Multiscale Context-Aware Feature Fusion Network for Land-Cover Classification of Urban Scene Imagery

Abubakar Siddique, Zhengzhou Li, *Member, IEEE*, Abdullah Azeem, Yuting Zhang, and Bitong Xu

***Abstract*—Recently, several land-cover classification models have achieved great success in terms of both accuracy and computational performance. However, it remains challenging due to interclass similarities, intraclass variations, scale-related inaccuracies, and high computational complexity. First, these methods fail to establish a correlation among different feature maps during multiscale feature extraction, leading to interclass similarities and intraclass variations. Second, they underutilize feature interdependencies of the context contained in each layer of the encoder–decoder architecture, causing scale-related inaccuracies. Third, they cause checkerboard artifacts and blurry edges, which can negatively impact the accuracy and generate segmentation map at increased computational cost. To address these problems, this article proposes a novel multiscale context-aware feature fusion network (MCN) for high-resolution urban scene images. MCN mainly consists of three modules: First, a multiscale feature enhancement module for backbone network to extract rich spatial information dynamically by incorporating dense correlation among feature maps with different receptive fields; second, multilayer feature fusion module as skip connections to produce a single high-level representation of the local–global context by capturing low-, mid-, and high-level interdependencies at different encoder–decoder stages; and third, pixel-shuffle decoder to reduce the blurry edges and checkerboard artifacts while upsampling with reduced number of parameters. Experiments on three high-resolution aerial and satellite urban scene datasets show that MCN consistently outperforms the mainstream land-cover classification models. Specifically, MCN achieves an OA of 93.51 on Potsdam, 90.18 on Vaihingen, and an mIoU of 73.73 on DeepGlobe.**

***Index Terms*—Attention mechanism, multiscale context aggregation, remote sensing, semantic segmentation, similarity fusion, urban scene images.**

Manuscript received 16 May 2023; revised 28 July 2023; accepted 24 August 2023. Date of publication 30 August 2023; date of current version 20 September 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61675036 and in part by the Chinese Academy of Sciences Key Laboratory of Beam Control Fund under Grant 2017LBC006. *(Corresponding author: Zhengzhou Li.)*

Abubakar Siddique, Abdullah Azeem, Yuting Zhang, and Bitong Xu are with the School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China (e-mail: abubakar@cqu.edu.cn; abdullah.azeem@cqu.edu.cn; hangyuting@cqu.edu.cn; xubitong@cqu.edu.cn).

Zhengzhou Li is with the School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China, and also with the Key laboratory of Optical Engineering, Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu 610209, China (e-mail: lizhengzhou@cqu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3310160

I. Introduction

SEMANTIC segmentation is typically regarded as a problem of land-cover classification and aims at providing a category label for each pixel in an urban scene image. It plays a vital role in land-cover mapping [1], [2] urban planning, change detection [3], road extraction, and environmental protection. With the advancement of sensor technology, plenty of high-resolution (HR) urban scene images have been captured. Urban scene images with rich potential semantic content and abundant spatial details can provide data support for segmentation. However, large-scale variation, unbalanced distribution of the ground objects and their categories, interclass similarities, intraclass variations, and difficulty in extracting comprehensive feature information have challenged land-cover classification approaches to accurately identify and segment objects in HR urban scene images.

Earlier methods had made some progress employing the traditional feature design and convolution neural networks (CNNs), making the data-driven feasible for multiscale feature learning for urban scene images. Conventional methods often use handcrafted features (such as spectral, spatial, and textural) and traditional machine learning methods (such as support vector machine and random forest) to segment the urban scene images. However, the traditional methods depend on handcrafted features, consistently achieving unsatisfactory performance. Thus, designing good feature extractors for multiscale stimuli for solving urban scene image semantic segmentation problems is essential. It requires feature extractors to use larger receptive fields to identify context at multiscale.

Unsurprisingly, CNNs learn multiscale features through a stack of convolutional operators. This ability of CNNs leads to effective representations for solving several vision tasks (i.e., classification, object detection, and semantic segmentation). Compared with traditional methods, CNN-based approaches have shown tremendous success in semantic segmentation. In modern computer vision systems, CNNs are the most common choice of visual encoders. The most popular encoders include AlexNet [4], VGGNet [5], ResNet [6], InceptionNet-v4 [7], and InceptionResNet-v2 [7]. Recently, CNNs have been challenged by vision transformers (ViTs) [8], which have demonstrated good performance in several vision tasks. ViTs are designed either locally or globally and can gather information from a larger region using a larger window size. In contrast, these methods are usually computationally heavy and require significant

memory to capture the global context. Contrastingly, attention mechanisms are widely adopted for semantic segmentation due to their advantages in acquiring long-range context information. DANet [9] and CBAM [10] are state-of-the-art approaches that incorporate spatial and channel attention mechanisms to enhance the feature representation abilities of models.

With the widespread adoption of CNNs for vision tasks, significant advancements have been made for natural image semantic segmentation, producing spectacular results. Fully convolutional network (FCN) [11] is pioneering work in semantic segmentation tasks, employing FCNs without fully connected layers for end-to-end dense pixel prediction. U-Net [12] introduces skip connections in encoder–decoder architecture and U-Net++ [13] dense skip connections to bridge the semantic gap between encoder–decoder feature maps. PSP-Net [14] leverages the pyramid pooling module (PPM) and DeepLabV3+ [15] atrous spatial pyramid pooling (ASPP) to steadily segment targets at multiscale to capture object and image context.

The above-mentioned CNN-based semantic segmentation methods have three limitations. First, these methods construct a multiscale model using single-sized fixed convolutional kernels, dilated convolution with different dilation rates, or pooling grids without enabling correlation among different feature maps. Second, they underutilize feature interdependencies of the context contained in each layer of the encoder–decoder architecture, causing scale-related inaccuracies. Finally, these methods usually use deconvolution or bilinear-based upsampling techniques to recover image resolution, causing checkerboard artifacts and blurry edges at increased computational cost. Thus, it is essential to design good feature extractors to simultaneously identify local and global contexts. Redesigned skip connections to prevent the loss of fine- and coarse-grained details at shallower and deeper layers, and upsampling techniques to fully fuse feature information at reduced computational cost for HR urban scene image's semantic segmentation tasks.

To mitigate the above-mentioned limitations, this work proposes an alternative simple yet effective multiscale context-aware feature fusion network (MCN) for HR urban scene images. Concretely, MCN includes three fundamental components: multiscale feature enhancement (MFE) module, multilayer feature fusion (MLF) module, and pixel-shuffle decoder (PSD) module. MFE is exploited for the backbone network to identify the local and global context of ground objects while suppressing the background noise and capturing complementary features by enabling correlation among different levels of feature maps. MLF is introduced as skip connections where features from different MFE layers are merged before the supervision to produce a single high-level representation of the input data by leveraging the strength of each layer in capturing low-, mid-, and high-level features. The network can better segment the image into various semantic classes by merging these learned representations from all layers. PSD is used to enhance the resolution of the feature maps in the decoder and fully fuse feature information from various scales. Unlike other upsampling methods, PSD can better fix blurry edges and checkerboard artifacts with fewer parameters while improving network speed and accuracy.

The contributions to this work are given as follows.

1) A novel MCN with three fundamental modules is proposed to solve the interclass similarities, intraclass variations, scale-related inaccuracies, and high computational complexity issues.
2) We adopted the well-extracted multiscale information to identify local and global contexts simultaneously. Redesigned skip connections to leverage the strengths of each layer in capturing different levels of features and upsampling technique to fully fuse feature information at various receptive fields while reducing the number of parameters.
3) Extensive experiments are conducted on the ISPRS 2-D semantic labeling datasets and DeepGlobe to demonstrate the effectiveness of MCN, which yields notable performance gains compared with the existing architectures but with much fewer parameters.

II. Related Works

This section discusses similar techniques commonly employed in the semantic segmentation of HR urban scene images, including multiscale feature learning, skip connections, upsampling methods, and visual attention mechanisms.

A. Encoder–Decoder Architecture

One of the first semantic segmentation efforts using CNNs is an FCN [11]. However, during downsampling, FCN reduces spatial information by a larger factor. Thus, during upsampling, it becomes difficult to reproduce fine details even after using transpose convolution, which results in coarse output. To tackle this issue, Ronneberger et al. [12] introduce skip connections in the encoder–decoder module. However, due to the fixed receptive field of convolutional kernels, U-Net [12] suffers from extracting multiscale features. Zhao et al. [14] introduce an effective PPM to capture multiscale features by applying pooling operations at different grids. However, the pooling-based approach (i.e., PPM) may lose pixel-level fine detail information because distinct pixels may use the same contextual information. Chen et al. [15] introduced DeepLabv3+, which utilizes a more effective ASPP module that deploys multiple parallel filters with different dilation rates to capture multiscale features without adding extra computational costs. However, it can only manage scale variation to some degree. In addition, sparse sampling leads to spatial information loss, and a larger dilation rate produces gridding artifacts.

1) Multiscale Feature Learning: The accurate semantic segmentation of urban scene imagery requires the multiscale feature information of the region of interest. Extracting ground object features at various scales can help address interclass similarities and intraclass variances for diverse situations. The AlexNet [4] stacks filters sequentially and achieves significant performance gain over traditional methods. VGGNet [5] stacks filters of smaller kernels to increase the network depth and receptive field. Although VGGNet provides a more robust multiscale feature representation than AlexNet, due to stacking filters directly, AlexNet and VGGNet had a relatively fixed receptive field

for each layer. Szegedy et al. [7] introduced InceptionNet-v4 with different filter sizes in parallel to increase receptive fields and InceptionResNetv2 with inception and residual connections to enhance the efficiency of multiscale feature learning. SegFormer [16] presents a hierarchical transformer encoder to extract multiscale features and employs a multilayer perception (MLP) decoder to aggregate information from different layers. In urban scene semantic segmentation, Xu et al. [17] proposed a network to solve the problem of existing backbones in extracting multiscale features due to a large downsampling factor. In [18], an ensemble learning paradigm is employed to adaptively fuse the features from different scales and a pointwise convolution method to reduce the parameters while improving the model's accuracy. Extracting essential contextual information about ground objects requires CNN models to process features at various scales for effective semantic segmentation. In summary, to adequately utilize the rich spatial information in HR urban scene images and improve feature extraction's robustness among diverse and complex ground scenes, we introduce kernels of different sizes (i.e., 3×3, 5×5, and 7×7) in our work.

2) Skip Connections: Skip connections were introduced to solve different problems in different architectures, such as ResNets [6], for degradation and U-Net [12] for encoder–decoder architecture to prevent the loss of fine-grained details (i.e., object boundaries). U-Net++ [13] introduces dense skip connections to replace the plain skip connections in U-Net to bridge the semantic gap between encoder–decoder feature maps. However, due to the skip connections scheme and fixed receptive field of each layer, it is challenging for both U-Net and U-Net++ to model the global multiscale context for HR urban scene images. In urban scene semantic segmentation, 2DSegFormer [19] designed dilated residual connections as skip connections to further increase the receptive field of deep feature maps. MAResU-Net [20] redesigned skip connections in U-Net based on linear attention mechanism and ResNet. MSCA-Net [21] presents skip connections with atrous convolution to deal with the segmentation problems of multiscale urban scene images. MACU-Net [22] introduces skip connections with a channel attention mechanism to combine the multiscale features. We found that skip connections are helpful for several reasons in HR urban scene image segmentation. First, residual connections [6] allow the network to learn more complex feature mappings and facilitate faster convergence without being affected by vanishing gradient problems. Second, skip connections [12] are essential for encoder–decoder-based architectures as they allow the decoder to directly access the feature maps from corresponding levels of the encoder, thus preserving fine-grained details that might, otherwise, be lost as the spatial resolution decreases. It makes U-Net [12] particularly effective for tasks, such as image segmentation, where precise localization of object boundaries is essential. In summary, we used a residual connection in MFE and redesigned kip connections as MLF.

3) Upsampling Methods: CNNs are popular and highly performant choices for dense-level prediction. One commonly required component in CNNs is to increase the low-resolution feature maps for network visualization. Interpolation and deconvolution are the most common upsampling methods for recovering spatial information from convolution or max-poling layers. Interpolation upsampling methods include the nearest neighbor, bicubic, and bilinear interpolation. These methods lack the "learnable" aspect, blur the images, and aliasing distortions. Deconvolution upsamples low-resolution images using learnable kernels while improving upsampling during training. However, "uneven overlap" can easily occur during deconvolution, meaning that the convolutional kernel operates more in some places than others, causing checkerboard artifacts. Deconvolution was first proposed in FCN [11] and has been used in later segmentation models, i.e., U-Net [12]. In contrast to interpolation and deconvolution, Shi et al. [23] introduced the parameter and checkerboard artifacts free upsampling method, i.e., pixel shuffle (PS), also known as subpixel convolution for single-image super-resolution (SISR), which was later used in semantic segmentation tasks. PS provides a larger receptive field to capture more contextual information with minimal loss of information while maintaining the quality of generated segmentation. In urban scene semantic segmentation, many methods based on PS have been proposed. Chen et al. [24] proposed an end-to-end semantic segmentation network by inserting a shuffling layer in DeepLab architecture. They designed a field-of-view method to enhance the prediction while using an ensemble method to improve the model performance. Zhang et al. [25] proposed a network that simultaneously solves the super-resolution semantic segmentation and super-resolution image reconstruction by using low-resolution images to generate an HR segmentation image. We found that methods having upsampling layers, i.e., deconvolution or bilinear interpolation cause images to be distorted by checkerboard artifacts. In summary, the proposed work uses a PS operation in the decoder to improve the resolution of output feature maps and leverage the advantage of the MCN in capturing multiscale context. Using PS further improved the network speed and accuracy while alleviating the edge blur and artifacts caused by information loss.

B. Visual Attention Mechanism

The attention mechanism can improve the saliency representation of important features while suppressing interference from redundant features. SPOL [26] analyzed the importance of shallow features and used global average pooling to suppress background noise. In urban scene semantic segmentation, SCAttNet [27] combines channel and spatial attention mechanisms to refine the feature map. Zhang et al. [28] proposed a network to adaptively recalibrate feature responses and simultaneously aggregate global information along the channel and spatial dimensions to improve feature representation. Li et al. [29] propose a dual-channel scale-aware segmentation network with position and channel attention. PGNet [30] uses transformer-based architecture to fully leverage the long-range dependencies and global contextual information to segment objects of varying sizes. Ding et al. [31] proposed a network that utilizes CNN encoder and global–local attention-based transformer decoder to model global and local information. Gao et al. [32] proposed a network that used a dual-branch encoder based on CNN transformer to model local and global semantic information and

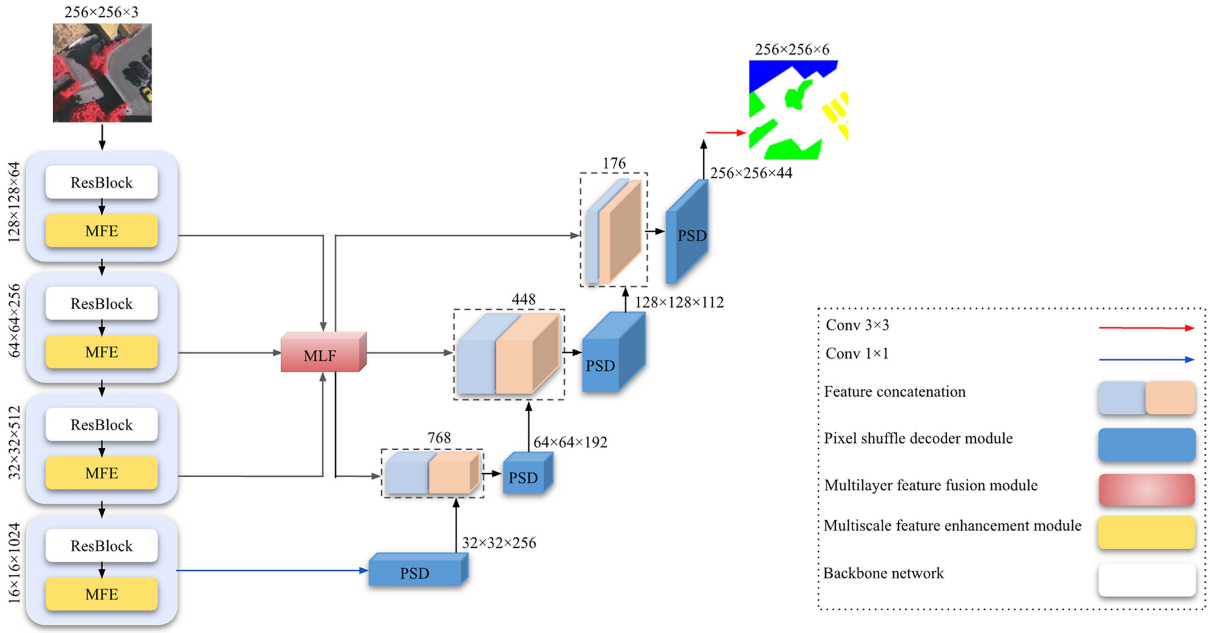

Fig. 1. Overview of the proposed segmentation method MCN, including MFE module for backbone network, MLF module as skip connections, and PSD module as a decoder.

designed a multilayer dense connectivity network as a decoder to aggregate the dual-branch semantic information. MANet [33] uses a multiscale strategy and self-attention mechanisms to aggregate relevant contextual features. ABCNet [34] uses a spatial path and a contextual path to extract contextual and fine-grained information to increase the segmentation accuracy of the network. BIBED-Seg [1] proposed a block-in-block edge detection network using an attention mechanism. In summary, an attention mechanism can improve the object region features while suppressing interference from redundant features. As a result, we designed MFE and MLF for feature enhancement and correlation modeling using an attention mechanism.

One of the fundamental concepts underlying these strategies is the multilevel context to enhance segmentation prediction. Although these methods can prevent global contextual information loss, they are computationally expensive and redundant while collecting rich and multiscale contextual information. The following shows that the MCN provides comparable results or excels on benchmark methods with fewer parameters.

III. Methodology

Fig. 1 shows MCNs three-module architecture. The first module is MFE (see Fig. 2), which takes input from ResNet50, the backbone network. Initially, MFE utilizes 3×3, 5×5, and 7×7 convolution kernels to extract feature information at low, middle, and high levels, respectively. Second, MFE refines the extracted multiscale features using an attention mechanism (see Fig. 3), which helps to decrease the influence of redundant background feature information. Finally, it uses a similarity function to capture the complementary features by establishing the correlation among different levels of feature maps. The second module is MLF (see Fig. 5), which merges the shallow and deep layers in the encoder–decoder network by using a larger receptive field and attention mechanism to deal with various layers simultaneously and improve the multiscale representation capability at a finer grained level. The third module is PSD (see Fig. 7), which alleviates the blurry edges and checkerboard artifacts while upsampling with reduced parameters. The proposed MCN is an end-to-end segmentation network that uses hierarchical processing to refine feature information and improve segmentation performance to obtain accurate semantic segmentation results.

A. MFE Module

In recent years, many efforts have been made to improve the performance of CNNs from convolutional operations and bottleneck layers to more efficient architectures. The most common way in CNNs to enlarge the receptive field is to stack smaller kernels than a single larger kernel. According to the theory of effective receptive field (ERF) [35], ERF is proportion to $O(\sqrt{l})$. ERF size grows linearly with kernel size k and sublinearly over the neural network's layers (depth). Moreover, increasing the network depth introduces optimization difficulty, where the design of larger kernels requires fewer layers to obtain larger ERF and avoids the optimization problems introduced by increasing the network depth. Therefore, we argue using a single larger kernel (i.e., 5×5 or 7×7) rather than stacking smaller kernels. Unlike natural images, urban scene images usually cover large areas; therefore, it is essential to consider the context of the objects in the image when performing semantic segmentation. Larger kernels can capture more global spatial context by considering a larger area around each pixel. This allows the network to

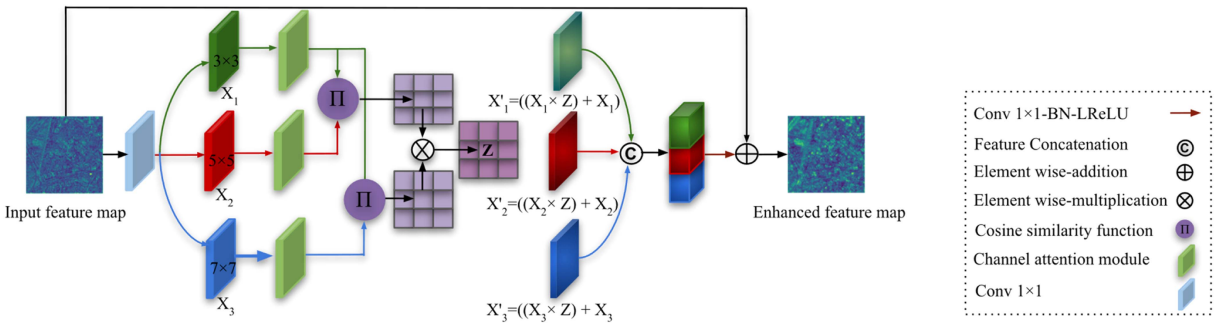

Fig. 2. Structure of MFE.

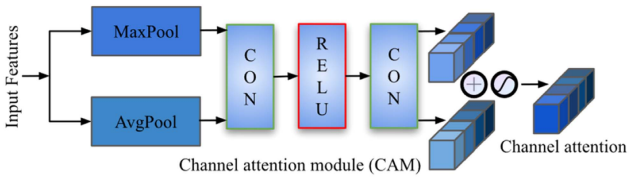

Fig. 3. Given an intermediate feature map, the operation process of the CAM is divided into two parts, including average-pooled features and max-pooled features, to capture both the overall importance and the most discriminative features of each channel. After pooling, the two 1-D vectors are sent to the multilayer perceptron network and added to generate 1-D channel attention. Then, the channel attention is multiplied by the input elements to obtain the refined feature map.

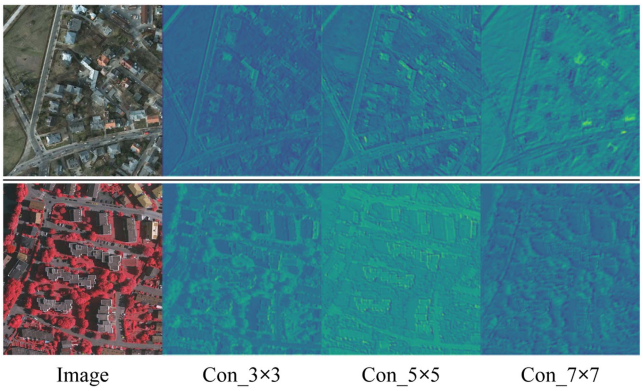

Fig. 4. Feature maps are selected from the first layer of a trained MCN by convolutional filters of different sizes. It is indicated that smaller convolutional filters are more effective at extracting fine structures, such as the sharp corners of buildings and intricate patterns of vegetation. In comparison, coarse structures respond to larger filters. (From Top to Bottom) First and second rows demonstrate the Potsdam and Vaihingen train set, respectively.

consider the relationships between different objects and features in the scene, which can help reduce interclass similarities. In addition, urban scene images contain fine-grained details, such as small objects and textures, that can be difficult to capture using larger kernels. We found that larger kernels can capture these details even better while maintaining a larger receptive field to capture global context, as shown in Fig. 4. To summarize, extracting essential contextual information about ground objects requires CNN models to process features at various scales for effective semantic segmentation. While larger kernels (i.e., 5 × 5 or 7 × 7) can detect coarser features, and smaller kernels (i.e., 3 × 3) can capture fine-scale details in urban scene images. Different kernel sizes, such as 3 × 3, 5 × 5, and 7 × 7, can effectively capture features at various scales and leverage local and global contexts in urban scene semantic segmentation. Thus, to adequately utilize the rich spatial information and improve the robustness of feature extraction among diverse and complex ground scenes, we introduce larger kernels in our work and designed MFE, as shown in Fig. 2.

To extract multiscale features from urban scene images for the input feature map, we use convolution kernels of different sizes, specifically 3 × 3, 5 × 5, and 7 × 7. In each convolution branch, we also introduce a 1 × 1 convolution to reduce the number of channels and control the calculation parameters. This allows us to efficiently extract multiscale features while minimizing computational cost. Overall, the process for extracting multiscale features can be described as follows:

$$\left.\begin{array}{l} X_1 = \vartheta_{k=3}\ \left(\vartheta_{k\ =\ 1}\left(X\right)\right) \\ X_2 = \vartheta_{k=5}\ \left(\vartheta_{k\ =\ 1}\left(X\right)\right) \\ X_3 = \vartheta_{k=7}\ \left(\vartheta_{k\ =\ 1}\left(X\right)\right) \end{array}\right\} \tag{1}$$

where $\vartheta(\cdot)$ denotes the convolution, and k is the size of different kernels.

At each convolutional layer, a group of filters expresses neighborhood spatial connectivity patterns along input channels. These filters are extremely useful in learning edges and a particular texture in the images, making CNNs produce image representations that capture hierarchical patterns. These representations can be strengthened by explicitly modeling the interdependencies of their convolutional feature channels. Channel attention provides a weight for each channel to enhance those particular channels, which are essential for feature learning. To improve the extracted multiscale features and mitigate the influence of overlapping background feature information, we utilize a channel attention module (CAM) to learn how to prioritize various features for calibration. This process of feature calibration is described as follows:

$$\mathrm{CA}\left(X_i^{'}\right) \ = \ \delta\left(\Theta\left(\mathrm{AP}\left(X_i\right)\right) + \Theta\left(\mathrm{MP}\left(X_i\right)\right)\right) \tag{2}$$

where δ represents a sigmoid function, Θ represents an MLP network, AP represents the average pooling, MP represents the max pooling, $X_{i\ =\ 1,2,3}$ represents the three input feature

maps, and CA($X^{'}_{i\,=\,1,2,3}$) represents the three generated channel attention maps.

It is essential to correctly utilize information at various scales to achieve precise semantic segmentation of urban scene images. Due to the inherent distinctions between the three branch feature maps, simply adding or concatenating them could cause more unnecessary information or data repetition in the final output. It is, therefore, essential to consider other techniques that can blend the feature maps in an efficient way while minimizing these issues. To capture the complementary features of three branches, we introduce a cosine similarity function that captures the similarity of objects among different feature maps. The MFE employs cosine similarity to quantify the relevance of the three branch feature maps. The obtained similarity scores are then normalized and used as weights to combine the maps into a single multiscale representation. This combined representation is refined and integrated into the final feature vector through optimization. For three branches of channel attention maps (i.e., $X^{'}_1, X^{'}_2,\ X^{'}_3$), we introduce the cosine similarity function to first compute the similarity between each pair of feature maps representing distinct channels, $(X^{'}_1, X^{'}_2)$ and $(X^{'}_1, X^{'}_3)$. Next, we multiply the similarity score with an elementwise product of the respective feature maps. This process is repeated for all three branches to create their corresponding feature maps. This approach allowed us to focus on specific channels and highlight their importance within the overall feature map. The feature similarity computation can be expressed as follows:

$$Z\ = \frac{X^{'}_1 \cdot X^{'}_2}{||X^{'}_1\,\|*\|\,X^{'}_2||} \otimes \frac{X^{'}_1 \cdot X^{'}_3}{||X^{'}_1\,\|*\|\,X^{'}_3||} \tag{3}$$

where $\otimes$ represents the elementwise multiplication, $(\cdot)$ represents the dot product, $(*)$ represents the cross product, and $(||\ ||)$ represents the length of two vectors $(X^{'}_1,\ X^{'}_2)$ and $(X^{'}_1,\ X^{'}_3)$, respectively.

When a neural network uses different branches to extract features from an input, the resulting feature maps may have differences due to variations. This can make it challenging to completely understand the input by correlating these feature maps. To overcome this limitation, it is essential to consider these distinctions between feature maps and find ways to integrate them more effectively. Otherwise, the accuracy and effectiveness of the network's predictions or results may suffer. To address this issue, a gate mechanism is implemented in this study. This mechanism aims to optimize similar features and merge them with feature maps at various scales. The gate unit used in this research is different from previous work [39], which usually uses the sigmoid function to restrict the value between 0 and 1. Instead, we employ a ReLU function as it accelerates the training process of MCN and avoids potential issues with gradient dispersion

$$\left.\begin{array}{l} X^{''}_1 = \sigma\left((Z \otimes X_1) + X_1\right) \\ X^{''}_2 = \sigma\left((Z \otimes X_2) + X_2\right) \\ X^{''}_3 = \sigma\left((Z \otimes X_3) + X_3\right) \end{array}\right\} \tag{4}$$

where $\sigma(\cdot)$ represents the ReLU, and $\otimes$ represents the elementwise multiplication.

For the feature map obtained by different branches, the fusion feature $X^{h\times w\times c}_c$ is obtained by feature concatenation, where $c\ =\ 3\cdot c/4$

$$X_c = \text{concat}\ (X^{''}_1, X^{''}_2, X^{''}_3). \tag{5}$$

Convolutional kernels with trainable weights are repeatedly applied to feature maps to extract new features. During feature extraction, the input of a layer depends on the weights of the previous layer. Small changes in image batches or shallow feature maps accumulate and amplify along the network depth by making training layers fit these distribution changes rather than the valuable and actual content. As a result, a neural network suffers from covariant shift, decreasing performance and training speed. Batch normalization (BN) can avoid covariant shifts by normalizing the feature map along the channel direction. BN keeps the representation capacity of the neural network by retranslating and rescaling the normalized feature map. Therefore, we use BN and LReLU to increase the numerical stability and activate the output nonlinearly. Moreover, the 1×1 convolution is introduced to get $X^{h\times w\times c}$ feature map, where $c\ =\ c$

$$\text{MFE}\,(X)\ =\ \sigma\left(\gamma(\vartheta_{k\,=\,1}\,(X_c)) + \beta\right. \tag{6}$$

where scaling (γ) and shifting (β) are two trainable parameters of the BN layer. σ denotes the activation function of LReLU.

A residual network connection is added to achieve constant training. It takes activation from one layer and feeds it to a layer far deeper in the network, facilitating the training and learning of more complex features. Residual connections allow gradients to flow backward during backpropagation. The performance will not degrade even if some data are lost during feature extraction, as it will flow through residual connections during forward propagation

$$X^E_i =\ \text{MFE}\,(X) + X \tag{7}$$

where $X^{E(h\times w\times c)}_i$ is the final enhanced feature map at encoder stages *i* (i.e., MFE-1, MFE-2, MFE-3, and MFE-4).

B. MLF Module

Features from deeper layers are high in semantic details, while features from shallow layers are less semantic but contain more local information that helps in defining object boundaries more accurately. In U-Net [12] and similar architectures that use plain skip connections, these deep and shallow features are supervised directly, pushing the network to learn better representations. Unfortunately, direct form of supervision does not prove to be very beneficial for urban scene semantic segmentation due to the following limitations: First, limited receptive field of shallow features can lead to less semantic information and introduce more noise, whereas a larger receptive field can help to get more accurate segmentation results; second, merging shallow and deep features through direct concatenation can result in a large number of parameters and computations where indirect supervision in which features from different layers are merged prior to supervision can be beneficial for urban scene semantic

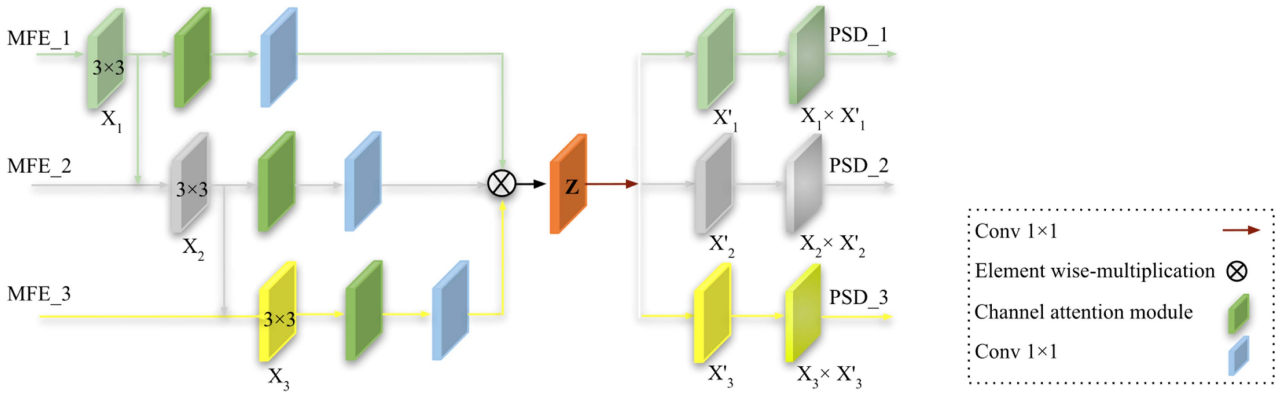

Fig. 5. Shows the structure of MLF architecture. A multiplicative mechanism correlates the different branches (MFE-1, MFE-1, and MFE-3) that separately learn representations of the input data. The gradients in one branch can be affected by the performance of other branches, as the error signals propagate through the network during training. Errors in one branch can be amplified when multiplied by the activations from another branch. This can lead to large gradients, and the final prediction will be wrong. For instance, different branches, such as MFE-2 and MFE-3, are dependent on MFE-1, which conveys that the different branches have some level of interaction and interdependence. When MFE-1 learns better representations, it can improve the other branches' performance, ultimately leading to more accurate predictions when all the branches are fused together.

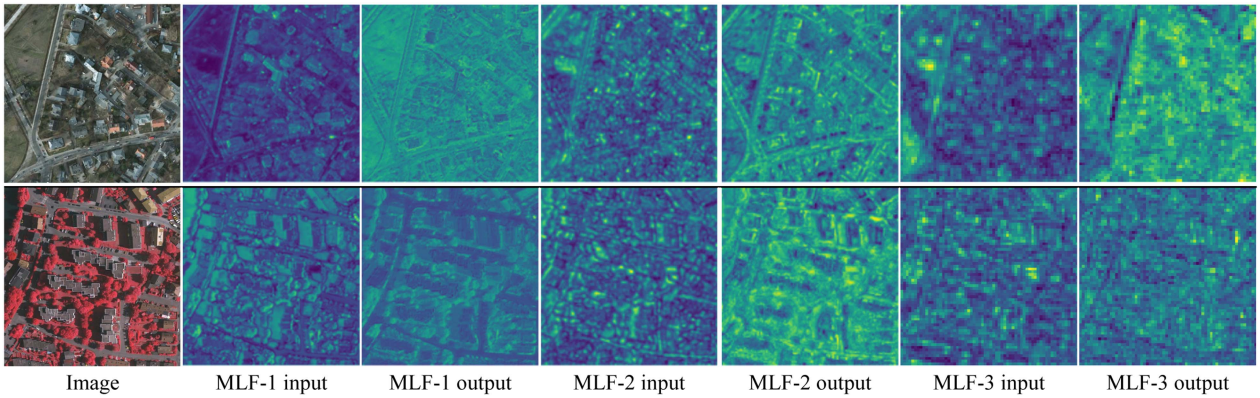

Fig. 6. (From Top to Bottom) First and second rows demonstrate the Potsdam and Vaihingen train set, respectively. The figure displays visualizations of input and output features from three MLF modules, respectively. The input image in each MLF module represents low-, mid-, and high-level features from the MFE modules, while the output image represents the processed features.

segmentation at a lower computational cost. Despite significant advancements made by prior skip connections oriented approaches; two major challenges remain.

1) Deep Coarse Features: Due to the successive downscaling operations, the feature maps over the past few layers become severely coarse (i.e., 8×8 in ResNet50), resulting in a loss of spatial resolution. Despite their limited impact on classification accuracy, these coarse feature maps can significantly affect object localization. However, when the feature maps are too coarse, it becomes difficult to precisely localize objects since their exact position within the image is unclear. Even if the network correctly classifies an image containing an object, it will struggle to localize it if the feature maps are too coarse. To tackle this problem, techniques, such as upsampling or deconvolution, can be utilized to recover some of the lost spatial information, allowing for more precise object localization even with coarse feature maps. However, this comes at the cost of increased computational cost. Therefore, we utilized an alternative upsampling method (PSD) to recover some of the lost spatial information better but with reduced computational cost.

2) Shallow and Deep Features: Without additional boundary information from input data, it is hard to refine objects' complete and sharp boundaries. Shallow layers capture low-level features, such as edges, corners, and textures, making object boundaries sharper. On the other hand, deeper layers capture higher level features that are most abstract and semantic, such as object categories. Thus, effectively combining both shallow and deep features is critical for achieving precise semantic segmentation of HR urban scene images. Using an appropriate strategy can lead to a more effective generation of feature maps, benefiting both shallow and deep features, just as MLF. Compared with plain skip connections, we designed MLF (see Fig. 5) as skip connections, where features of different layers (i.e., MFE-1, MFE-2, and MFE-3) are merged before the supervision to produce a single, high-level representation of the input data. Our approach leverages the strengths of each layer's features,

such as shallower layers capturing low-level features (i.e., edges and textures), middle layers capturing mid-level features (i.e., shapes and patterns), and deeper layers capturing high-level features (i.e., object categories). Combining these learned representations from all three MFE layers allows the network to better segment the image into various semantic classes, as shown in Fig. 6. Our MLF is helpful in specific scenarios, such as improved segmentation accuracy, robustness to noise, better generalization, more flexible design, and reduced computational complexity. MLF improves the overall network performance by emphasizing low-level features embedded in shallow layers and high-level features embedded in deep layers.

For the input feature maps from MFE layers, $X^E \in \mathbb{R}^{h_1 \times w_1 \times c_1}$, $X^E \in \mathbb{R}^{h_2 \times w_2 \times c_2}$, and $X^E \in \mathbb{R}^{h_3 \times w_3 \times c_3}$, three parallel 3×3 convolutions are utilized in a hierarchical way to achieve $X_1 \in \mathbb{R}^{h_1 \times w_1 \times c_1}$, $X_2 \in \mathbb{R}^{h_2 \times w_2 \times c_2}$, and $X_3 \in \mathbb{R}^{h_3 \times w_3 \times c_3}$ with an increased receptive field for subsequent layers, where *h*, *w*, and *c* represent the height, width, and channels of feature map, respectively. As mentioned above, the limited receptive field of shallow features can lead to less semantic information and introduce more noise. In contrast, a larger receptive field can better segment larger spatial context and global information in urban scene images. Therefore, to increase the receptive field of MFE layers, we utilized three convolutional filters of size 3×3 and hierarchically connected different filters to improve the multiscale representation ability and variation of receptive fields at a more granular level to capture details, as given in (8). Each feature map has a corresponding 3×3 convolutional filter ϑ. We denote the output of ϑ by X_1, X_2, and X_3. With each iteration of the function, the output feature map's receptive field can be increased. Each, except for $X^E \in \mathbb{R}^{h_1 \times w_1 \times c_1}$, can inherit features from all preceding feature maps. Concretely, $X^E \in \mathbb{R}^{h_1 \times w_1 \times c_1}$ continues to propagate forward to X_1. X_1 is upsampled to the same resolution and added with the following feature map $X^E \in \mathbb{R}^{h_2 \times w_2 \times c_2}$, and then fed into ϑ. So, ϑ obtains the information of both $X^E \in \mathbb{R}^{h_1 \times w_1 \times c_1}$ and $X^E \in \mathbb{R}^{h_2 \times w_2 \times c_2}$, and so on

$$\left\{ \begin{array}{l} X_1 = \sigma\left(\gamma\left(\vartheta_{k=3}\left(X^E\right)\right) + \beta\right) \\ X_2 = \sigma(\gamma(\vartheta_{k=3}\left(X_1 + X^E\right)) + \beta) \\ X_3 = \sigma(\gamma(\vartheta_{k=3}\left(X_2 + X^E\right)) + \beta) \end{array} \right\}. \tag{8}$$

The presence of noise in feature data can severely hinder the accuracy of predictive models. Our approach differs from [10], which only focuses on one layer at a time. Instead, we deal with multiple layers simultaneously in a unique way. Before fusing features from different layers, we deploy CAM to filter out irrelevant features across various layers simultaneously and improve the quality of the processed data. By leveraging the interdependence of features, channel attention allows for efficient information processing and enhanced learning capabilities within the network. Our findings suggest that this method is highly effective in improving the accuracy of predictive modeling, mainly when dealing with complex data, such as urban scene images.

For all three feature maps, CAM is exploited to achieve $X_1 \in \mathbb{R}^{h_1 \times w_1 \times c_1}$, $X_2 \in \mathbb{R}^{h_2 \times w_2 \times c_2}$, and $X_3 \in \mathbb{R}^{h_3 \times w_3 \times c_3}$ with the same shape $\mathbb{R}^{1 \times c}$ given as follows:

$$\mathrm{CA}\left(X_i'\right) = \delta\left(\Theta\left(\mathrm{AP}\left(X_i\right)\right) + \Theta\left(\mathrm{MP}\left(X_i\right)\right)\right) \tag{9}$$

where δ represents the sigmoid function, Θ represents the multilayer perceptron network, AP represents the average pooling, MP represents the max-pooling, $X_{i=1,2,3}$ represents the three input feature maps, and $\mathrm{CA}(X'_{i=1,2,3})$ represents the three generated channel attention maps.

After CAM, we employ three parallel 1×1 convolutions to achieve X_1'', X_2'', and X_3''. As a result, a latent representation $Z = X_1'' \cdot X_2'' \cdot X_3''$ is obtained using elementwise multiplication, where $c' = 256$. The latent representation obtained through elementwise multiplication results in a coupled feature representation. As a result, we can conduct channel attention for several layers simultaneously

$$Z = \vartheta_{k=1}\left(X_1'\right) \otimes \vartheta_{k=1}\left(X_2'\right) \otimes \vartheta_{k=1}\left(X_3'\right) \tag{10}$$

where $\otimes$ represents the elementwise multiplication.

Finally, the latent representation is multiplied by X_1, X_2, and X_3. Based on the latent representation, our MLF performs channel attention for each relevant layer via the multiplicative operation. The variation of receptive fields and concatenation strategy can enhance the efficiency of convolutions in processing features. This approach allows for efficient multilayer feature refinement, thereby improving the performance of the proposed network

$$\left\{ \begin{array}{l} X_1^F = X_1 \otimes (\vartheta_{k=1}(Z)) \\ X_2^F = X_2 \otimes (\vartheta_{k=1}(Z)) \\ X_3^F = X_3 \otimes (\vartheta_{k=1}(Z)) \end{array} \right\} \tag{11}$$

where $X_i^{F\ (h \times w \times c)}$ is the final fused feature map at skip connections' stages *i* (i.e., MLF-1, MLF-2, and MLF-3).

C. PSD Module

PS [23] was initially introduced for SISR, where we aim to train a CNN that generates super-resolved images at the original resolution. Without adding extra parameters and computation costs, PS provides another way to fit semantic segmentation for large-scale urban scene images under memory limits. On this basis (see Fig. 7), we design the PSD module to leverage the advantage of the MCN in capturing multiscale information. PSD upsamples by rearranging pixels of the feature map and reducing the number of channels by four times, which significantly reduces parameters of the subsequent convolution, followed by a composite function comprising three different operations (3×3, BN, and PReLU) and concatenation with corresponding MLF layers from the skip connections path. To achieve dense-level prediction, high-level features generated at the last encoder stage MFE_4 are upsampled and concatenated with skip connections stage MLF_3. Take PSD_3 as an example; Fig. 1 shows how to generate feature maps. First, refined feature maps of corresponding skip connections path MLF_3 are concatenated with decoder stage PSD_3. Next, a PS operation with a scaling factor $r = 2$ is applied. It rearranges pixels of feature map of the shape $X^{h \times w \times r2}$ to an upsampled feature map of shape $X^{rh \times rw \times c/r2}$ without the

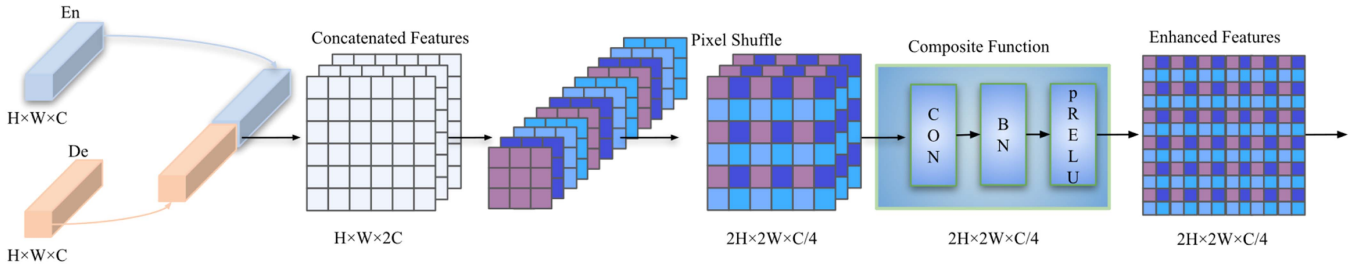

Fig. 7. Structure of PSD.

loss of information and reduced number of parameters

$$\bar{X}^{\,2h\times 2w\times c/4}_{\mathrm{De}(s)} = \mathrm{PS}\left(X^{h\times w\times c}_{\mathrm{De}(s-1)} \oplus X^{h\times w\times c}_{\mathrm{MLF}(s)}\right) \tag{12}$$

where PS is a pixel shuffle operator that rearranges pixels in a periodic shuffling manner to upscale the features map. By retaining feature information, PS can better alleviate the edge blur and artifacts caused by information loss. $X^{h\times w\times c}_{\mathrm{MLF}(s)}$ is the current MLF stage, $X^{h\times w\times c}_{\mathrm{De}(s-1)}$ is the previous decoder stage, and $\bar{X}^{\,2h\times 2w\times c/4}_{\mathrm{De}(s)}$ is an intermediate upsampled feature map with PS at stage s.

Upsampled feature maps are further convolved with one standard convolution 3×3 to lessen the aliasing distortion, and nonlinearity is added to generate the final feature map at the decoder stage s

$$X^{2h\times 2w\times c/4}_{\mathrm{De}(s)} = \sigma(\gamma(\bar{X}^{\,2h\times 2w\times c/4}_{\mathrm{De}(s)} * \vartheta_{k=1}) + \beta) \tag{13}$$

where $X^{2h\times 2w\times c/4}_{\mathrm{De}(s)}$ is the final generated feature map at decoder stages s (i.e., PSD-1, PSD-2, PSD-3, and PSD-4). γ and β are two trainable parameters of the BN layer. σ denotes the activation function of PReLU.

IV. Experimental Results and Analysis

A. Semantic Segmentation Datasets

The performance of MCN is assessed using three publicly available datasets, Potsdam, Vaihingen, and DeepGlobe.

1) Potsdam Dataset: The Potsdam dataset consists of 38 HR aerial images, each with an average size of 6000 × 6000 pixels and a ground sampling distance (GSD) of 5 cm. Potsdam RGB images are annotated with six distinct landscape classes: impervious surface (road), building, low vegetation, trees, car, and clutter, where clutter class is not considered in the assessment. For Potsdam, we split images into training, validation, and testing sets with 23, 1, and 14 images, respectively.

2) Vaihingen Dataset: The Vaihingen dataset consists of 33 HR aerial images, each with an average size of 2494 × 2064 pixels and a GSD of 9 cm. Vaihingen IRRG images are annotated with six distinct landscape classes: impervious surface (road), building, low vegetation, trees, car, and clutter, where clutter class is not considered in the assessment. For Vaihingen, we split images into training, validation, and testing sets with 15, 1, and 17 images, respectively.

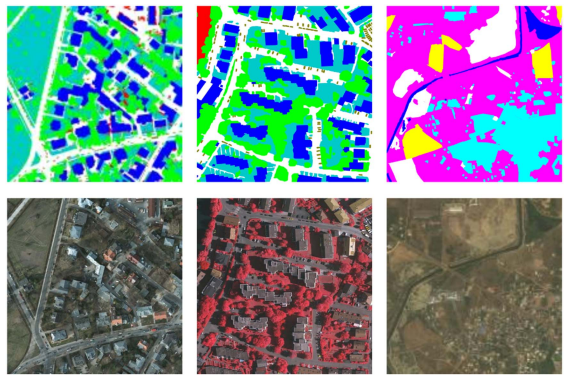

Fig. 8. Sample images from land-cover classification datasets, Potsdam, Vaihingen, and DeepGlobe, respectively.

3) DeepGlobe Dataset: The DeepGlobe dataset consists of 803 HR satellite images, each with an average size of 2448 × 2448 pixels and a GSD of 50 cm pixels. These images are annotated with seven distinct landscape classes: forest land, urban land, barren land, agriculture land, rangeland, water, and unknown, where unknown class is not considered in the assessment. Following [36], we split images into training, validation, and testing sets with 454, 207, and 142 images, respectively.

B. Evaluation Metrics

We use the following assessment metrics to evaluate the performance MCN, overall accuracy (OA), mean intersection over union (mIoU), and mean $F1$-score (mF1), which can be defined as follows:

$$\mathrm{OA} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}} \tag{14}$$

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}} \tag{15}$$

$$F1 = 2 \times \frac{\mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{16}$$

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \quad \mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{17}$$

where TP, TN, FP, and FN denote the true positive, true negative, false positive, and false negative, respectively.

TABLE I
ABLATION STUDIES OF MFE, MLF, AND PSD ON ISPRS DATASETS

MFE	MLF	PSD	Potsdam mF1 (%)	Potsdam OA (%)	Potsdam mIoU (%)	Vaihingen mF1 (%)	Vaihingen OA (%)	Vaihingen mIoU (%)
X	X	X	88.46	88.57	79.67	83.82	86.99	72.79
✓	X	X	91.74	91.79	84.87	87.84	89.19	78.58
X	✓	X	90.81	91.05	83.67	88.06	89.11	79.31
X	X	✓	89.88	90.16	81.44	86.87	88.56	77.36
✓	✓	X	91.98	92.03	85.30	88.96	89.99	80.39
✓	✓	✓	92.85	93.51	86.81	89.25	90.18	80.86

TABLE II
MODEL COMPLEXITY OF MFE, MLF, AND PSD MODULES ON ISPRS DATASETS

MFE	MLF	PSD	Potsdam Train Time (s)/epoch	Potsdam Inference (s)	Vaihingen Train Time (s)/epoch	Vaihingen Inference (s)	Param	Flops	Model Size
X	X	X	100	0.47	92	0.07	32.52 M	10.73 G	130.4 MB
✓	X	X	150	1.32	115	0.17	22.41 M	30.00 G	89.60 MB
X	✓	X	135	2.80	110	0.42	13.30 M	24.71G	54.60 MB
X	X	✓	90	1.08	72	0.16	10.65 M	09.72 G	44.60 MB
✓	✓	X	156	1.24	130	0.13	22.23 M	39.91 G	92.20 MB
✓	✓	✓	123	1.20	102	0.12	19.56 M	18.51 G	78.90 MB

C. Implementation Details

The experiments were conducted on a single NVIDIA RTX 3090 GPU, utilizing the PyTorch framework. We chose the U-Net [12] as the baseline model and employed ResNet50 [6] as the backbone network for the MCN. In our work, we utilized only the first three bottleneck layers of the pretrained ResNet50 to reduce the number of trainable parameters. For optimization, we employed the Adam optimizer with AMSGrad [37], using a weight decay of 2×10^{-5}. In addition, we applied polynomial decay (L) to $1-$ cur_iter/max_iter)$^{0.9}$, where the maximum number of iterations was set to 10^8. We also set $2\times L$ for all bias parameters. The initial learning rate was set to $8.5 \times 10^{-5}/\sqrt{2}$ for the ISPRS dataset and $8.5 \times 10^{-4}/\sqrt{2}$ for the DeepGlobe dataset. We implemented a stepwise schedule method to decrease the learning rate and improve the training process. For the ISPRS dataset, we reduced the learning rate by a factor of 0.85 after every 15 epochs. Similarly, for the DeepGlobe dataset, we employed a reduction in the learning rate by a factor of 0.85 after every 4 epochs. During training and validation, we used randomly sampled 5000 patches of size 256 $\times$ 256 from the ISPRS and DeepGlobe datasets. These patches were augmented by mirroring and flipping, each with a 50% probability. To improve the predictions, we employed test time augmentation (TTA) by averaging the predictions of overlapping TTA regions. To handle the problem of imbalanced data in the ISPRS dataset, we utilized a cross-entropy loss function that incorporated median frequency balancing weights, as described by the equation. Conversely, the DeepGlobe dataset employed a cross-entropy loss function

$$L = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} l_c^{(n)} \log\left(p_c^{(n)}\right) W_c \tag{18}$$

$$W_c = \frac{\text{median}(\{f_c | c \in C\})}{f_c} \tag{19}$$

where N denotes the number of samples in a minibatch, W_c is the class weight, fc is the pixel frequency, ${p_c}^{(n)}$ is the probability of sample n, ${I_c}^{(n)}$ is the label of sample n in class c, and C is set for all classes.

D. Ablation Study

To verify the effectiveness of MCN, we conducted extensive ablation experiments on ISPRS datasets using different settings. Tables I and II present the ablation experiments of different modules of MCN, while Tables III and IV present the ablation experiments of MCN using different upsampling methods.

1) Effectiveness of MFE Module: MFE is designed for extracting rich spatial information among various and complex urban objects while suppressing background noise and enabling correlation among different levels of feature maps. Table I presents that compared with the baseline, MFE module increases the mF1 by 3.28/4.02%, OA by 3.22/2.20%, and mIoU by 5.20/5.79% on ISPRS datasets, respectively. Fig. 9 shows the visualization results of MFE module on ISPRS datasets, which produced the overall better segmentation results, but there are instances where some pixels are misclassified, such as low veg category.

2) Effectiveness of MLF Module: MLF is designed as skip connections to prevent the loss of fine- and coarse-grained details at shallower and deeper layers. MLF makes the utmost low-level features embedded in shallow layers and high-level features embedded in deep layers by improving the overall network's performance. Table I presents that compared with the baseline, MLF module increases the mF1 by 2.35/4.24%, OA by 2.48/2.12%, and mIoU by 4.00/6.52% on ISPRS datasets, respectively. Furthermore, when comparing MFE+MLF with the MLF module, there was a further improvement of 0.24/1.12%

TABLE III
ABLATION STUDIES OF MCN WITH DECONVOLUTION, BILINEAR, AND PSD-BASED UPSAMPLING METHODS ON ISPRS DATASETS

MFE	MLF	Upsampling	Potsdam			Vaihingen		
			mF1 (%)	OA (%)	mIoU (%)	mF1 (%)	OA (%)	mIoU (%)
✓	✓	Deconvolution	91.49	91.49	84.42	88.38	89.50	79.43
✓	✓	Bilinear	91.98	92.03	85.30	88.96	89.99	80.39
✓	✓	PSD	92.85	93.51	86.81	89.25	90.18	80.86

TABLE IV
MODEL COMPLEXITY OF MCN WITH DECONVOLUTION, BILINEAR, AND PSD-BASED UPSAMPLING METHODS ON ISPRS DATASETS

MFE	MLF	Upsampling	Potsdam		Vaihingen				
			Train Time (s)/epoch	Inference (s)	Train Time (s)/epoch	Inference (s)	Param	Flops	Model Size
✓	✓	Deconvolution	104	1.31	94	0.17	23.12 M	18.51 G	93.2 MB
✓	✓	Bilinear	108	1.24	99	0.13	22.23 M	22.23 G	92.2 MB
✓	✓	PSD	97	1.20	88	0.12	19.56 M	18.51 G	78.9 MB

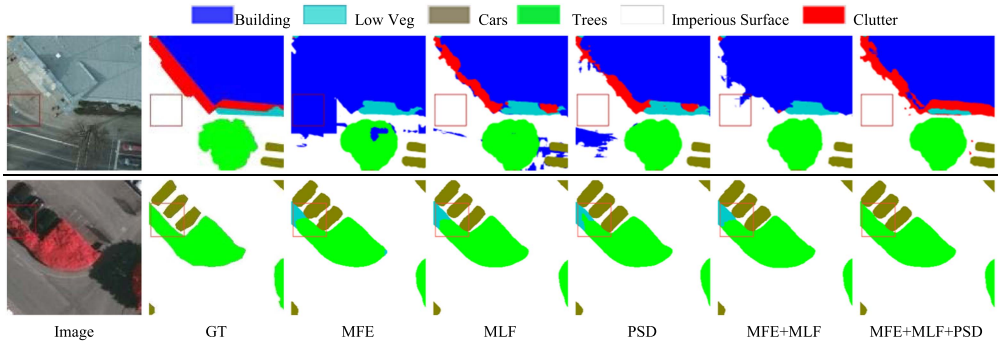

Fig. 9. Qualitative comparisons of MFE, MLF, and PSD modules. (From Top to Bottom) First and second rows demonstrate the Potsdam and Vaihingen test dataset, respectively.

in mF1, 0.24/0.80% in OA, and 1.12/1.81% in mIoU on ISPRS datasets, respectively. Fig. 9 shows the visualization results of MLF and MFE+MLF modules on ISPRS datasets. Compared with the baseline method, MLF demonstrates noticeable enhancements in segmentation results. Similarly, MFE+MLF significantly improves over MLF by effectively reducing the number of misclassified pixels while generating sharp boundaries.

3) Effectiveness of PSD Module: PSD is designed for the decoder to improve the resolution of the feature maps and fully fuse feature information of different receptive fields while better fixing blurry edges and checkerboard artifacts with the reduced number of parameters. Table I presents that compared with the baseline, PSD module increases the mF1 by 1.42/3.05%, OA by 1.59/1.57%, and mIoU by 1.77/4.57% on ISPRS datasets, respectively. Compared with MFE+MLF and PSD modules, MFE+MLF+PSD further improves the semantic segmentation accuracy, with mF1, OA, and mIoU reaching 92.85/89.25%, 93.51/90.18%, and 86.81/80.86% on ISPRS datasets, respectively. Fig. 9 shows the visualization results of PSD and MFE+MLF+PSD modules on ISPRS datasets. Compared with the baseline method, PSD demonstrates noticeable enhancements in segmentation results. Similarly, MFE+MLF+PSD exhibits significant improvements over MFE+MLF and PSD by classifying the correct number of classes while reducing the checkerboard artifacts and blurry edges.

4) MCN With Different Upsampling Methods: Deconvolution and bilinear are the most common upsampling methods for recovering spatial information from convolution or maxpoling layers. In contrast to both upsampling techniques, we designed PSD based on subpixel convolution, which can better fix blurry edges and checkerboard artifacts with fewer parameters while further improving network speed and accuracy. Table III presents the quantitative results of ISPRS datasets, respectively. Compared with deconvolution, MCN with PSD improved the segmentation accuracy in terms of mF1 by 1.36/0.87%, OA by 2.02/0.68%, and mIoU by 2.39/1.43%. While compared with bilinear, MCN with PSD improved the segmentation accuracy in terms of mF1 by 0.87/0.29%, OA by 1.48/0.19%, and mIoU by 1.51/0.47%. Fig. 10 shows that MCN with PSD can better alleviate the edge blur and artifacts caused by information loss than other upsampling methods.

5) Model Complexity: Considering that the complexity of the model is significant to assess the metric of a framework, we report the training time for each epoch, inference time, parameters, flops, and model size of different modules and upsampling methods in Tables II and IV, which demonstrates that the design of MCN is computationally efficient. Table II presents the model complexity of MCN using different modules. Tables I and II present that compared with the baseline, MCN outperforms ISPRS datasets in terms of mF1 by 4.39/5.43%, OA by 4.94/3.19%, and mIoU by 7.14/8.07% while reducing

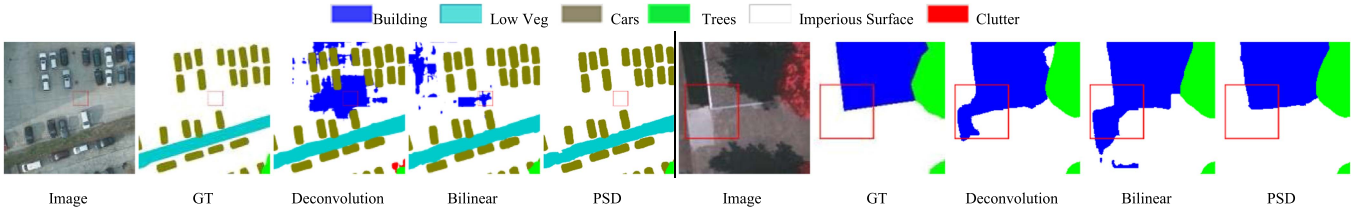

Fig. 10. Qualitative comparisons of MCN with deconvolution, bilinear, and PSD-based upsampling methods. (From Left to Right) row demonstrates the Potsdam and Vaihingen test dataset, respectively.

TABLE V
EXPERIMENTAL RESULTS OF DIFFERENT METHODS ON POTSDAM TEST SET

			F1_score (%)							
Models	Publication	Backbone	Imp. Surf.	Building	Low Veg.	Tree	Car	mF1 (%)	OA (%)	mIoU (%)
U-Net [12]	MICCAI15	ResNet50	90.39	92.20	84.59	83.96	91.72	88.57	88.46	79.67
SCAttNet [27]	GRSL20	ResNet50	90.04	94.05	84.05	79.75	89.06	78.42	87.97	68.31
EFCNet-UNet [18]	GRSL21	-	74.01	78.42	66.78	61.34	76.47	80.49	81.77	80.49
STransFuse [32]	JSTARS21	ResNet34	88.25	91.46	79.04	85.51	77.14	78.67	86.07	66.66
MANet [33]	TGRS21	ResNet50	91.86	93.81	88.52	88.57	92.78	91.10	91.11	83.74
MAResU-Net [20]	GRSL21	ResNet50	91.30	93.34	87.77	88.56	91.21	90.44	90.61	82.60
ABCNet [34]	ISPRS21	ResNet50	88.28	90.37	81.65	75.51	91.68	85.50	85.30	84.64
WiCoNet [31]	TGRS21	ResNet50	92.50	96.53	87.03	87.31	95.13	91.70	90.24	84.93
MSCA-Net [21]	J. Sens21	-	91.70	94.10	84.50	89.20	91.80	90.30	89.30	76.10
FSHRNet [17]	TGRS22	HR-Net	92.16	96.59	86.12	86.94	91.53	90.67	89.82	83.16
MACU-Net [22]	GRSL22	-	88.28	90.37	81.65	75.51	91.68	85.50	85.30	84.64
SERNet [28]	Remote Sens22	ResNet50	91.75	95.30	73.74	82.26	92.13	87.04	90.29	76.76
ColNet [25]	TGRS22	ResNet50	90.10	93.85	82.83	82.30	89.79	87.77	88.49	78.50
C-PNet [38]	TGRS23	-	92.10	95.80	83.80	86.80	91.30	91.50	90.10	-
RSSFormer [50]	TIP23	RSS-B	93.82	96.04	86.87	86.75	**96.82**	89.81	91.25	-
EG-UNet [40]	TGRS23	-	92.56	85.17	86.07	**91.14**	92.63	89.51	89.42	81.17
EMRT [42]	TGRS23	ResNet50	90.87	94.86	84.12	85.23	90.32	73.62	88.12	73.59
LiANet [41]	TGRS23	ResNet50	**95.33**	96.10	79.79	77.14	91.37	87.95	91.99	-
UperNet [44]	arXiv23	ViT-G12	92.76	96.93	85.88	89.02	96.02	92.12	92.58	-
MCN		ResNet50	95.13	**97.54**	**89.67**	89.57	92.35	**92.85**	**93.51**	**86.81**

Boldface indicates the best performance.

the number of parameters by 66% and model size by 65%. Table IV presents the model complexity of MCN using different upsampling methods. These results indicate that compared with deconvolution, MCN with PSD reduces the number of parameters and model size by 18%.

Compared with deconvolution, MCN with PSD has the same flops but is 7 s faster and requires less training time of 11 s. Compared with bilinear, MCN with PSD is 11 s faster and reduces the number of parameters by 13%, flops by 20%, model size by 16%, and training time of 4.

E. Comparison Methods

To conduct a quantitative comparison, we have carefully selected a comprehensive set of benchmark methods that are specifically designed for semantic segmentation of urban scene imagery. Note that all experimental findings are provided by the source code or the author in Tables V–VIII

1) CNN-Based Context Aggregation Networks: Collaborative network with PS layer (ColNet) [25], class perception network (C-PNet) [38], ensemble full CNN-based network (EFCNet-UNet) [18], one-shot neural architecture search for a backbone network (RSBNet) [39], feature-selection network with hypersphere embedding (FSHRNet) [17], and deep feature enhancement method for land cover (EG-UNet) [40].

2) CNN-Based Attentional Networks: Lightweight attention network (LiANet) [41], segmentation network with spatial and channel attention (SCAttNet) [27], dual-channel scale-aware network with position and channel attention (DSPCANet) [29], attentive bilateral contextual network (ABCNet) [34], multi-attention network (MANet) [33], and squeeze and excitation residual network (SERNet) [28].

3) Transformer and With or Without CNN-Based Context Enhancement Networks: Fusing swin transformer and CNN-based network (STransFuse) [32], wide-context transformer network (WiCoNet) [31], positioning guidance network (PGNet) [30], enhancing multiscale representations with transformer network (EMRT) [42], distilling segmenters from CNNs and transformers (DSCT) [43], a billion-scale foundation model (UperNet) [44], and foreground saliency enhancement network (RSSFormer) [45].

4) Segmentation Models Based on Redesigned Skip Connections: Multistage attention network (MAResU-Net) [20], semantic segmentation network using multiscale skip

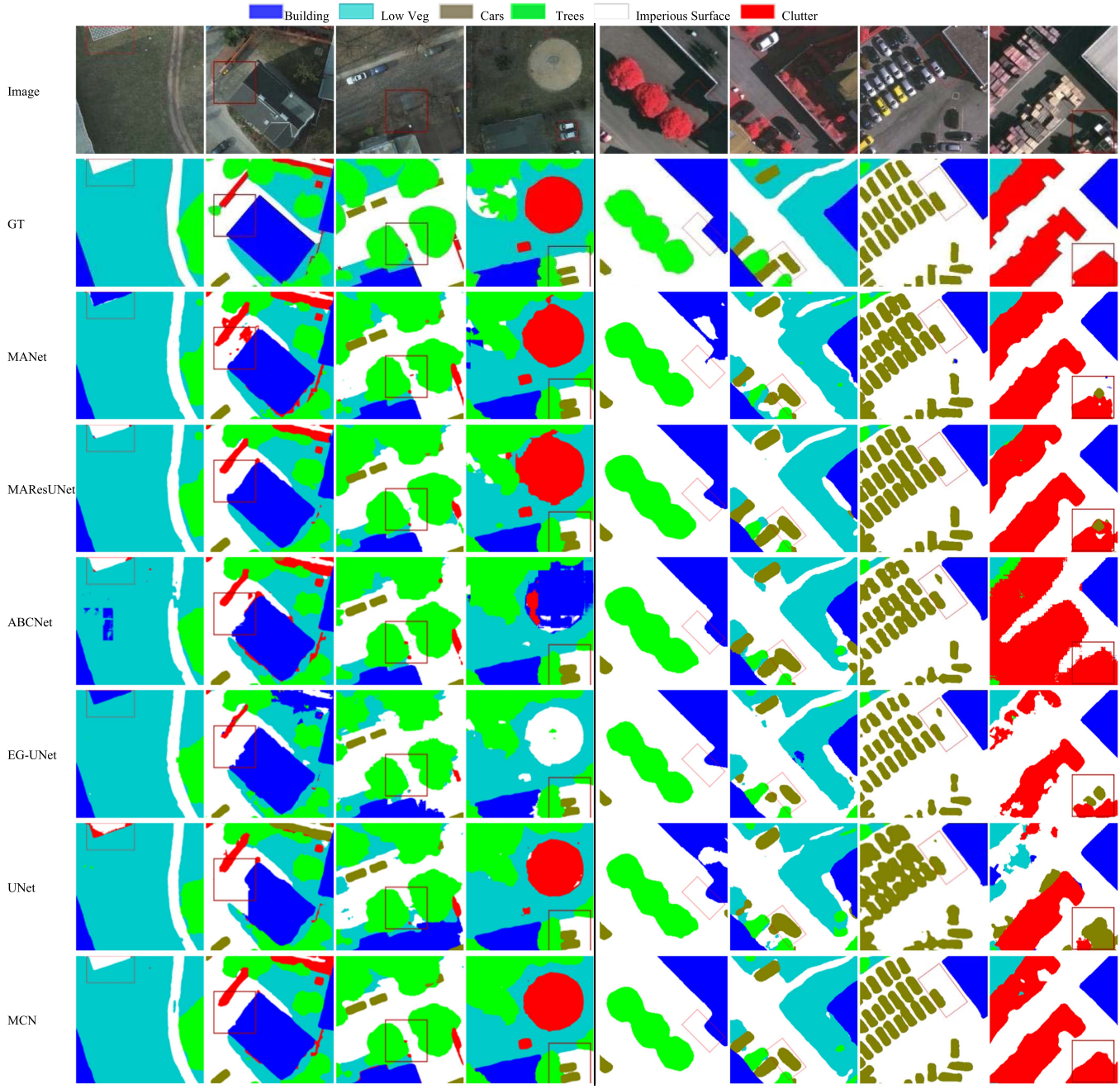

Fig. 11. Visualization results of different land-cover classification methods on ISPRS datasets. (From Left to Right) First four and second four columns demonstrate the Potsdam and Vaihingen test dataset, respectively.

connection (MSCA-Net) [21], segmentation network for fine-resolution remotely sensed images (MACU-Net) [22] and 2-D transformer model (2DsegFormer) [19].

5) Segmentation Models Designed for Ultrahigh-Resolution Images (UHR): Collaborative global–local network (GLNet) [36], progressive semantic segmentation network (MagNet) [45], integrating shallow and deep features network (ISDNet) [46], patch proposal network (PPN) [47], image segmentation via locality-aware contextual correlation network (LCC) [48], and one model is enough for image semantic segmentation (OME) [49].

F. Comparative Study

1) Results on Potsdam Dataset: Table V presents the experimental results of different methods on test sets of Potsdam. Our model on Potsdam outperforms the existing land-cover classification methods with an OA of 93.51%, mF1 of 92.85%, and mIoU of 86.81%. Regarding details, our model ranks first in the $F1$-score for (building and low veg) and second for (imperious surface and trees) subclasses. Fig. 11 shows the visualization results of MANet [33], MAResUNet [20], ABCNet [34], EG-UNet [40], U-Net [12], and proposed method MCN on test set of Potsdam. Compared with popular semantic segmentation

TABLE VI
EXPERIMENTAL RESULTS OF DIFFERENT METHODS ON THE VAIHINGEN TEST SET

			F1_score (%)							
Models	Publication	Backbone	Imp. Surf.	Building	Low Veg.	Tree	Car	mF1 (%)	OA (%)	mIoU (%)
U-Net [12]	MICCAI15	ResNet50	89.67	91.71	77.91	86.78	73.03	83.82	86.99	72.79
SCAttNet [27]	GRSL20	ResNet50-	89.13	90.32	80.04	80.31	70.50	82.06	85.47	70.20
EFCNet-UNet [18]	GRSL21	-	77.50	83.71	62.75	75.71	51.08	81.87	85.46	70.14
STransFuse [32]	JSTARS21	ResNet34	88.25	91.46	79.04	85.51	77.14	78.67	86.07	66.66
MANet [33]	TGRS21	ResNet50	92.13	94.63	81.86	88.95	86.98	88.91	89.90	80.31
MAResU-Net [20]	GRSL21	ResNet50	92.28	94.93	81.93	88.95	84.99	88.62	90.03	79.88
ABCNet [34]	ISPRS21	ResNet50	90.58	92.61	80.22	87.95	75.41	85.35	88.23	75.01
DSPCANet [29]	JSTARS21	-	91.37	93.37	79.60	81.21	76.56	87.19	90.13	77.66
MACU-Net [22]	GRSL22	-	90.34	92.61	79.48	87.68	79.68	85.96	88.07	75.77
SERNet [28]	Remote Sens22	Resnet50	91.79	91.64	81.68	81.64	75.68	84.48	88.19	72.69
FSHRNet [17]	TGRS22	HR-Net	90.34	94.31	81.37	87.19	80.10	86.66	88.38	76.86
PGNet [30]	Remote Sens22	ResNet50	90.61	93.54	72.39	84.44	81.31	72.56	86.32	62.67
2DsegFormer [19]	TGRS22	MyT-B0	90.96	94.50	81.44	87.20	81.29	87.08	88.85	77.49
EG-UNet [40]	TGRS23	-	**92.87**	87.74	78.92	**90.84**	81.48	86.37	88.16	76.40
EMRT [42]	TGRS23	ResNet50	89.16	92.68	78.69	86.54	80.12	81.38	86.97	69.79
DSCT [43]	TGRS23	ResNet101-SwinT	87.26	91.09	74.87	85.65	72.60	82.29	84.90	70.55
RSBNet [39]	Neurocomputing23	RSBNet-large	92.52	**95.18**	81.28	89.78	76.30	86.77	89.25	-
MCN		ResNet50	92.52	94.96	**82.40**	88.98	**87.40**	**89.25**	**90.18**	**80.86**

Boldface indicates the best performance.

TABLE VII
EXPERIMENTAL RESULTS OF DIFFERENT METHODS ON DEEPGLOBE TEST SET

			IoU_score (%)								
Models	Publication	Backbone	Urban	Agri.	Range.	Forest	Water	Barren	mIoU (%)	OA (%)	mF1 (%)
U-Net [12]	MICCAI15	ResNet50	77.96	82.92	44.19	78.99	79.85	61.64	70.93	89.10	80.41
MANet [33]	TGRS21	ResNet50	78.76	84.54	49.37	78.09	80.12	57.00	71.31	87.13	78.05
MAResU-Net [20]	GRSL21	ResNet50	77.81	87.08	48.33	79.83	80.04	**62.72**	72.64	88.01	79.16
ABCNet [34]	ISPRS21	ResNet50	76.12	81.68	43.82	78.06	78.54	58.68	69.48	86.75	77.98
EG-UNet [40]	TGRS23	-	71.67	85.22	35.14	78.60	67.32	50.42	64.73	84.91	77.18
MCN		ResNet50	**79.76**	**87.58**	**50.13**	**80.50**	**81.72**	62.69	**73.73**	**90.56**	**89.60**

Boldface indicates the best performance.

methods, we can observe that MCN can better handle situations with a shadow or complex texture and generate complete shapes of objects, such as buildings, trees, cars, and imperious surfaces with clear boundary separating objects.

2) Results on Vaihingen Dataset: Table VI presents the experimental results of different methods on test sets of Vaihingen. Our model on Vaihingen outperforms the existing land-cover classification methods with an OA of 90.18%, mF1 of 89.25%, and MIoU of 80.86%. Regarding details, our model ranks first in the $F1$-score for (low veg and car), and second for (building and imperious surface) subclasses. MCN has better capability in handling highly imbalanced classes, such as a car with an increased receptive field in Table VI. Notably, on Vaihingen, the $F1$-score for the car class is 87.40%. The visualization results, as shown in Fig. 11, compare the performance of MCN with five other semantic segmentation methods (MANet [33], MAResNet [20], ABCNet [34], EG-UNet [40], and U-Net [12]), on the test set of the Vaihingen. The findings indicate that MCN outperforms these methods in handling challenging scenarios involving shadows or intricate textures, generating precise shapes for objects, such as buildings, trees, low veg, and roads, and accurately distinguishing between objects with clear boundaries (i.e., cars).

3) Results on DeepGlobe Dataset: Tables VII and VIII present the experimental results of different methods on test sets of DeepGlobe. Our model on DeepGlobe outperforms the existing land-cover classification methods with an mIoU of 73.73%, OA of 90.56%, and mF1 of 89.60%. Regarding details, our model ranks first in the IoU score for all subclasses except the barren class. Fig. 12 shows the visualization results of MANet [33], MAResUNet [20], ABCNet [34], EG-UNet [40], U-Net [12], and proposed method MCN on the test set of DeepGlobe. DeepGlobe consists of classes with similar visual features and irregular shapes, making their classification challenging. However, our MCN accurately distinguishes between these land-cover categories, resulting in precise segmentation

TABLE VIII
QUANTITATIVE COMPARISON OF DIFFERENT UHR METHODS ON DEEPGLOBE TEST SET

Models	Publication	mIoU (%)
GLNet [36]	CVPR19	71.60
PPN[47]	AAAI20	71.90
MagNet-Fast [45]	CVPR21	71.85
MagNet [45]	CVPR21	72.96
LCC [48]	ICCV21	73.50
ISDNet [46]	CVPR22	73.30
OME [49]	TGRS23	68.28
MCN		**73.73**

Boldface indicates the best performance.

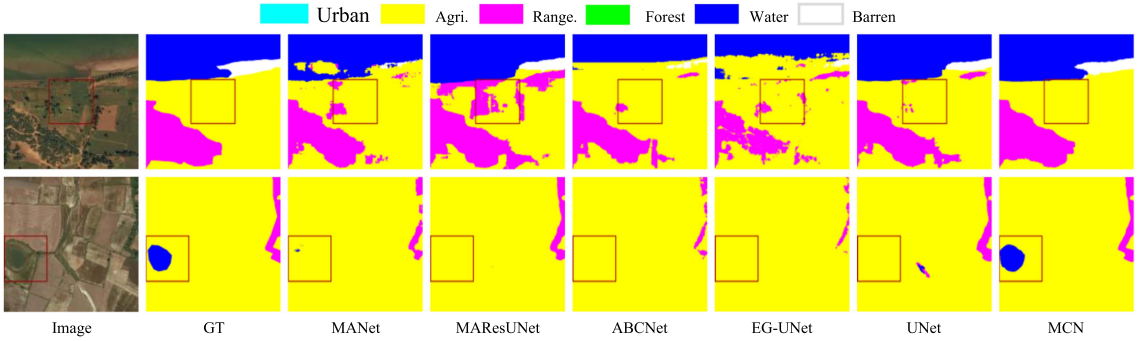

Fig. 12. Visualization results of different land-cover classification methods on the DeepGlobe test dataset.

TABLE IX
ISPRS PERFORMANCE COMPARISON OF VARIOUS METHODS USING RESNET50 BACKBONE

Models	Parameters	OA (%)	
		Potsdam	Vaihingen
MAResU-Net [20]	102 M	90.61	90.03
EMRT [42]	54.85 M	88.12	86.97
ABCNet [34]	45.65 M	85.30	88.23
PGNet [30]	42.67 M	-	86.32
WiCoNet [31]	38.24 M	90.24	-
MANet [33]	35.86 M	91.11	89.90
U-Net [12]	32.52 M	88.46	86.99
ColNet [25]	28.57 M	88.49	-
SCAttNet [27]	24.62 M	87.97	85.47
LiANet [41]	24.59 M	91.99	**-**
MCN	**19.56 M**	**93.51**	**90.18**

Boldface indicates the best performance.

results. These visualized results further convincingly validate the effectiveness of our MCN on satellite images.

4) Discussion: Our model consistently outperforms CNN and transformer-based networks in terms of OA, mF1, and mIoU on well-established benchmark datasets, such as ISPRS and DeepGlobe, demonstrating competitive performance despite utilizing fewer parameters. Using MFE, MCN can effectively capture local and global contextual information. MFE enables the extraction and encoding of features from multiple levels of abstraction, allowing the model to perceive intricate details and holistic scene understanding. Using MLF, MCN can effectively combine features extracted from different levels of abstraction to improve the accuracy and performance of the model. Fusing these features leads to the generation of sharper boundaries, distinguishing between different objects or classes in an image. In addition, our model benefits from PSD. Incorporating PSD enhances the model's ability to generate precise and artifact-free segmentations while minimizing computational requirements.

5) ISPRS Datasets Performance Comparison: Table IX presents a comparison of various methods that employ the ResNet50 backbone. The comparison is based on two key aspects: overall accuracy and the number of parameters. Compared with these methods, MCN achieved an OA of 93.51% on Potsdam and 90.18% on Vaihingen datasets with 19.56 million parameters, maintaining both the number of parameters and high accuracy simultaneously.

V. CONCLUSION

This article proposes a novel MCN with three fundamental modules to solve the interclass similarities, intraclass variations, scale-related inaccuracies, and high computational complexity issues. MFE is introduced as a feature enhancement module for the backbone network to identify the local and global context of ground objects while suppressing the background noise and capturing complementary features by enabling correlation among different levels of feature maps. MLF is introduced as skip connections where features from different MFE layers are merged before the supervision to produce a single high-level representation of the input data by leveraging the strength of each layer in capturing different levels of features. PSD is introduced as a decoder, which can better fix blurry edges and checkerboard artifacts with fewer parameters while improving network speed and accuracy. Ablation studies and comparative experiments conducted on the ISPRS datasets demonstrate the effectiveness of the proposed method. On all three Potsdam, Vaihingen, and DeepGlobe, MCN achieves the best performance compared with the existing land-cover classification models with fewer parameters.

REFERENCES

[1] B. Sui, Y. Cao, X. Bai, S. Zhang, and R. Wu, "BIBED-Seg: Block-in-block edge detection network for guiding semantic segmentation task of high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1531–1549, 2023, doi: 10.1109/JSTARS.2023.3237584.

[2] X. Zheng, Q. Ma, L. Huan, X. Xie, H. Xiong, and J. Gong, "Semantic-aware region loss for land-cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4139–4152, 2023, doi: 10.1109/JSTARS.2023.3265365.

[3] Z. Lv, F. Wang, G. Cui, J. A. Benediktsson, T. Lei, and W. Sun, "Spatial–spectral attention network guided with change magnitude image for land cover change detection using remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4412712, doi: 10.1109/TGRS.2022.3197901.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, pp. 84–90, 2017.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[7] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.

[8] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[9] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[10] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

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

[12] 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.*, 2015, pp. 234–241.

[13] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested U-Net architecture for medical image segmentation," in *Proc. 4th Int. Workshop Deep Learn. Med. Image Anal.*, 2018, pp. 3–11.

[14] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[15] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[16] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.

[17] H. Xu, X. Tang, B. Ai, F. Yang, Z. Wen, and X. Yang, "Feature-selection high-resolution network with hypersphere embedding for semantic segmentation of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4411915.

[18] L. Chen, X. Dou, J. Peng, W. Li, B. Sun, and H. Li, "EFCNet: Ensemble full convolutional network for semantic segmentation of high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8011705, doi: 10.1109/LGRS.2021.3076093.

[19] X. Li, Y. Cheng, Y. Fang, H. Liang, and S. Xu, "2DSegFormer: 2-D transformer model for semantic segmentation on aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4709413.

[20] R. Li, S. Zheng, C. Duan, J. Su, and C. Zhang, "Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8009205.

[21] B. Ma and C.-Y. Chang, "Semantic segmentation of high-resolution remote sensing images using multiscale skip connection network," *IEEE Sensors J.*, vol. 22, no. 4, pp. 3745–3755, Feb. 2022.

[22] R. Li, C. Duan, S. Zheng, C. Zhang, and P. M. Atkinson, "MACU-net for semantic segmentation of fine-resolution remotely sensed images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8007205, doi: 10.1109/LGRS.2021.3052886.

[23] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1874–1883.

[24] K. Chen, K. Fu, M. Yan, X. Gao, X. Sun, and X. Wei, "Semantic segmentation of aerial images with shuffling convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 173–177, Feb. 2018.

[25] Q. Zhang, G. Yang, and G. Zhang, "Collaborative network for super-resolution and semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4404512, doi: 10.1109/TGRS.2021.3099300.

[26] J. Wei, Q. Wang, Z. Li, S. Wang, S. K. Zhou, and S. Cui, "Shallow feature matters for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5993–6001.

[27] H. Li, K. Qiu, L. Chen, X. Mei, L. Hong, and C. Tao, "SCAttNet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 905–909, May 2021.

[28] X. Zhang et al., "SERNet: Squeeze and excitation residual network for semantic segmentation of high-resolution remote sensing images," *Remote Sens.*, vol. 14, no. 19, 2022, Art. no. 4770.

[29] Y.-C. Li, H.-C. Li, W.-S. Hu, and H.-L. Yu, "DSPCANet: Dual-channel scale-aware segmentation network with position and channel attentions for high-resolution aerial images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8552–8565, 2021.

[30] B. Liu, J. Hu, X. Bi, W. Li, and X. Gao, "PGNet: Positioning guidance network for semantic segmentation of very-high-resolution remote sensing images," *Remote Sens.*, vol. 14, no. 17, 2022, Art. no. 4219.

[31] L. Ding et al., "Looking outside the window: Wide-context transformer for the semantic segmentation of high-resolution remote sensing images," 2021, *arXiv:2106.15754*.

[32] L. Gao et al., "STransFuse: Fusing Swin transformer and convolutional neural network for remote sensing image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10990–11003, 2021.

[33] R. Li et al., "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607713.

[34] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 181, pp. 84–98, 2021.

[35] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, vol. 29.

[36] W. Chen, Z. Jiang, Z. Wang, K. Cui, and X. Qian, "Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8924–8933.

[37] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of Adam and beyond," 2019, *arXiv:1904.09237*.

[38] L. Sun et al., "Which target to focus on: Class-perception for semantic segmentation of remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4404213, doi: 10.1109/TGRS.2023.3278133.

[39] C. Peng, Y. Li, R. Shang, and L. Jiao, "RSBNet: One-shot neural architecture search for a backbone network in remote sensing image recognition," *Neurocomputing*, vol. 537, pp. 110–127, 2023.

[40] G. Zhou, J. Xu, W. Chen, X. Li, J. Li, and L. Wang, "Deep feature enhancement method for land cover with irregular and sparse spatial distribution features: A case study on open-pit mining," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4401220, doi: 10.1109/TGRS.2023.3241331.

[41] R. Guan, M. Wang, L. Bruzzone, H. Zhao, and C. Yang, "Lightweight attention network for very high-resolution image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4403514, doi: 10.1109/TGRS.2023.3272614.

[42] T. Xiao, Y. Liu, Y. Huang, M. Li, and G. Yang, "Enhancing multiscale representations with transformer for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5605116, doi: 10.1109/TGRS.2023.3256064.

[43] Z. Dong, G. Gao, T. Liu, Y. Gu, and X. Zhang, "Distilling segmenters from CNNs and transformers for remote sensing images' semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5613814, doi: 10.1109/TGRS.2023.3290411.

[44] K. Cha, J. Seo, and T. Lee, "A billion-scale foundation model for remote sensing images," vol. 19, pp. 1–5, 2022, Art. no. 8011705, doi: 10.1109/LGRS.2021.3076093.

[45] C. Huynh, A. T. Tran, K. Luu, and M. Hoai, "Progressive semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16755–16764.

[46] S. Guo et al., "ISDNet: Integrating shallow and deep networks for efficient ultra-high resolution segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4361–4370.

[47] T. Wu, Z. Lei, B. Lin, C. Li, Y. Qu, and Y. Xie, "Patch proposal network for fast semantic segmentation of high-resolution images," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12402–12409.

[48] Q. Li, W. Yang, W. Liu, Y. Yu, and S. He, "From contexts to locality: Ultra-high resolution image segmentation via locality-aware contextual correlation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7252–7261.

[49] Z. Li, X. Zhang, and P. Xiao, "One model is enough: Toward multi-class weakly supervised remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4503513, doi: 10.1109/TGRS.2023.3290242.

[50] 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, doi: 10.1109/TIP.2023.3238648.

Abubakar Siddique received the master's degree in information and communication engineering from the Chongqing University of Posts and Telecommunications, Chongqing, China, in 2019. He is currently working toward the Ph.D. degree with the School of Microelectronics and Communication Engineering, Chongqing University, Chongqing, China.

His research interests include computer vision and remote sensing image analysis.

Zhengzhou Li (Member, IEEE) received the B.S. degree in photoelectron from Northeastern University, Shenyang, China, in 1998, and the M.S. degree in physics' electronics and the Ph.D. degree in optical engineering from the Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, China, in 2001 and 2004, respectively.

He was a Postdoctoral Fellow with the Department of Ophthalmology, Harvard Medical School, USA. He is a Professor with the School of Microelectronics and Communication Engineering, Chongqing University, Chongqing, China. His research interests include target detection and tracking, and image and signal processing for remote sensing.

Abdullah Azeem received the master's degree in information and communication engineering from the Chongqing University of Posts and Telecommunications, Chongqing, China, in 2020. He is currently working toward the Ph.D. degree with the School of Microelectronics and Communication Engineering, Chongqing University, Chongqing, China.

His research interests include computer vision and remote sensing image analysis.

Yuting Zhang received the master's degree in control engineering from the Southwest University of Science and Technology, Mianyang, China, in 2021. He is currently working toward the Ph.D. degree with the School of Microelectronics and Communication Engineering, Chongqing University, Chongqing, China.

His research interests include infrared small target detection and extended target tracking.

Bitong Xu received the B.S. degree in communication engineering from the Harbin University of Science and Technology, Harbin, China, in 2020. She is currently working toward the master's degree in information and communication engineering with Chongqing University, Chongqing, China.

Her research interests include deep learning and remote sensing image processing.