	90.9	81.5	89.4
GFFNet (ours)	87.3	92.1	71.7	80.6	77.2	91.2	81.8	89.9

TABLE II

COMPARATIVE RESULTS ON POTSDAM DATASET, INCLUDING THE IOU FOR EACH CLASS, AS WELL AS THE OVERALL EVALUATION METRICS FOR ALL CLASSES, MEASURED IN PERCENTAGE (%).

Method	Imp. surf.	Buildings	Low veg.	Trees	Cars	OA	mIoU	F1
FCN-8s [32]	85.9	89.7	75.5	77.1	90.3	91.3	83.7	90.2
PSPNet [5]	87.6	93.8	76.4	79.3	92.8	92.5	86.0	91.5
DeepLabV3+ [6]	87.1	92.9	77.2	78.5	91.7	92.1	85.5	91.3
DANet [9]	88.2	93.9	77.3	80.2	92.5	92.7	86.4	91.8
SETR [17]	86.6	91.3	75.2	78.8	89.4	90.9	84.3	90.7
SegFormer [36]	88.5	93.5	78.4	80.1	92.9	92.9	86.7	92.0
HRCNet [21]	88.4	92.6	77.8	79.4	92.6	92.6	86.2	91.7
LANet [35]	87.7	93.1	77.5	78.9	91.8	91.8	85.8	91.6
RSSFormer [28]	88.3	93.6	78.2	80.3	91.5	92.8	86.4	91.9
GFFNet (ours)	89.7	94.1	78.9	80.6	92.7	93.1	87.2	92.4

TABLE III

COMPARATIVE RESULTS ON GID-15 DATASET, INCLUDING THE IOU FOR EACH CLASS, AS WELL AS THE OVERALL EVALUATION METRICS FOR ALL CLASSES, MEASURED IN PERCENTAGE (%).

Method	IDL	UR	RR	TL	PF	IL	DC	GL	AF	SL	NM	AM	RV	LK	PN	OA	mIoU	F1
FCN-8s [32]	61.2	68.5	56.4	57.3	70.7	69.8	67.5	42.3	78.1	74.4	44.2	66.1	71.5	83.2	68.3	79.3	65.3	72.4
PSPNet [5]	65.6	71.7	59.1	62.6	79.5	76.2	71.1	45.5	83.3	78.9	52.5	72.1	76.3	85.6	71.2	82.1	70.1	75.6
DeepLabV3+ [6]	66.2	72.1	61.2	60.3	77.5	75.1	69.8	46.2	82.1	79.4	51.3	73.1	74.5	84.7	72.3	81.9	69.7	75.1
DANet [9]	67.5	73.2	62.6	61.9	80.8	78.5	72.3	48.2	82.8	79.1	52.3	74.2	77.5	87.1	70.8	83.2	71.3	76.2
SETR [17]	62.4	69.5	58.3	61.1	76.2	75.8	68.9	45.1	79.6	76.8	49.2	68.3	72.4	85.2	68.9	81.4	67.8	74.5
SegFormer [36]	68.3	73.5	62.3	62.9	81.4	79.2	74.3	47.9	84.1	80.2	51.8	75.1	78.6	89.2	72.1	85.6	72.1	78.3
HRCNet [21]	66.9	74.1	61.5	61.7	79.2	77.4	73.3	46.5	83.9	78.2	53.4	73.8	76.9	88.7	72.6	84.5	71.2	77.8
LANet [35]	66.3	71.5	60.9	61.2	80.9	76.9	73.6	45.8	83.6	79.3	51.9	74.8	77.2	87.7	71.5	82.8	70.9	75.6
RSSFormer [28]	67.8	73.6	62.5	63.1	82.1	79.8	75.5	47.2	84.3	80.5	52.6	75.2	78.1	89.5	72.3	85.6	72.3	78.4
GFFNet (ours)	68.1	73.9	62.2	66.2	83.3	81.1	76.4	48.1	85.7	81.6	56.1	76.4	82.1	91.9	71.8	87.1	73.7	79.9

The results for the Vaihingen, Potsdam, and GID-15 datasets are presented in Table I, Table II and Table III, respectively. Notably, the proposed GFFNet consistently outperforms other methods in terms of the three evaluation metrics on all datasets. Furthermore, regarding mIoU, GFFNet achieves the highest performance improvement compared to other methods, with gains of 0.2%, 0.5%, and 1.4% on these datasets, respectively. Several factors contribute to this performance enhancement. Firstly, as the datasets progressively increase in scale, our GFFNet, with its global contextual learning branch, demonstrates its advantage, as it is capable of effectively lever-

aging larger datasets compared to local methods. Secondly, the GID-15 dataset, in particular, presents more complex land cover types, which intensifies the reliance on contextual information for accurate segmentation. Figure 7 and Figure 8 showcase some segmentation results on the three datasets. It can be observed that GFFNet performs exceptionally well, particularly in scenarios where it is difficult to accurately determine land cover types in a single image due to the lack of contextual information, such as the river in the second set of images in Figure 8, where GFFNet demonstrates a significant advantage.

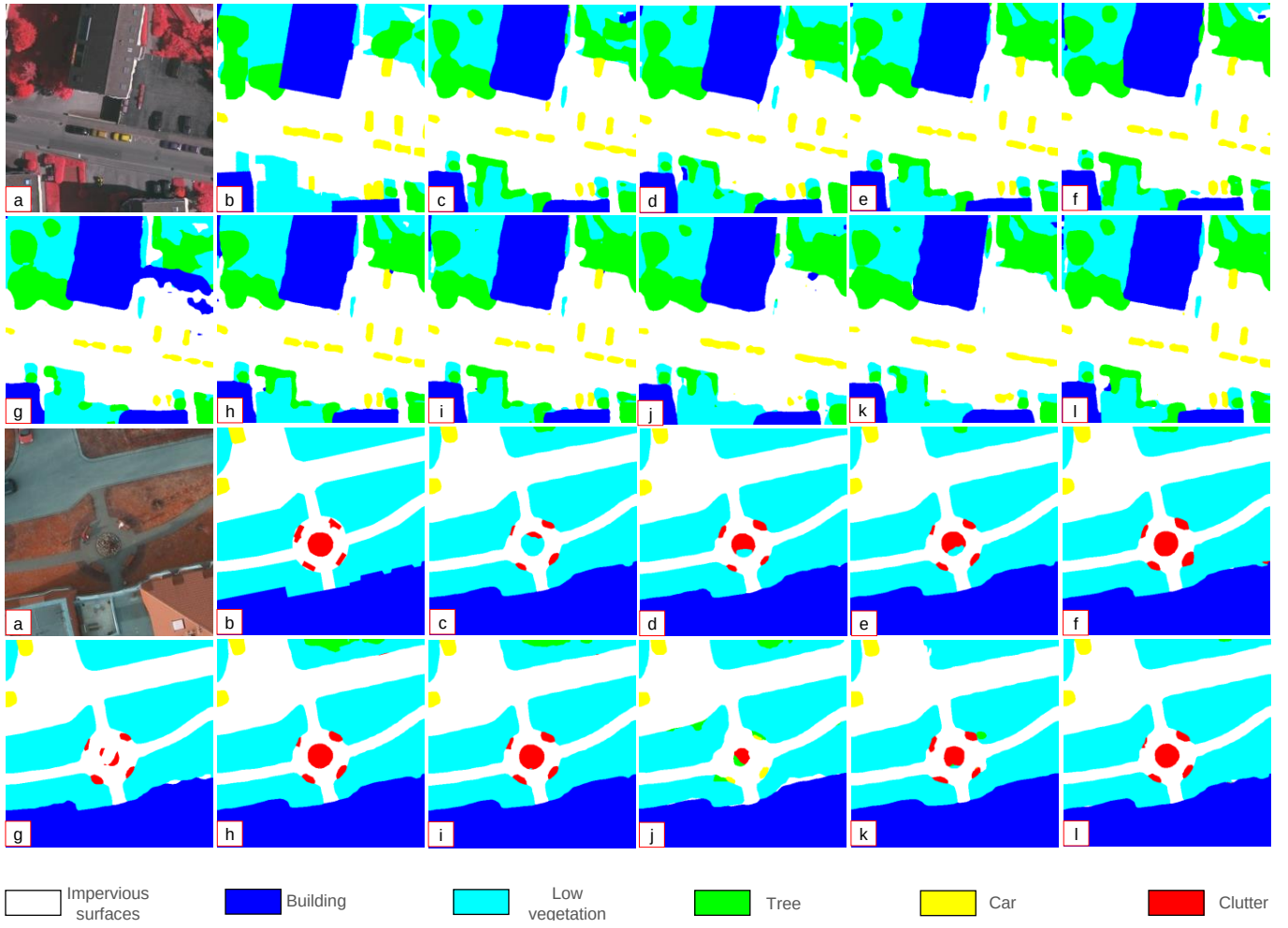


Fig. 7. Visual comparisons on the ISPRS dataset, with the top and bottom sets of images coming from the Vaihingen and Potsdam datasets, respectively. (a) input image, (b) ground truth, (c) FCN-8s, (d) PSPNet, (e) DeepLabV3+, (f) DANet, (g) SETR, (h) SegFormer, (i) HRCNet, (j) LANet, (k) RSSFormer, and (l) GFFNet (ours).

When considering specific categories, GFFNet demonstrates superior performance on the ISPRS dataset, surpassing other methods in terms of mIoU for the Impervious surfaces, Building, and Tree categories. Similarly, on the GID-15 dataset, GFFNet outperforms other methods in mIoU for ten categories, including traffic land (TL), paddy field (PF), irrigated land (IL), dry cropland (DC), arbor forest (AF), shrub land (SL), Natural meadow (NM), artificial meadow (AM), river (RV), and lake (LK). Notably, these categories often encompass expansive geographical regions, and GFFNet's ability to capture global contextual information contributes to its superior segmentation capabilities for these specific classes.

It's important to note that the Transformer-based SETR model demonstrates relatively poorer performance on the ISPRS dataset, at times even falling below the baseline set by the original FCN model. While its performance on the GID-15 dataset surpasses that of FCN, it still lags behind classic convolution-based segmentation models such as PSPNet and DeepLab. This observation indirectly emphasizes that, when working with remote sensing datasets, solely relying on Transformers may not guarantee superior results, particularly due to limitations in dataset size. Achieving optimal perfor-

mance might necessitate the incorporation of domain-specific inductive biases. For example, models like SegFormer, which introduce multiscale features, have shown notably improved performance.

TABLE IV
COMPARISON OF MODEL SIZE AND COMPUTATIONAL COST.

Method	Backbone	Params (M)	GFLOPs
FCN-8s [32]	ResNet-101	68.4	378.2
PSPNet [5]	ResNet-101	67.9	351.7
DeepLabV3+ [6]	ResNet-101	62.4	310.7
DANet [9]	ResNet-101	69.0	380.2
LANet [35]	ResNet-101	63.5	325.6
SETR [17]	VIT-L	306.3	292.5
SegFormer [36]	MIT-B5	81.8	72.6
HRCNet [21]	HRNetV2-W48	76.2	158.4
RSSFormer [28]	HRNetV2-W48	72.9	169.1
GFFNet (ours)	HRNetV2-W48	74.3	189.6

To assess and compare the computational efficiency of

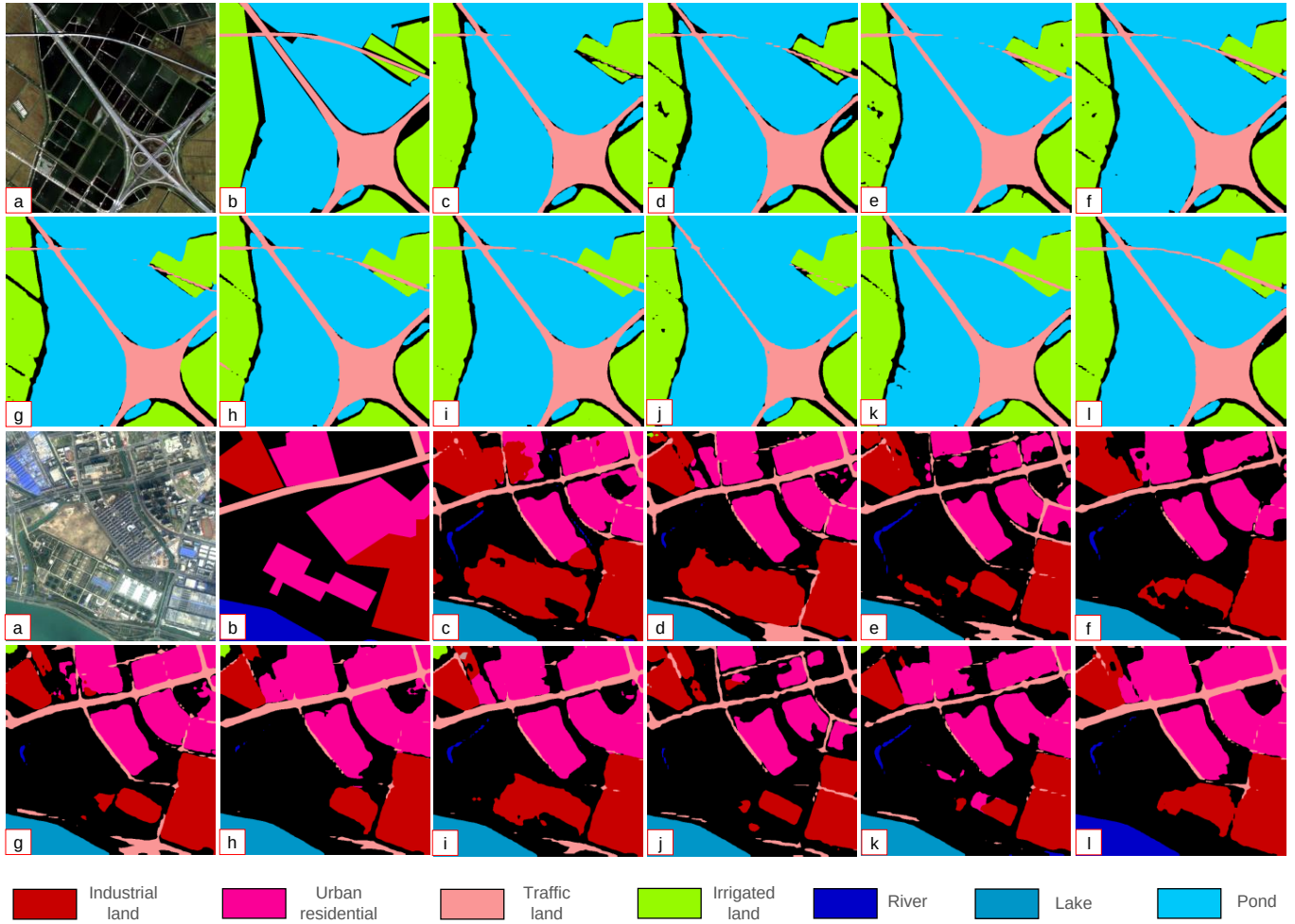


Fig. 8. Visual comparisons on the GID-15 dataset. (a) input image, (b) ground truth, (c) FCN-8s, (d) PSPNet, (e) DeepLabV3+, (f) DANet, (g) SETR, (h) SegFormer, (i) HRCNet, (j) LANet, (k) RSSFormer, and (l) GFFNet (ours).

different models, we conducted an analysis of their parameter counts and computational costs, which are listed in Table IV. In this context, "Params" indicates the spatial complexity of a model, while "GFLOPs" reflects its time complexity. The model configurations are based on the GID-15 experiment. It's noteworthy that when compared to models utilizing ResNet-101 [37] as their backbone network, our proposed GFFNet demonstrates a slight increase in Params but a notable advantage in computational efficiency. Furthermore, in comparison to Transformer-based models like SegFormer and models explicitly tailored for remote sensing images like HRCNet, our model exhibits competitive performance in terms of both parameter count and computational cost.

C. Ablation study

The innovative group Transformer and CFF modules constitute the cornerstone elements within the GFFNet architecture. To thoroughly evaluate the contributions of these modules to the overall effectiveness of the network, we conducted an extensive series of ablation experiments using the GID-15 dataset.

Group Transformer As illustrated in Figure 9, we examined four distinct network configurations: (a) utilizing only

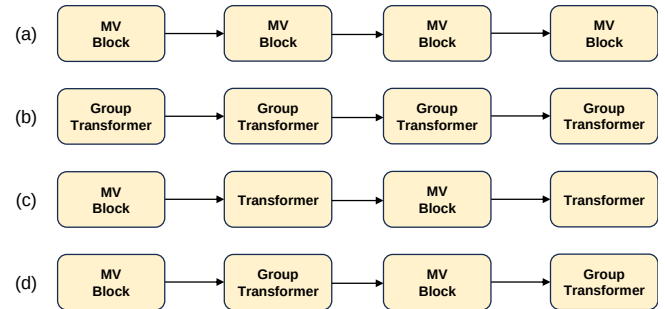


Fig. 9. Network configurations of ablation study of group Transformer. (a) use MobileNet blocks; (b) use group Transformer; (c) use MobileNet blocks and vanilla Transformers alternately; and (d) use MobileNet blocks and group Transformers alternately.

TABLE V
ABLATION RESULTS OF GROUP TRANSFORMER ON GID-15 DATASET

Network configuration	OA	mIoU	Fscore
structure (a)	84.4	71.2	77.8
structure (b)	85.2	71.9	78.5
structure (c)	84.5	71.2	77.9
structure (d)	87.1	73.7	79.9

TABLE VI
ABLATION RESULTS OF CFF MODULE ON GID-15 DATASET

Feature fusion mode	OA	mIoU	Fscore
Addition	81.4	69.9	76.8
Average	82.3	70.2	77.1
Max-out	78.2	68.5	74.6
Multiplication	71.7	64.1	69.8
Concatenation	83.0	71.6	77.9
CFF (ours)	87.1	73.7	79.9

MobileNet blocks; (b) employing only the group Transformer; (c) alternating MobileNet blocks with traditional Transformers; and (d) employing an alternating arrangement of MobileNet blocks and group Transformers. The findings, as presented in Table V, are noteworthy. Constructing a network architecture solely based on MobileNet blocks resulted in minimal gains in overall accuracy. This deficiency can be attributed to the limited receptive field of CNNs, which struggle to capture extensive contextual relationships within large Remote Sensing Images (RSIs). Similarly, relying exclusively on the group Transformer led to only marginal improvements in performance. This limitation arises from the absence of CNN convolutional kernel fusion, causing the feature map to represent merely a localized sampling of global features, lacking rich semantic details. Incorporating traditional Transformers in an alternating arrangement with MobileNet blocks did not yield significant performance enhancements either. We hypothesize that this outcome is due to the inherent spatial displacement introduced by vanilla Transformers, disrupting the spatial coherence of the feature map. Given the relative scarcity of remote sensing datasets, maintaining spatial context is crucial. In contrast, our GFFNet effectively addresses these challenges. Through the strategic integration of the group Transformer, we not only reduce computational demands but also preserve spatial contextual information within the feature maps. By seamlessly fusing this approach with MobileNet modules, we empower the network to understand global context comprehensively, resulting in superior performance.

CFF Module To assess the effectiveness of our CFF module, we conducted a comparative analysis against other commonly used feature fusion techniques. These techniques include: (a) Addition: element-wise summation of the two feature maps; (b) Average: element-wise averaging of the two feature maps; (c) Max-out: element-wise selection of maximum values between the two feature maps; (d) Multiplication: element-wise multiplication of the two feature maps; (e) Concatenation: channel-wise concatenation of the two feature maps; and (f) the cross feature fusion module as proposed in this study. As illustrated by Table VI, it is evident that our CFF module significantly outperforms traditional feature fusion approaches, thereby considerably enhancing the model's performance.

Global Context Learning Branch The global context learning branch within GFFNet serves as an essential and versatile feature extraction mechanism. Customized for semantic segmentation applications, this specialized branch can seam-

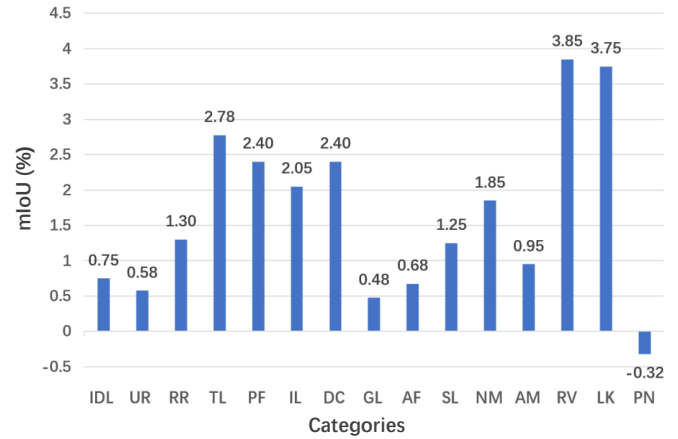


Fig. 10. Average performance (mIoU) improvement with global context learning branch

lessly integrate into various existing segmentation frameworks to capture intricate global contextual information—an advantage especially beneficial when dealing with cropped data. To empirically validate the utility of this feature extraction structure, we integrated it into several established models, including FCN, PSPNet, DeepLabV3+, and HRNet, and subjected them to rigorous testing on the GID-15 dataset. The results, as presented in Table VII, unequivocally demonstrate that the incorporation of the global context learning branch significantly enhances segmentation accuracy across the board. This leads to substantial mIoU improvements of 1.6%, 1.1%, 1.3%, and 2.1% for each respective model. Furthermore, we systematically quantified the average performance improvement across diverse categories, as illustrated in Figure 10. As evident from the visual data, GFFNet brings about notable enhancements in 14 out of the 15 evaluated categories, highlighting its comprehensive effectiveness. These improvements are particularly prominent in categories characterized by extensive spatial scales, such as traffic land (TL), paddy field (PF), irrigated land (IL), dry cropland (DC), river (RV), and lake (LK). A slight performance decrease in the pond (PN) category can be attributed to its smaller spatial dimensions, rendering it less reliant on global contextual cues. Additionally, this minor impact could also stem from the category's relatively limited representation within the dataset.

V. DISCUSSION

The discussion section aims to provide a comprehensive analysis of our study's implications, comparing it with existing methods, acknowledging its limitations, and suggesting future research directions.

A. Significance and implications of the study

Our study introduces the global feature fusion network (GFFNet), a novel approach that effectively integrates global and local contextual information for semantic segmentation of RSIs. The key significance of GFFNet lies in its ability to handle the large-scale and high-dimensional nature of RSIs, a challenge that has been a bottleneck in traditional methods. By

TABLE VII
ABLATION RESULTS ON GID-15 DATASET, † MEANS ADD THE GLOBAL CONTEXT LEARNING BRANCH, MEASURED IN PERCENTAGE (%).

Method	IDL	UR	RR	TL	PF	IL	DC	GL	AF	SL	NM	AM	RV	LK	PN	OA	mIoU	F1
FCN-8s [32]	61.2	68.5	56.4	57.3	70.7	69.8	67.5	42.3	78.1	74.4	44.2	66.1	71.5	83.2	68.3	79.3	65.3	72.4
FCN-8s†	62.6	69.3	58.9	60.2	72.3	71.2	69.1	43.2	78.8	75.9	45.3	66.7	75.1	86.8	67.7	81.2	66.9	74.1
PSPNet [5]	65.6	71.7	59.1	62.6	79.5	76.2	71.1	45.5	83.3	78.9	52.5	72.1	76.3	85.6	71.2	82.1	70.1	75.6
PSPNet†	66.7	72.2	60.6	63.9	81.1	78.4	73.8	46.2	83.1	79.4	53.4	71.7	79.1	88.5	70.1	84.6	71.2	77.5
DeepLabV3+ [6]	66.2	72.1	61.2	60.3	77.5	75.1	69.8	46.2	82.1	79.4	51.3	73.1	74.5	84.7	72.3	81.9	69.7	75.1
DeepLabV3+†	65.9	72.7	61.6	62.4	80.3	77.2	72.1	45.8	82.9	80.3	52.7	72.9	77.8	87.4	72.4	83.7	71.0	77.2
HRNet [31]	67.3	73.5	61.4	61.5	79.7	78.6	73.4	47.4	84.3	79.5	52.1	72.6	76.4	86.1	71.5	84.1	71.0	77.7
GFFNet (ours)	68.1	73.9	62.2	66.2	83.3	81.1	76.4	48.1	85.7	81.6	56.1	76.4	82.1	91.9	71.7	87.1	73.7	79.9

leveraging a Group Transformer structure coupled with group convolution, GFFNet excels in extracting global contextual features from large-scale RSIs. This is particularly vital for RSIs where objects like traffic land, field, and rivers require both local detail and global context for accurate segmentation.

The integration of the cross feature fusion (CFF) module further refines the model's performance by merging local spatial features with global contextual information. This integration is pivotal in segmenting large-scale objects, which is evident from the superior performance of GFFNet on datasets like ISPRS and GID-15. The capability to process large images without cropping them into smaller patches is a substantial advancement, reducing information loss and improving segmentation accuracy.

B. Comparison with existing methods

The traditional CNN-based methods, while effective in extracting local features, often falter in capturing the global context, particularly in large-scale RSIs. Transformer-based models, despite their success in other domains, have limitations in the context of RSIs due to their inherent requirement for large annotated datasets and their lack of inductive biases specific to spatial information in images. GFFNet addresses these gaps by combining the strengths of CNNs and Transformers, thereby extracting both local and global contextual information effectively.

C. Limitations and future research directions

Despite the promising results, GFFNet has certain limitations that open avenues for future research. First, the computational complexity, although reduced compared to traditional Transformers, still poses challenges for very large-scale RSIs. Optimizing the model further to reduce computational demands without compromising performance is an essential area for future work. Additionally, while GFFNet demonstrates robust performance on certain land cover types, its effectiveness across diverse and more complex landscapes needs further exploration.

Another potential direction is the adaptation of GFFNet for real-time applications. The current model, primarily designed for accuracy, may not meet the computational efficiency requirements for real-time processing. Developing a more lightweight version of GFFNet could expand its applicability to time-sensitive remote sensing tasks.

Lastly, integrating multi-spectral data could significantly enhance the model's utility in remote sensing. Most RSIs contain a wealth of spectral information that is not fully utilized in the current model. Incorporating these data could improve segmentation accuracy, especially in distinguishing between land cover types with similar textures but different spectral signatures.

VI. CONCLUSION

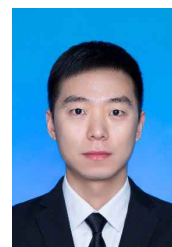
In conclusion, this study presents a novel GFFNet for semantic segmentation of RSIs. GFFNet, featuring an innovative group Transformer and a cross feature fusion (CFF) module, adeptly combines global and local contextual information. Its lightweight design, alternating between group Transformers and CNNs, efficiently extracts comprehensive semantic details from large-scale RSIs. Our extensive evaluations on the ISPRS 2D and GID-15 datasets demonstrate GFFNet's superior performance over traditional methods, particularly in accuracy, mean intersection over union (mIoU), and F1 score.

The model shows exceptional prowess in handling complex land cover types and large-scale structures, excelling in categories like traffic land, fields, and rivers. Ablation studies confirm the significant contribution of each component in enhancing overall performance. However, the model's effectiveness varies with smaller-scale or less-represented categories. Future efforts will focus on balancing local spatial modeling with global context learning, aiming to refine GFFNet's performance across a broader range of categories.

REFERENCES

- [1] Q. Zhang and K. Seto, "Mapping urbanization dynamics at regional and global scales using multi-temporal dmsp/ols nighttime light data," *Remote Sensing of Environment*, vol. 115, no. 9, pp. 2320–2329, 2011.
- [2] M. Maboudi, J. Amini, S. Malihi, and M. Hahn, "Integrating fuzzy object based image analysis and ant colony optimization for road extraction from remotely sensed images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 138, pp. 151–163, 2018.
- [3] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, "Land cover mapping at very high resolution with rotation equivariant cnns: Towards small yet accurate models," *ISPRS journal of photogrammetry and remote sensing*, vol. 145, pp. 96–107, 2018.
- [4] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Systems with Applications*, vol. 169, p. 114417, 2021.
- [5] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

- [6] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
- [7] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.
- [8] C. Liu, L. Chen, F. Schroff, H. Adam, W. Hua, A. L. Y., and F. Li, "Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation," in *Computer Vision and Pattern Recognition*, 2019.
- [9] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.
- [10] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 603–612.
- [11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [12] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 4015–4026.
- [13] K. Chen, C. Liu, H. Chen, H. Zhang, W. Li, Z. Zou, and Z. Shi, "Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model," *arXiv preprint arXiv:2306.16269*, 2023.
- [14] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sensing of Environment*, vol. 237, p. 111322, 2020.
- [15] "Isprs 2d semantic labeling contest," <https://www.isprs.org/education/benchmarks/UrbanSemLab/semantic-labeling.aspx>.
- [16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [17] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.
- [18] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7262–7272.
- [19] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.
- [20] Y. Liu, B. Fan, L. Wang, J. Bai, S. Xiang, and C. Pan, "Semantic labeling in very high resolution images via a self-cascaded convolutional neural network," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 78–95, 2018.
- [21] Z. Xu, W. Zhang, T. Zhang, and J. Li, "Hrcnet: High-resolution context extraction network for semantic segmentation of remote sensing images," *Remote Sensing*, vol. 13, no. 1, p. 71, 2020.
- [22] W. Cheng, W. Yang, M. Wang, G. Wang, and J. Chen, "Context aggregation network for semantic labeling in aerial images," *Remote Sensing*, vol. 11, no. 10, p. 1158, 2019.
- [23] X. Chen, Z. Li, J. Jiang, Z. Han, S. Deng, Z. Li, T. Fang, H. Huo, Q. Li, and M. Liu, "Adaptive effective receptive field convolution for semantic segmentation of vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 4, pp. 3532–3546, 2020.
- [24] L. Gao, H. Liu, M. Yang, L. Chen, Y. Wan, Z. Xiao, and Y. Qian, "Stranfuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 10990–11003, 2021.
- [25] H. Wang, X. Chen, T. Zhang, Z. Xu, and J. Li, "Cctnet: Coupled cnn and transformer network for crop segmentation of remote sensing images," *Remote Sensing*, vol. 14, no. 9, p. 1956, 2022.
- [26] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, "A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.
- [27] 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 Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 196–214, 2022.
- [28] R. Xu, C. Wang, J. Zhang, S. Xu, W. Meng, and X. Zhang, "Rssformer: Foreground saliency enhancement for remote sensing land-cover segmentation," *IEEE Transactions on Image Processing*, vol. 32, pp. 1052–1064, 2023.
- [29] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
- [30] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
- [31] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.
- [32] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
- [33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18. Springer, 2015, pp. 234–241.
- [34] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.
- [35] L. Ding, H. Tang, and L. Bruzzone, "Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 426–435, 2020.
- [36] 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. 12077–12090, 2021.
- [37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.



Yong Cao received the B.S. degree in automation from Northeastern University, Shenyang, China, in 2016. He is currently working toward the Ph.D. degree in pattern recognition and intelligent system with the Institute of Automation, Chinese Academy of Sciences, Beijing, China.

His current research interests include remote sensing image interpretation, computer vision, semantic segmentation.



Chunlei Huo (Member, IEEE) received the B.S. degree in applied mathematics from Hebei Normal University, Shijiazhuang, China, in 1999, the M.S. degree in applied mathematics from Xidian University, Xi'an, China, in 2002, and the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2009.

He is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His current

research interests include remote sensing image processing, computer vision, pattern recognition, so on.



Shiming Xiang (Member, IEEE) received the B.S. degree in mathematics from Chongqing Normal University, Chongqing, China, in 1993, the M.S. degree from Chongqing University, Chongqing, China, in 1996, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2004. From 1996 to 2001, he was a Lecturer with the Huazhong University of Science and Technology, Wuhan, China. He was a Postdoctorate Candidate with the Department of Automation, Tsinghua University, Beijing, China,

until 2006.

He is currently a Professor with the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include pattern recognition, machine learning, and computer vision.



Chunhong Pan (Member, IEEE) received the B.S. degree in automatic control from Tsinghua University, Beijing, China, in 1987, the M.S. degree from the Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai, China, in 1990, and the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2000.

He is currently a Professor with the National Laboratory of Pattern Recognition of Institute of Automation, Chinese Academy of Sciences. His research interests include computer vision, image processing, computer graphics, and remote sensing.