A Cascaded Network With Coupled High-Low Frequency Features for Building Extraction

Xinyang Chen, Pengfeng Xiao, *Senior Member, IEEE*, Xueliang Zhang, *Senior Member, IEEE*, Dilxat Muhtar, and Luhan Wang

***Abstract*—Accurately extracting buildings from high-resolution remote sensing images is crucial for human productivity and livelihood in urban areas. Due to varying scales and indistinct boundaries of buildings, it is crucial to fully leverage the high- and low-frequency features in building extraction from remote sensing images. However, previous studies have solely relied on either low- or high-frequency features, leading to errors such as omissions or internal holes in the detected buildings at various scales. Although some studies have considered the integration between both high- and low-frequency features, they overlook the suitability of different network depths for extracting different frequency features. A novel network called Cascaded Inception Conv-Former Network (CICF-Net) is proposed in this study to solve these problems. It leverages the parallel combination of convolutional neural network and Transformer to efficiently extract high- and low-frequency features for building extraction. In the encoder, as the network depth grows, we gradually reduce the contribution of high-frequency branch and enhance the focus on low-frequency branch. Moreover, a cascaded fusion strategy is employed to extract and integrate multiscale high- and low-frequency features. Meanwhile, we propose gated convolution UperNet as the decoder, which utilizes recursive gated convolution to facilitate multilevel spatial interactions and better restoration of fine-grained spatial details for building segmentation. The proposed CICF-Net achieves competitive accuracies on three public benchmarks: Massachusetts Building Dataset, WHU Aerial Building Dataset, and Inria Aerial Image Labeling Dataset, with IoU of 75.17%, 91.45%, and 81.28%, respectively. This provides strong evidence of its effectiveness in building extraction, as it can accurately capture spatial details and context of buildings.**

***Index Terms*—Building extraction, deep learning, high-frequency feature, high-resolution remote sensing image, low-frequency feature.**

I. Introduction

THE accurate extraction of buildings from high-spatial resolution remote sensing images holds significant importance for urban planning [1], urban dynamic monitoring [2], [3], [4], [5], disaster response [6], [7], [8], and other related applications. However, extracting buildings from high-spatial resolution remote sensing images faces challenges due to substantial differences among buildings and complex backgrounds. Specifically, differences among buildings lie in their various scales as well as shapes. Various scales refer to the presence of buildings with different sizes, while diverse shapes indicate the structures of buildings, including rectangles, irregular polygons, circles, and so on. Meanwhile, it is difficult to distinguish buildings from complex backgrounds, resulting in imprecise segmentation boundaries. Therefore, achieving accurate and efficient extraction of buildings from high-spatial resolution remote sensing images remains a challenge.

When dealing with buildings that vary in scale and shape, it is important to pay attention to both low- and high-frequency information. The low-frequency components provide large-scale patterns, such as the overall shape and structure of the buildings, while the high-frequency components are needed to edges and textures providing fine-grained detail such as building boundaries [9], [10]. However, while high-frequency features are naturally correlated with local details and low-frequency features are inherently linked with global context [11], [12], [13], high- and low-frequency features are not entirely equivalent to global and local features. Local regions may also perceive blurry backgrounds rather than sharp edges and textures.

Therefore, the absence of high-frequency information can lead to omissions of small buildings and imprecise segmentation boundaries, while the lack of low-frequency information may bring out fragmented and discontinuous segmentation masks for large buildings with internal holes.

In order to tackle these challenges, it is necessary to explore methods that excel in extracting both high- and low-frequency features. Traditional methods that focus on high- and low-frequency features typically employed Fourier transform for filtering and processing in the frequency domain [14]. In recent years, with the development of deep learning techniques, deep networks have strong capabilities in representation and generalization. Among these, convolutional neural networks (CNNs) [15] and Transformers [16] are the most widely used. With the extensive research and applications of these two types of deep networks, a study decomposes original images into low- and high-frequency components using the discrete Fourier transform and evaluates model performance on each component, which confirms that Transformers outperform CNNs on low-frequency components but underperform on high-frequency components

Manuscript received 12 March 2024; revised 6 May 2024; accepted 15 May 2024. Date of publication 21 May 2024; date of current version 30 May 2024. This work was supported by the National Natural Science Foundation of China under Grant 42071297. *(Corresponding author: Pengfeng Xiao.)*

The authors are with the Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Key Laboratory for Land Satellite Remote Sensing Applications of Ministry of Natural Resources, School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China and also with the Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China (e-mail: xiaopf@nju.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3403882

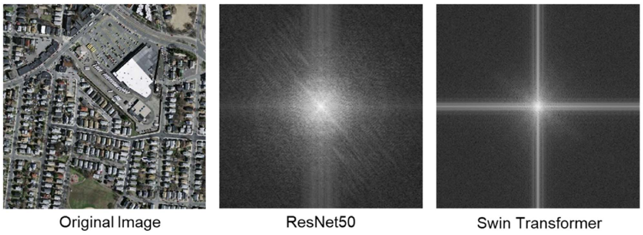


Fig. 1. High-resolution remote sensing image with buildings and the corresponding Fourier spectrum of its final output feature map extracted by ResNet50 and Swin Transformer.

[13]. We can observe this pattern from Fig. 1, where the Fourier spectrum of the output feature map from the Swin Transformer exhibits significantly more low-frequency components compared to ResNet50. Given the understanding of the role of high- and low-frequency features in remote sensing images, it is imperative to explore how CNNs and Transformers address these features differently.

CNNs have been proven to excel in extracting high-frequency features [13], [17], which have been introduced to various remote sensing tasks, especially for building extraction [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34]. Some multiscale networks based on CNNs have been proposed to extract building contours and precise boundaries at different scales, some of which follow HRNet for building segmentation and change detection tasks, achieving good results [18], [22], [35], [36], [37]. Several studies employed postprocessing strategies to refine the building segmentation map from coarse to fine, enhancing building contours through polygon regularization or using direction maps that indicate the direction of each pixel to the center of the nearest object it might belong to [23], [24]. Li et al. [34] have employed a strategy of multitask learning which employs CNN to simultaneously predict distances and directions, effectively improving the accuracy of building segmentation and boundary delineation. These studies have shown that CNNs excel at capturing details and local context in building extraction [38], [39]. But the fixed-size convolutional kernels impose limitations on capturing overall information, as they only have limited receptive fields, thereby lacking the capability to model long-range dependencies [40], [41], [42]. Hence, dilated convolutions have been designed to expand the receptive field, enabling models to capture a broader range of context information [20], [21], [22]. The atrous spatial pyramid pooling method has been applied to extract dense and multiscale features simultaneously [20], [27], [28], [29], [30]. However, these studies still encounter obstacles in capturing low-frequency features and global context.

In contrast to CNNs, Transformers utilize self-attention mechanisms for feature extraction and show exceptional proficiency in capturing global context and extracting low-frequency features [43], [44], [45], [46]. Vision Transformer (ViT) [47] and its variants have shown remarkable performance in image processing [48], [49], [50], [51]. General large models based on Transformers are highly-sought-after and yield strong performance, which leverage a vast number of parameters and progressive training to fully exploit remote sensing images and other multimodal data [52], [53]. But Transformers often have substantial computational requirements and are less effective in extracting high-frequency features compared to CNNs [54], [55]. Therefore, there are currently efforts in remote sensing image segmentation tasks that combine CNNs and Transformers to construct deep networks to complement their shortcomings and harness their advantages [56], [57], [58], [59]. The U-Net architecture is commonly utilized, either by combining a CNN-based encoder with a Transformer-based decoder [44] or incorporating a Transformer-based encoder, such as Swin Transformer for feature extraction [57]. However, most studies that incorporate Transformers into the CNN framework only utilize pure Transformers either in the encoder or decoder, neglecting the potential of combining the strengths of both methods in feature extraction.

On the other hand, there have been efforts to integrate CNN and Transformer within encoder or decoder. For instance, CNN just serves as a means to obtain tokens required for self-attention, while the primary feature extraction is accomplished through the Transformer in building extraction tasks [58]. Some methods employ independent Transformer and CNN branches for extracting features [59], [60], [61], which are fused and utilized in the decoder. An extension of U-Net in [44] designs a Transformer-based decoder with a global-local dual-branch attention module for contextual information across scales. These approaches either adopt two parallel and independent branches with intersections occurring only at the final fusion module, or solely employ CNN for preliminary feature extraction as a preparatory step for the Transformer. However, they overlook the hierarchical nature of the features in the network: shallow or deep layers of networks are more adept at capturing high-frequency (corners or edge/color conjunctions) or low-frequency features (overall patterns such as shapes and tone of objects) which are both essential for each respective layer [10], thereby failing to fully exploit the combined advantages of both methods. Additionally, there is a lack of continuous fusion of the two types of features throughout the process. A more effective multiscale feature fusion strategy should be applied to integrate the high- and low-frequency features obtained at different depths for handling the diverse scales of buildings.

With this motivation, we propose a novel network named Cascaded Inception Conv-Former Network (CICF-Net) for high-spatial resolution building extraction. Inspired by Inception Transformer [55], we designed an Inception Conv-Former (iConvFormer) Block to effectively combine the strengths of CNN and Transformer in extracting high- and low-frequency features at different depths. As the network depth grows, we gradually reduce the contribution of high-frequency branch and enhance the focus on low-frequency branch, resembling a frequency ramp, which ensures each depth level receives an appropriate allocation of channels. In addition, we also adopt a cascaded feature fusion strategy. Unlike the commonly used lightweight multiscale fusion module strategy (such as the feature pyramid network [62], [63], [64]), the cascaded framework allows our feature fusion module to have more parameters and occupy a more important weight in the model, which has been proven to achieve better performance [65]. And another study has proposed multiscale progressive segmentation, using three subnetworks to segment objects into small, large, and other scales gradually, which also allocates more parameters to multiscale feature fusion and demonstrates good performance in multiclass segmentation [66]. Meanwhile, in light of the intricate interplay between buildings themselves and their surroundings, we introduce the Gated Convolution UperNet (GCUperNet), which employs recursive gated convolutions [67] to enable multilevel interactions among features. Moreover, ViT-based models have been verified to face challenges during training, demanding a notably large amount of training samples compared to CNN-based models [13]. Therefore, we employ the self-supervised learning method named contrastive mask image distillation (CMID) [68] for pretraining on the Million-AID dataset [69] to overcome the data-hungry nature of deep learning models and improve the model's generalization capabilities. We opted CMID for its proficiency in extracting representations that are both global semantic separable and local spatial perceptible, achieved by combining the inductive biases of contrast learning and masked image modeling. This representation aligns well with our design for extracting high-and low-frequency features, crucial for building segmentation tasks that require simultaneous understanding of the relation between buildings and the broader scene, as well as between individual buildings. Furthermore, CMID demonstrates faster convergence during pretraining compared to other self-supervised methods, such as masked autoencoders [70], significantly reducing the additional computational overhead involved in the pretraining process.

The main contributions of this study are as follows.

1) We propose the iConvFormer Block, which combines high- and low-frequency features in parallel using a frequency ramp structure to gradually increase the channel ratio of low-frequency branches. This module is designed to address the issue of potential omissions of small buildings and discontinuity in large buildings caused by the lack of high- and low-frequency information.
2) A cascaded fusion framework is proposed to connect multiple iConvFormer Blocks and integrate multiscale features, which compose a comprehensive backbone network called CICF-Net. Furthermore, CICF-Net is pretrained using the self-supervised learning method.
3) A decoder called GCUperNet is introduced to enable multilevel spatial interactions at different scales, resulting in improved restoration of spatial detail. Experimental results on three building datasets show that the proposed method achieves outstanding accuracies.

II. Methodology

A. Overview of CICF-Net

We propose CICF-Net as a dedicated network for building extraction in high-spatial resolution images, which adopts an encoder-decoder structure. We designed corresponding modules for both the encoder and the decoder to address the challenges of building extraction.

In the overall architecture of the encoder, we incorporate the iConvFormer Block that extracts high- and low-frequency features, as well as a cascaded fusion structure allowing for effective feature fusion with a significant portion of the whole model's parameters (Fig. 2). A lightweight Transition Block and Focal Block are designed for the cascaded fusion structure, enabling the model to generate enriched multiscale features. Specific details of these modules of the encoder are provided in the corresponding subsections. The encoder consists of four stages with consistent forms. Prior to being input into each stage, the feature maps undergo patch embedding. In the case of the original remote sensing image input into the first stage, the patch embedding includes two convolution-and-normalization operations with a GELU activation applied in between. For patch embeddings between subsequent stages, a single convolution-and-normalization operation is applied. After each patch embedding, we obtain the intermediate feature map $\mathbf{X} \in \mathbb{R}^{N\times C}$ that can be fed into each corresponding stage. For each stage, N and C, respectively, represent the number of pixels ($H \times W$, where H and W represent the height and width of the feature maps) and embed dimensions of the feature maps.

Meanwhile, a frequency ramp structure is adopted to balance the channel proportions for high- and low-frequency feature extraction within each stage. Moreover, we introduce the GCUperNet as our decoder, which will be illustrated in the following subsection.

B. Inception Conv-Former for Building Extraction

We design an iConvFormer aggregator that utilizes multibranch and parallel architecture to couple high- and low-frequency features (Fig. 3). The architecture of iConvFormer aggregator is tailored to harness the efficient feature extraction capability of CNNs for high-frequency features in shallow layers and the robust long-range modeling ability of Transformers for low-frequency features in deep layers.

The feature map $\boldsymbol{X_{raw}} \in \mathbb{R}^{N\times C_{raw}}$ fed into the iConvFormer aggregator is split into two parts along the channel dimension, $\boldsymbol{X_{high}} \in \mathbb{R}^{N\times C_{high}}$ and $\boldsymbol{X_{low}} \in \mathbb{R}^{N\times C_{low}}$. Notably, the total number of channels $\boldsymbol{C_{raw}}$ of $\boldsymbol{X_{raw}}$ is equal to the sum of the number of channels C_{high} and C_{low} for each branch. These two parts are then, respectively, fed into the high- and low-frequency branches of the iConvFormer aggregator. The high-frequency branch employs CNNs for feature extraction, while the low-frequency branch utilizes the self-attention as its core operation.

Specifically, the high-frequency branch is designed to extract high-frequency features based on CNN and is divided into two sub-branches operating in parallel. The main sub-branch takes into account the detail-perception ability of convolutions by following the successful inverted bottleneck structure of ConvNeXt [71]. Additionally, considering the high sensitivity of the maximum filter for high-frequency information, we utilize max-pooling in the other sub-branch. This dual design targets the efficient extraction of high-frequency features. Particularly, we split $\boldsymbol{X_{high}}$, obtained through channel-wise averaging, into two parts, $\boldsymbol{X_{high_1}}$ and $\boldsymbol{X_{high_2}} \in \mathbb{R}^{N\times \frac{C_{high}}{2}}$, which are,

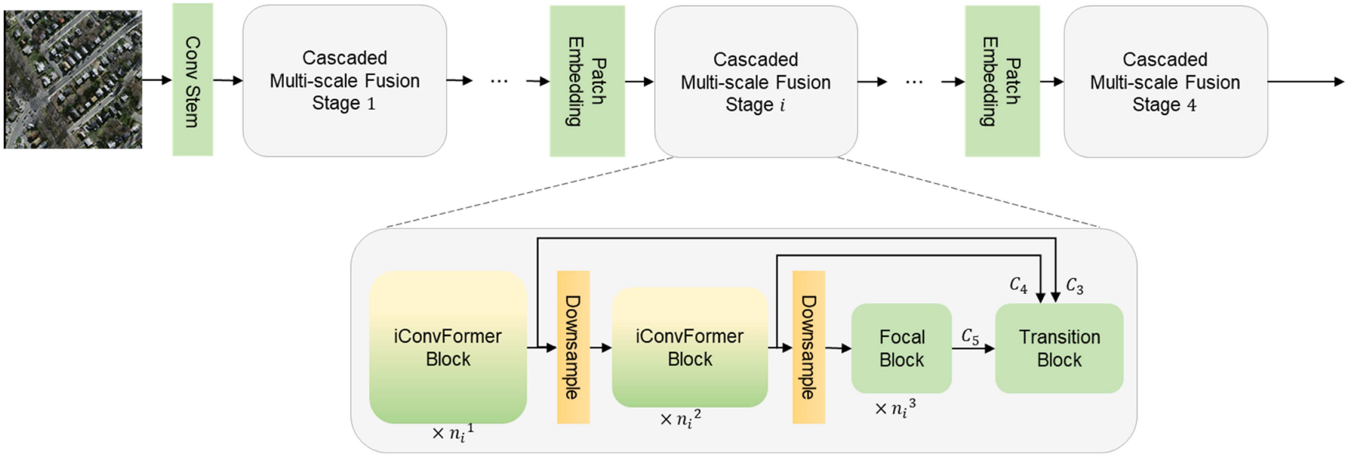


Fig. 2. Overview of the encoder of CICF-Net. The network consists of four stages, where the iConvFormer blocks are interconnected through a cascaded multiscale fusion strategy. C_3, C_4, and C_5 represent feature maps at three different scales. n_i denotes the number of blocks in stage i.

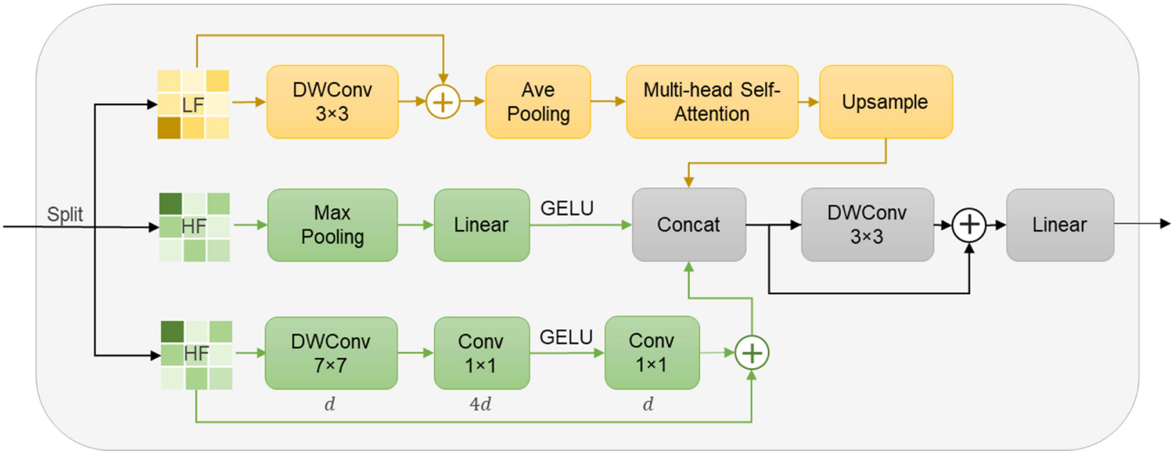


Fig. 3. Internal structure of iConvFormer aggregator. The input feature map is divided into parts for low-frequency branch (orange, LF) and high-frequency branch (green, HF) along the channel dimension. Within the high-frequency branch, it is further split into two sub-branches, each allocated half of its channel number. The bottom high-frequency sub-branch incorporates the ConvNeXt structure, which includes inverted bottleneck layers with channel expansion operations (from d to $4d$ and then back to d). The gray boxes illustrate the initial fusion of the high- and low-frequency branches within each iConvFormer block.

respectively, fed into two sub-branches of the high-frequency branch. As shown in Fig. 3, $\boldsymbol{X_{high_1}}$ is mainly processed by a depth-wise convolution with a large kernel size of 7×7 and two 1×1 convolution layers to implement the inverted bottleneck structure. This structure is designed with an expansion (four times wider) of the middle dimension compared to the beginning and the end, effectively avoiding the problem of information loss [71]. $\boldsymbol{X_{high_2}}$ undergoes a max-pooling and a linear layer to supplement any lost high-frequency details from the main sub-branch and suppress noise. The outputs of two sub-branches are denoted as $\boldsymbol{Y_{high_1}}$ and $\boldsymbol{Y_{high_2}}$.

Meanwhile, the low-frequency branch aims to capture global low-frequency information by leveraging multihead self-attention (MHSA), which owns the ability to model long-range dependencies and global context. However, the self-attention operation does not inherently consider patch order due to its permutation invariance. Therefore, we first apply a 3×3 depth-wise convolution with zero-padding to encode positional information [72], [73]. Then, we employ average-pooling as a sparsification method before MHSA to reduce significant computational demands, which enables self-attention operation to better focus on global low-frequency information. We define the result after sparsification as $\boldsymbol{X_{mhsa}} \in \mathbb{R}^{n \times Dim}$, where Dim represents the dimension of the input vector of MHSA. Based on the number of attention heads (h), each attention head evenly distributes the dimensions it operates on, leading to the creation of h sets of $\boldsymbol{X_{head_i}} \in \mathbb{R}^{n \times \frac{Dim}{h}} (i = 1, 2, \ldots, h)$. $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ vectors are obtained after h sets of individual linear projection transformations W_i^q, W_i^k, and W_i^v (Fig. 4)

$$\boldsymbol{Q} = \{Q_i\} = \{\boldsymbol{X_{head_i}} W_i^q\} \tag{1}$$

$$\boldsymbol{K} = \{K_i\} = \{\boldsymbol{X_{head_i}} W_i^k\} \tag{2}$$

$$\boldsymbol{V} = \{V_i\} = \{\boldsymbol{X_{head_i}} W_i^v\}. \tag{3}$$

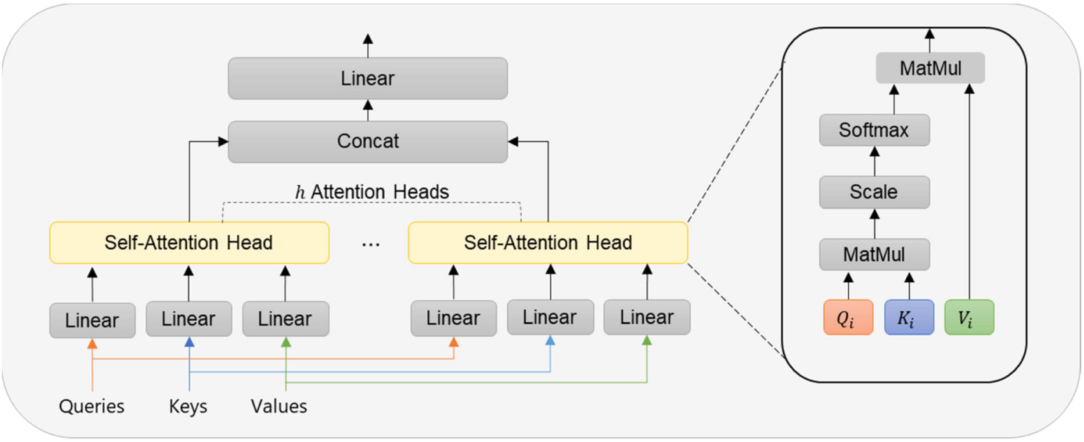


Fig. 4. Details of MHSA. Queries, keys, and values all refer to $\boldsymbol{X_{mhsa}}$, which undergo h sets of distinct linear projections and yield transformed vectors denoted as Q_i, K_i, and V_i $(i = 1, 2, \ldots, h)$. The h sets of Q_i, K_i, and V_i are then fed into self-attention heads to execute self-attention operations. The outputs of h individual self-attention computations are concatenated and transformed through another learnable linear projection to generate the final output.

These sets of Q_i, K_i, and V_i undergo self-attention operations when fed into their corresponding self-attention heads as (4). The dot-product result obtained through matrix multiplication (MatMul) needs to be divided by a scaling factor $\sqrt{d_{k_i}}$ (dimension of K_i) prior to applying softmax, which typically regulates the range of attention weights and mitigates the risk of gradient explosions. Then, the results of the h self-attention computations are concatenated and transformed through an additional linear projection for the ultimate output of MHSA as (5).

$$\text{head}_i = \text{SelfAttention}\left(Q_i, K_i, V_i\right) = \text{Softmax}\left(\frac{Q_i {K_i}^T}{\sqrt{d_{k_i}}}\right) V_i \tag{4}$$

$$MHSA\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right) = \text{Concat}\left(\text{head}_1, \text{head}_2, \ldots, \text{head}_h\right) W^o. \tag{5}$$

Subsequently, we adopt upsampling to restore the original resolution of $\boldsymbol{X_{low}}$ after MHSA and obtain the output of the low-frequency branch $\boldsymbol{Y_{low}}$. Here, we select bilinear interpolation as the upsampling method.

After extracting the high-frequency features $\boldsymbol{Y_{high_1}}$ $\boldsymbol{Y_{high_2}}$ and low-frequency features $\boldsymbol{Y_{low}}$, the iConvFormer aggregator first concatenates the three features along the channel dimension to obtain $\boldsymbol{Y_c}$ as (6). It then applies a depth-wise convolution with residual connection followed by a linear layer to produce the final output $\boldsymbol{Y_o}$ as (7). As a result, the final output $\boldsymbol{Y_o}$ incorporates the complementary high- and low-frequency features extracted by different branches.

$$\boldsymbol{Y_c} = \text{Concat}\left(\boldsymbol{Y_{high1}},\ \boldsymbol{Y_{high2}}, \boldsymbol{Y_{low}}\right) \tag{6}$$

$$\boldsymbol{Y_o} = \text{Linear}(\boldsymbol{Y_c} + DWConv\left(\boldsymbol{Y_c}\right). \tag{7}$$

Within each iConvFormer Block, the iConvFormer aggregator, along with residual connections, extracts high- and low-frequency features. And a multilayer perceptron (MLP) is then used to align the channel dimensions between adjacent blocks.

$$\boldsymbol{O_{mid}} = \boldsymbol{X_{raw}}\ + ICFA\left(LN\left(\boldsymbol{X_{raw}}\right)\right) \tag{8}$$

$$\boldsymbol{O} = \boldsymbol{O_{mid}}\ + MLP\left(LN\left(\boldsymbol{O_{mid}}\right)\right) \tag{9}$$

where $\boldsymbol{X_{raw}}$ denotes the input of the current iConvFormer Block, ICFA represents the iConvFormer aggregator, and LN stands for the layer normalization which is beneficial for stabilizing and accelerating the training process. Finally, we obtain the output of an iConvFormer Block $\boldsymbol{O}$. In addition, the core idea of iConvFormer involves creating high- and low-frequency branches as well as regulating the channel ratios of $\boldsymbol{X_{high}}$ and $\boldsymbol{X_{low}}$. Common image processing models capture basic local context by utilizing high-frequency details in shallower layers, while gradually shifting their focus toward capturing low-frequency features in deeper layers, enabling a global understanding of the input feature map. Therefore, we adopt a frequency ramp structure that incrementally allocates more channels to the low-frequency branch to fully leverage the capabilities of the high- and low-frequency branches at different depths of the network. Specifically, our encoder consists of four stages with different channel and spatial dimensions. For each stage, a pair of channel ratios ($\frac{C_{high}}{C_{raw}}$ and $\frac{C_{low}}{C_{raw}}$) are defined to balance the proportion of high- and low-frequency branches, where $\frac{C_{high}}{C_{raw}} + \frac{C_{low}}{C_{raw}} = 1$. $\frac{C_{high}}{C_{raw}}$ decreases gradually from shallower to deeper layers, while $\frac{C_{low}}{C_{raw}}$ increases.

C. Cascaded Multiscale Feature Extraction

A cascaded structure is implemented to connect iConvFormer Blocks within each stage for multiscale feature extraction and fusion. Moreover, we ensure that a greater proportion of parameters allocated for feature fusion can achieve improved performance. Feature extraction along with multiscale fusion are no longer independent but continuously interact with each other in the context of the structure.

This structure comprises two types of iConvFormer Blocks with different scales, a Focal Block, and a Transition Block. The high-resolution feature map of the ith stage is fed into ${n_i}^1$ iConvFormer Blocks to obtain feature C_3. After downsampled by a 3 × 3 convolution with a stride of 2, the feature C_4 is

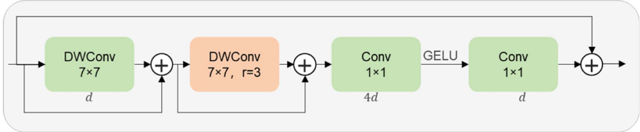


Fig. 5. Diagram of the Focal Block, where the second depth-wise convolution (orange) is a dilated convolution with the dilation rate of $r = 3$. The block utilizes inverted bottleneck layers with channel expansion operations (from d to $4d$ and then back to d).

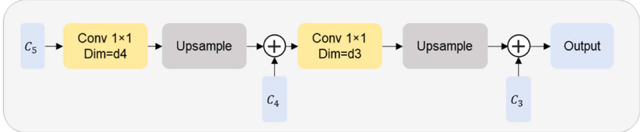


Fig. 6. Diagram of the Transition Block. The blue boxes C_3, C_4, and C_5 represent feature maps at different scales. The yellow boxes represent the channel reduction using a 1×1 convolution to achieve the desired dimension Dim, where d_4 and d_3 represent the number of C_4 channels and the number of C_3 channels, respectively. Upsampling is performed using bilinear interpolation with a scale factor of 2.

obtained via ${n_i}^2$ iConvFormer Blocks. Following another round of downsampling, the feature map is then fed into ${n_i}^3$ Focal Blocks, generating C_5. Finally, we can input C_3, C_4, and C_5 into the Transition Block for multiscale feature fusion.

The Focal Block is a variant of the required block structure and our iConvFormer Block is based on the bottom high-frequency branch. Thus, similar to ConvNeXt, the internal structure of a Focal Block includes a depth-wise convolution with a large kernel size and two 1×1 convolution layers (Fig. 5). And the former 1×1 convolution implements a channel expansion operation, increasing the number of feature channels by four times. In addition, the Focal Block adds a 7×7 dilated convolution with a dilation rate of 3 and incorporates two skip connections in the middle. The design of Focal Block allows fine-grained local features to interact with coarse-grained global features again.

In a Transition Block, features of different scales, namely C_5, C_4, and C_3, gradually upsample and fuse (Fig. 6). This design is similar to the fusion module in FPN, but without extra convolutions to transform the summed features. The fused features produced by the Transition Block are the final output of each stage.

D. Gated Convolution UperNet

Capturing spatial interactions at different levels and scales is crucial in building extraction, such as the interactions between buildings and backgrounds, as well as among different buildings. Therefore, we consider incorporating the decoder module used for image segmentation in UperNet [10]. It can facilitate the interaction of features at different scales, enabling effective representation of spatial interactions in building extraction.

The original UperNet used for segmentation is based on the design principles of pyramid pooling module (PPM) [74] and feature pyramid network (FPN) [62]. However, the original UperNet only employs single-order convolutional layers (order = 1), limiting its ability to capture simple spatial interactions at each level. Although decoders based on Transformers can achieve spatial interactions through multihead attention, they tend to have a large number of parameters and lack the inductive bias of convolutional operations, making it relatively more challenging to capture complex spatial interactions. Therefore, we introduce the recursive gated convolution (g^nConv) as an improvement to UperNet, yielding the gated convolution UperNet (GCUperNet) as the decoder (Fig. 7). g^nConv can realize multilevel spatial interactions using a highly efficient implementation with gated convolutions and recursive designs.

In the original UperNet, PPM is appended after the last layer of the backbone network and then fed into the top-down branch of FPN. We replace the original convolutional layers in UperNet with g^nConv, including those in the bottleneck layers of the PPM Head and FPN module. Meanwhile, a 1×1 convolution layer is applied to ensure consistent feature dimensions across all feature maps for subsequent fusion. All layers generated by downsampling in FPN are concatenated, followed by the bottleneck layer of FPN to yield the fused feature map, which is then input into the segmentation head with a convolutional kernel size of 1. We employ g^nConv with order = 2 to fulfill the spatial interaction requirements for building extraction (Fig. 7). The input of g^nConv $(n = 2)$ with a dimension of C undergoes an initial projection through a 1×1 convolution and is divided into two parts along dimension: $2C - \frac{C}{2}$ and $\frac{C}{2}$. The former applies a depth-wise convolution of which result is split into two parts with dimensions of $\frac{C}{2}$ and C, while the latter is multiplied with the part with a dimension of $\frac{C}{2}$ from the depth-wise convolution and then projected into an outcome with a dimension of C. This outcome is further multiplied by the remaining part of the depth-wise convolution and projected into the final output of g^nConv $(n = 2)$ with a dimension of C as well. This module enables us to capture spatial interactions between backgrounds and buildings, as well as among different buildings.

E. Loss Functions

We introduce four loss functions and adopt a joint loss formulation to constrain the training process of CICF-Net. The joint loss function, denoted as L, can be defined as follows:

$$L = a \cdot L_{CE} + b \cdot L_{Dice} + c \cdot L_{Lovasz} + d \cdot L_{BF1} \tag{10}$$

where L_{CE} refers to the cross-entropy loss; L_{Dice} denotes the dice loss [75]; L_{Lovasz} corresponds to the Lovasz loss [76]; and L_{BF1} represents the boundary loss [77]. The first three loss functions are all based on intersection over union (IoU) and are designed to address the significant class imbalance between background and building samples in semantic segmentation tasks. Meanwhile, the Lovasz loss prioritizes the shape of objects, aiding in capturing the geometric structure and continuity. It incorporates a certain level of boundary control by considering the pixel order along the object boundaries, ensuring fine boundaries for building extraction to a certain extent. In addition, we introduce the boundary loss, which measures the accuracy of building boundary detection using the Euclidean distance from

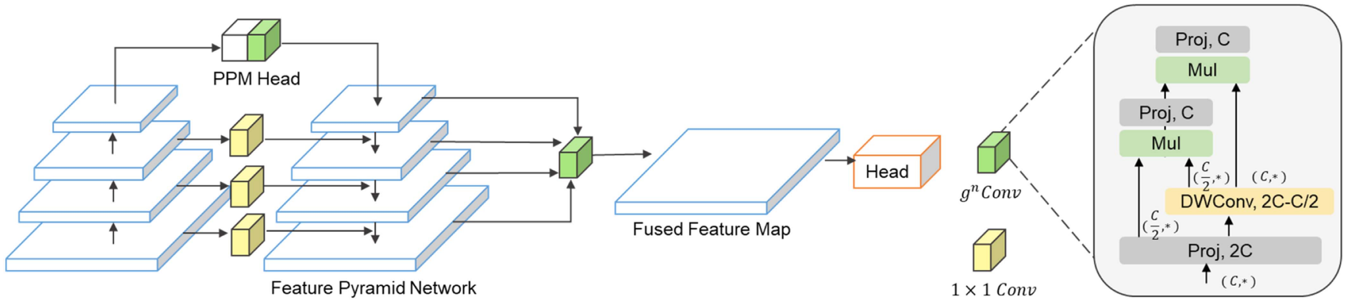


Fig. 7. Overview of GCUperNet. $g^n Conv$ (green blocks) is employed in the bottleneck layers of the PPM head and FPN, utilizing gated convolutions and recursive designs for efficient spatial interactions in building extraction (order = 2). A 1×1 convolution operation (yellow blocks) is applied to downsampled features before being fed into FPN.

pixels to boundaries. As a result, the incorporation of boundary loss L_{BF1} can effectively penalize boundary misalignment.

III. Experimental Results

A. Experimental Setup

1) Datasets: We conduct experiments to evaluate the performance of CICF-Net, using three publicly available building datasets: the Massachusetts Building Dataset, the WHU Aerial Building Dataset, and the Inria Aerial Image Labeling Dataset. These datasets provide a diverse set of building images that can be used to assess the performance of our model in building extraction tasks.

Massachusetts Building Dataset: The dataset consists of 137 aerial images with a spatial resolution of 1 m. Each image has a size of 1500×1500 pixels, covering a total area of 2.25 km^2. The dataset includes images from both urban and suburban areas of Boston, featuring buildings of varying sizes, shapes, textures, and roof colors. It is split into a training set of 137 images, a validation set of 4 images, and a test set of 10 images, each with pixel-level ground truth labels. Following the official dataset partition, we crop the training and validation images and their labels into 512 × 512 pixels with a stride of 256 pixels, resulting in an overlap rate of 0.5 for both the training and validation sets. The testing set does not have any overlap, with a size of 512 × 512 pixels as well. As a result, we obtain 3425 tiles for training, 100 tiles for validation, and 90 tiles for testing.

WHU Aerial Building Dataset: The dataset is an aerial image subset of the WHU Building Dataset [78], which includes an aerial subset and a satellite subset. The original aerial data comes from the New Zealand Land Information Services website and covers an area of 450 km^2, including 220 000 independent buildings. The dataset comprises a vast majority of original aerial data, containing 8189 tiles of RGB aerial images, each with 512 × 512 pixels and a spatial resolution of 0.3 m. The dataset is further divided into a training set (4736 tiles, 1 30 500 buildings), a validation set (1036 tiles, 14 500 buildings), and a test set (2416 tiles, 42 000 buildings). The dataset partition follows the official guidelines for experiments.

Inria Aerial Image Labeling Dataset: The dataset comprises 360 aerial images collected from five cities (Austin, Chicago, Kitsap, Tyrol, and Vienna), spanning a total area of 810 km^2, with half allocated to the training set and the other half to the test set [79]. The images capture diverse urban landscapes, ranging from highly densely populated metropolitan financial districts to alpine towns. Due to the unavailability of labels for the test set, our experiments solely rely on the original training set, where each city consists of 36 images with a size of 5000 × 5000 pixels and a spatial resolution of 0.3 m. Following the official partition of the original training set, the 1–5 images from each city are selected for testing, while the remaining 31 images from each city are used for training and validation. We crop these large images and their labels into 512 × 512 pixels. The ratio of the numbers of the training tiles to the validation tiles is 4:1. In the end, we obtained 12 400 tiles for training, 3100 tiles for validation, and 2500 tiles for testing.

TABLE I
Network Parameter Settings for the Encoder in the Main Experiments of CICF-Net

Parameters	Settings for per stage
Embedding dimensions	96, 192, 384, 512
Total heads	3, 6, 12, 16
Attention heads	1, 3, 8, 15
Number of blocks $(n_i{}^1, n_i{}^2, n_i{}^3)$	(1,2,1), (2,2,2), (4,6,4), (2,2,2)

2) Network Settings: The hyperparameters of the proposed CICF-Net include the network depth and the channel ratio between the low- and high-frequency branches in the iConvFormer Block. The network depth is determined by the number of iConvFormer Blocks and Focal Blocks in each stage. The depth of stage i is the sum of $n_i{}^1$, $n_i{}^2$, and $n_i{}^3$. We set the number of total heads and the overall width (number of embedding dimensions) for each stage based on the base scale of general models. As for the channel ratio setting in the iConvFormer Block, we conduct an experimental analysis in Section III-C. We control the channel ratio between the low- and high-frequency branches through the number of "heads," with a total of (3, 6, 12, and 16) heads for the two branches in the four stages. The main experimental parameter settings are presented in Table I.

The ratio between the number of attention heads and the number of total heads determines the proportion of the low-frequency branch. Specifically, the low-frequency branch ratios for the

TABLE II
ABLATION STUDY FOR EACH BRANCH WITHIN THE ICONVFORMER BLOCK

ConvNeXt	Maxpool	Transformer	IoU (%)	F1 (%)
✓			71.75	83.55
✓	✓		72.36	83.96
✓	✓	✓	**73.35**	**84.63**

TABLE III
ABLATION STUDY FOR THE ENTIRE ENCODER OF CICF-NET

iConvFormer	Cascaded fusion	Pretrained	IoU (%)	F1 (%)
✓			73.35	84.63
✓	✓		74.56	85.43
✓	✓	✓	**75.17**	**85.83**

four stages are 1/3, 1/2, 2/3, and 15/16, respectively, indicating the increasing proportion of the low-frequency branch as we progress through the stages.

The experiments in this study are conducted on an NVIDIA Tesla V100 GPU with 32 GB memory, using Python 3.9.0 and PyTorch 1.12.1. Since the performance of deep learning models heavily relies on the amount of data, we employ the CMID [68] for self-supervised pretraining the encoder of the proposed model on the Million-AID dataset for 200 epochs. Subsequently, we perform fine-tuning on the three building datasets for 100 epochs. During the training phase, we apply a cyclic cosine annealing learning rate strategy and use the Adan optimizer [80] with a weight decay of 0.02 and a base learning rate of 3e-4. Gradient clipping is utilized to stabilize the training process. The batch size is set to 8. Data augmentation techniques such as random flipping, random cropping, data normalization, and photometric distortion are applied during the training phase. According to an automated grid search for parameter adjustment, we set the weight ratios of the four losses in (5) as 1:1:5:0.5 to control the convergence of CICF-Net, resulting in the final formulation of L as follows:

$$L = L_{CE} + L_{Dice} + 5 \times L_{Lovasz} + 0.5 \times L_{BF1}. \qquad (11)$$

In all subsequent experiments, we employ the following widely used evaluation metrics to assess the performance of our proposed CICF-Net in building extraction: F1-score (F1) and intersection over union (IoU).

B. Component Analysis

1) Ablation Study for the Encoder: Despite its small scale, the Massachusetts Building Dataset offers a comprehensive coverage of various types of buildings. Therefore, it can be an ideal choice for conducting ablation experiments to assess the performance of the internal modules in CICF-Net. First, we evaluate the performance of the modules within the iConvFormer Block to verify the effectiveness of our method in synergistically leveraging CNN and Transformer for extracting high- and low-frequency features. Second, we perform ablation analyses on the entire encoder component, which encompass evaluating the improvement achieved by pretraining the encoder using CMID. In this part, we ensure fairness in the ablation experiments by consistently utilizing our modified GCUperHead as the decoder in segmentation tasks.

The quantitative results of the ablation experiments of iConvFormer Block are presented in Table II. We assess the performance of combining the high- and low-frequency branches within the iConvFormer Block. Specifically, we employ ConvNeXt as the primary sub-branch for extracting high-frequency features, which serves as the baseline. Moreover, we introduce a max-pooling sub-branch to enhance the capture of high-frequency features, leveraging the high sensitivity of the maximum filter and resulting in a 0.61% improvement in IoU and a 0.41% improvement in F1-score on the Massachusetts Building Dataset. Expanding on the high-frequency branch, we further integrate a Transformer branch for extracting low-frequency features, contributing to additional improvements of 0.99% in IoU and 0.67% in F1-score on the same dataset. It is worth noting that there are no pretrained model weights used in our ablation experiments of iConvFormer. Our experimental results offer evidence supporting the effectiveness of iConvFormer Block in successfully parallelly combining CNN and Transformer for extracting high and low-frequency features.

Based on the complete iConvFormer Block, we incorporate a cascaded fusion strategy. Table III presents the quantitative results of ablation experiments on the encoder component of CICF-Net. The cascaded fusion approach establishes a connection between multiple iConvFormer Blocks to extract and merge multiscale high- and low-frequency features, thus forming the complete encoder of CICF-Net. Compared to the experiment without the cascaded multiscale feature fusion module, it achieves noteworthy improvements of 1.21% in IoU and 0.80% in F1-score on the Massachusetts Building Dataset. Deep learning models, particularly Transformer-based models, heavily rely on the availability of large-scale datasets to support the training process. Typically, many models are pretrained on large-scale image datasets such as ImageNet and then fine-tuned for specific downstream tasks. Following this practice, we utilize a highly effective self-supervised pretraining method tailored for remote sensing tasks [68], to pretrain the CICF-Net encoder. The pretraining is conducted using the Million-AID dataset, a comprehensive dataset comprised of millions of images. Quantitative experimental results confirm the benefits of self-supervised pretraining followed by fine-tuning on the Massachusetts Building Dataset. Finally, we obtained excellent segmentation results, achieving the SOTA accuracies with an IoU of 75.17% and an F1-score of 85.83%.

2) Evaluation on the Decoder: We compare our decoder with the original UperNet, so as to demonstrate the effectiveness of the proposed GCUperNet for building extraction. We use $g^n Conv$ with order $= 2$ to capture spatial interactions between backgrounds and buildings as well as among different buildings, so as to facilitate accurate building extraction. Table IV shows the quantitative comparison results between GCUperNet and the original UperNet. Notably, the two experiments are both based on the encoder of CICF-Net using CMID for a fair comparison.

TABLE IV
COMPARATIVE EXPERIMENTS OF THE ORIGINAL UPERNET AND THE MODIFIED GCUPERNET

Decoder	IoU (%)	F1 (%)
UperNet	74.40	83.32
GCUperNet	**75.17**	**85.83**

TABLE V
EXPERIMENTAL INVESTIGATION OF THE OPTIMAL RATIOS OF LOW-FREQUENCY BRANCH CHANNELS (C_{low}) OF FOUR STAGES

Total heads	Attention heads	C_{low} ratio	IoU (%)	F1 (%)
3,6,12,16	1,2,6,12	(1/3, 1/3, 1/2, 3/4)	73.93	85.01
	1,3,8,15	(1/3, 1/2, 2/3, 15/16)	**74.56**	**85.43**
	2,4,9,15	(2/3, 2/3, 3/4, 15/16)	74.16	85.16

Sum of C_{low} and C_{high} equals the embedding dimension at each stage.

The improved GCUperNet exhibits superior accuracy on the Massachusetts Building Dataset, highlighting the effectiveness of substituting convolutional operations in the bottleneck layers of PPM and FPN modules with recursive gated convolutions in building extraction tasks.

C. Analysis of Channel Ratio

In CICF-Net, the channel ratio of high- and low-frequency branches in the iConvFormer Block plays a vital role in the method. To explore the impact of the channel ratio on the effectiveness of building extraction and determine the optimal ratio, we conduct a series of experiments. In the experimental setup, we use the number of "heads" to control the channel ratio between the high- and low-frequency branches. The numbers of "heads" are designed to facilitate the calculation of channel ratios. By adjusting the allocation of attention heads to the low-frequency branch, we are able to control the channel ratios of C_{low}. The numbers of total heads and attention heads for each stage are determined based on the embedding dimensions as well as the desired channel ratios. We derive the desired ratios based on factors of embedding dimension (total number of channels) per stage, ensuring that embedding dimension is divisible by the denominator of these ratios. Hence, the number of total heads is determined as the least common multiple of the denominators of these desired C_{low} ratios.

We conducted three experiments, and the specific configurations are presented in Table V. None of these experiments involved loading CMID pretrained weights. It can be observed that the first two experiments maintain an identical channel ratio for the low-frequency branch in the initial stage. In the second experiment, C_{low} ratios are increased in the subsequent three stages compared to the first experiment, resulting in a 0.63% improvement in IoU. In contrast, the third experiment increases C_{low} ratios in the first three stages while keeping it unchanged in the final stage compared to the second one, leading to a decrease in accuracy. This suggests that the second experiment combines high- and low-frequency features more effectively. It prioritizes low-frequency features more in deeper layers compared to the first experiment and emphasizes high-frequency features more in shallower layers compared to the third one, achieving an optimal tradeoff. Consequently, we adopt the parameter settings from the second experiment for all other CICF-Net experiments presented in this article.

TABLE VI
QUANTITATIVE COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON THE MASSACHUSETTS BUILDING DATASET

Method	IoU (%)	F1 (%)
U-Net	70.71	82.84
Deeplab V3+	68.92	81.60
Swin Transformer	72.11	83.80
UNetFormer	71.67	83.50
MA-FCN	73.80	84.93
EU-Net	73.93	85.01
MHA-Net	74.46	85.36
CBRNet	74.55	85.42
BCTNet	75.04	85.74
CICF-Net	**75.17**	**85.83**

D. Comparison With the State-of-the-Art Methods

We compare CICF-Net with several SOTA methods on the Massachusetts Building Dataset, the WHU Aerial Building Dataset, and the Inria Aerial Image Labeling Dataset. Three well-known and widely-used classical segmentation models are included in our comparison: U-Net [81], DeeplabV3+ [82], Swin Transformer [48], and UNetFormer [44]. Swin Transformer has shown exceptional performance in various natural image tasks. The decoder used in conjunction with the Swin Transformer is UperHead. UNetFormer is a notable network designed specifically for remote sensing, achieving excellent results on multiclass datasets, such as Potsdam Dataset. ResNet50 is used as the backbone for feature extraction in U-Net and DeeplabV3+, while ResNet18 is utilized in UNetFormer based on its official configuration. Additionally, we compare our results with that of the SOTA methods reported in published papers, which include MA-FCN [23], EU-Net [20], MAP-Net [18], MHA-Net [22], CBR-Net [24], ASF-Net [29], E-D-Net [30], SST [58], BuildFormer [61], and BCTNet [60]. Among these, SST and BCT-Net are Transformer-based methods, while UNetFormer, BCT-Net are dedicated to combining CNN and Transformer. Moreover, the majority of these comparative methods indeed incorporate multiscale approaches as well as utilizing multiscale features, which include MA-FCN, EU-Net, MAP-Net, MHA-Net, CBR-Net, ASF-Net, and BCT-Net. The quantitative accuracy metrics for these approaches are directly sourced from their respective publications.

We perform a quantitative comparison analysis between CICF-Net and other selected methods on the Massachusetts Buildings Dataset (Table VI). The results clearly demonstrate that CICF-Net achieves SOTA performance levels, as evidenced by its superior IoU and F1-scores compared to existing methods. We also select several representative images for visualization

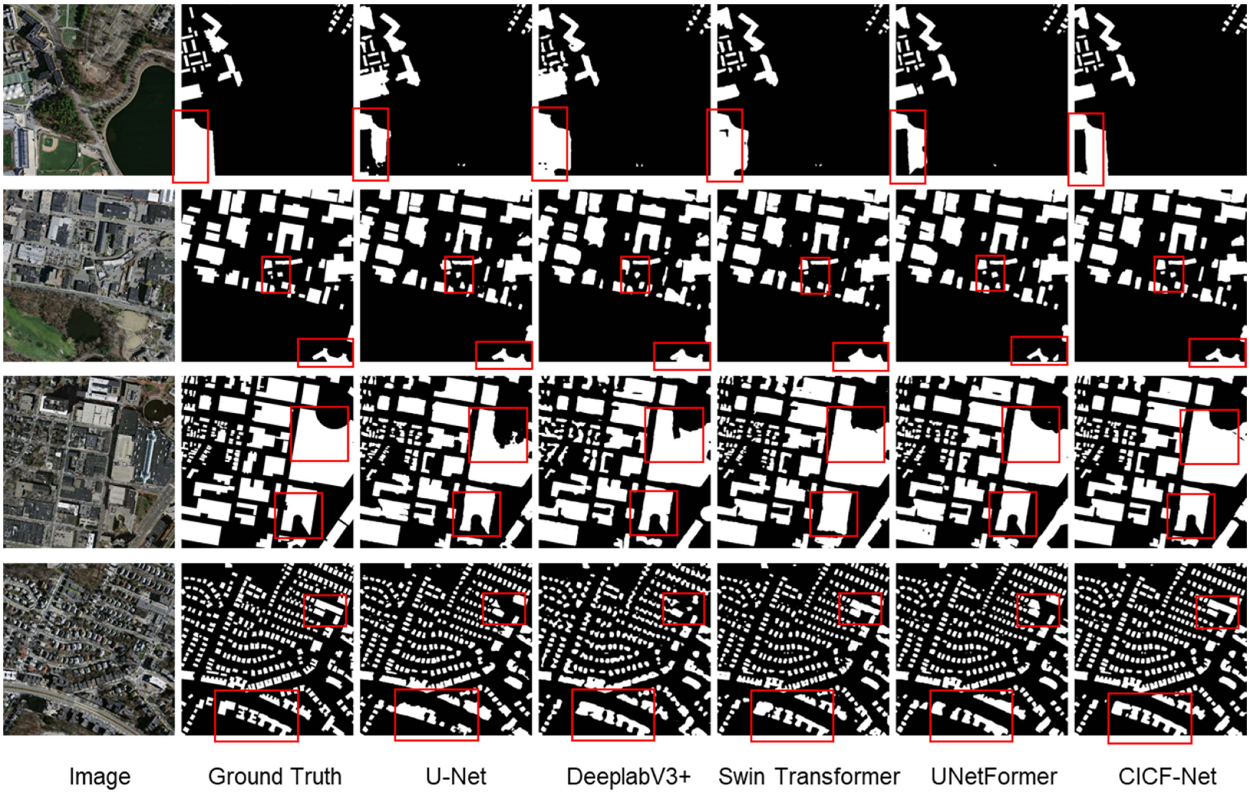


Fig. 8. Visualization comparisons of different methods for building extraction on the Massachusetts Building Dataset.

analysis (Fig. 8). The visualization results show that CICF-Net produces smoother boundaries and more regularly shaped building segmentation masks, while the other four methods tend to produce rounder building shapes. In the bottom left corner of the first row, within the red box, there is a stadium that should not be classified as a building based on its spectral characteristics. DeeplabV3+ and Swin Transformer both identify the central part as a building, whereas U-Net and UNetFormer correctly identify that the central part is not a building but fails to fully segment the building surrounding the stadium. Only CICF-Net accurately segments this particular building based on its spectral features. In the second row, Swin Transformer fails to detect the small building in the red box of the upper portion and inaccurately segments the boundaries of the building in the red box of the bottom right corner. Meanwhile, UNetFormer also struggles to completely extract this building, showing less precision in both shape and boundaries. Notably, CICF-Net accurately extracts both of these buildings. Regarding the third row, the upper red box highlights a building with a U-shaped opening. Neither of the pure convolutional methods successfully segments the entire building, and the boundary detected by the Swin Transformer is less precise when compared to the proposed CICF-Net. Moving to the lower red box, CICF-Net exhibits precise and straight boundaries for the building, whereas Swin Transformer mistakenly classifies its surrounding pixels as buildings. UNetFormer has achieved comparable results in the two boxes compared to CICF-Net, but still falls short in capturing the details of other buildings. Turning to the examples in the bottom row of Fig. 8, the upper red box features a relatively larger rectangular building. None of the two pure convolutional methods successfully detect it or accurately capture its extent, responding only to a limited region. And they both demonstrate poor building extraction results in the lower red box. The same goes for UNetFormer using a CNN-based backbone. In contrast, CICF-Net exhibits superior performance by accurately extracting this complete building and delineating the boundary, outperforming Swin Transformer.

Meanwhile, the proposed method achieves the SOTA accuracy on the WHU Aerial Dataset (Table VII). Similarly, visualization analysis is performed (Fig. 9) to provide support for our findings. In the first row of Fig. 9, only CICF-Net achieved an accurate prediction, whereas the other four methods more or less misclassified the nonbuilding pixels inside the two red boxes as buildings. In the second row, U-Net introduces two small nonbuilding regions within the red box, causing minor holes in the segmentation mask. Both DeeplabV3+ and Swin Transformer display jagged boundaries, with the latter incorrectly predicting vehicles on the road as buildings. What is more, UNet-Former mistakenly segments the large impervious surface in the red box as buildings. In the case of the third row, U-Net falls short in capturing the complete building in the upper right portion of the red box, while Swin Transformer overlooks the presence of small-size buildings in the lower left part. A small shaded area along the upper edge of the red box is mistakenly classified as building pixels by both Swin Transformer and UNetFormer. Regarding the bottom row, only CICF-Net excels

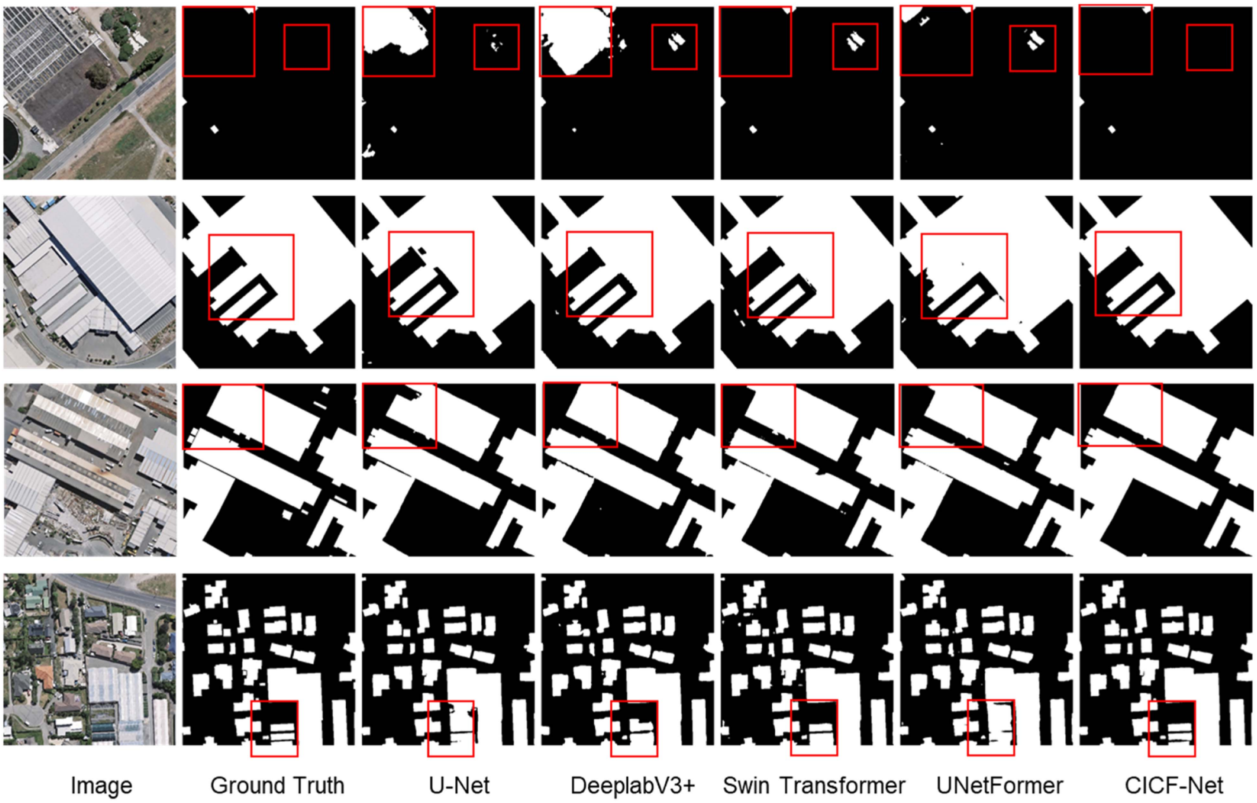


Fig. 9. Visualization comparisons of different methods for building extraction on the WHU Aerial Building Dataset.

TABLE VII
QUANTITATIVE COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON THE WHU AERIAL BUILDING DATASET

Method	IoU (%)	F1 (%)
U-Net	87.76	93.50
Deeplab V3+	87.83	93.52
Swin Transformer	90.27	94.89
UNetFormer	88.60	93.95
MA-FCN	90.70	95.15
EU-Net	90.56	95.04
MAP-Net	90.86	95.21
CBRNet	91.40	95.51
SST	90.48	94.97
BCTNet	91.15	95.37
BuildFormer	91.44	95.53
CICF-Net	**91.45**	**95.53**

TABLE VIII
QUANTITATIVE COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON THE INRIA AERIAL IMAGE LABELING DATASET

Method	IoU (%)	F1 (%)
U-Net	73.01	84.40
Deeplab V3+	76.81	86.88
Swin Transformer	78.86	88.18
UnetFormer	76.47	86.67
MA-FCN	79.67	88.68
EU-Net	80.50	89.20
ASF-Net	80.20	-
E-D-Net	79.78	-
CBRNet	81.10	89.56
SST	79.42	87.99
CICF-Net	**81.28**	**89.67**

in delineating the boundaries among the three buildings within the red box, consisting of two buildings with blue-green roofs and one building with a white roof.

Furthermore, quantitative comparisons and visualization of the segmentation results on the Inria Aerial Image Labeling Dataset show that CICF-Net consistently demonstrated superior performance (Table VIII). In the red box of the first row in Fig. 10, DeeplabV3+ and CICF-Net stand out as the only methods that detect two small concealed buildings. Moving to the two red boxes in the second row, CICF-Net showcases the most precise segmentation boundaries and shapes. In the third row, all methods identify the holes within the bottom red box though the ground truth incorrectly labeled the green space in the central hole as building pixels. Regarding the upper red box, UNet fails to perceive the dark gray roof, while DeeplabV3+ produces incomplete segmentation results. The red box in the bottom row

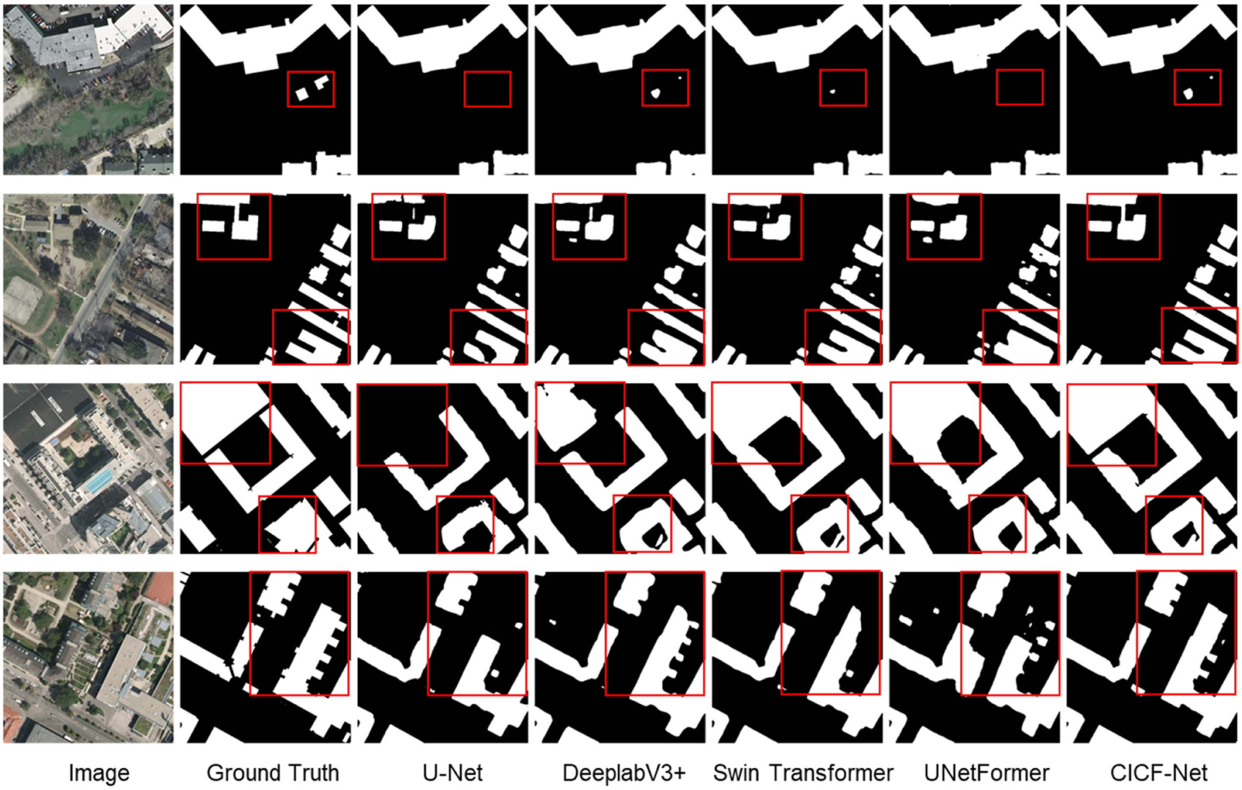


Fig. 10. Visualization comparisons of different methods for building extraction on the Inria Aerial Image Labeling Dataset.

encompasses two groups of buildings. While CICF-Net does not fully extract the small buildings on the right, it presents the most direct and nonrounded building boundaries. As for the other group, CICF-Net accurately segments its shape. This leads to a clear and precise segmentation result.

E. Visualization of Feature Maps

We visualize the feature maps of the four stages of CICF-Net on the Massachusetts Building Dataset, providing a more intuitive depiction of the combination of high- and low-frequency features. All visualization heatmaps are generated using Grad-CAM [83], a gradient-based deep network visualization method. GradCAM can visualize feature maps from any layer, which can show the differences in features at different stages. The heatmaps can not only intuitively demonstrate the receptive fields captured by the model but also reflect the differences in high- and low-frequency features extracted at different stages of the model, since the receptive field correlates with the frequency characteristics of the feature heatmaps. The colors in the heatmaps from red to blue represent perception from strong to weak. It can be observed that the feature maps from the four stages follow the logic designed in our CICF-Net: progressing from shallow to deep layers, our method shifts from perceiving small sharp high-frequency features to broader low-frequency features (Fig. 11). In the initial two stages, feature extraction primarily emphasizes local areas with intense changes, capturing sharp high-frequency information such as lines and edges. However, it encounters challenges in paying attention to larger buildings within shallow layers. Typically, CICF-Net becomes more adept at positioning these large buildings accurately by the third stage, as demonstrated inside the red boxes in Fig. 11. As the depth increases, the boundaries of the feature maps become smoother and rounder, with a more global receptive field. At this point, the low-frequency branch of our method plays a more crucial role, allowing better attention to low-frequency information such as overall spatial range and structure. The perceptual areas and contours of large buildings inside the red boxes become complete and more distinct.

Removing the base maps of the feature heatmaps and presenting their Fourier spectrums, we can intuitively observe the differences in frequency characteristics of feature extraction across the stages (Fig. 12). We select examples that align with Fig. 11 for demonstration. The bright areas are more dispersed, with many horizontal and vertical stripes in the Fourier spectrum of feature maps in shallow stages, reflecting the presence of more sharp edge features (such as building and street boundaries), which indicate typical high-frequency features in the spectrum. On the other hand, the contrast between bright and dark areas becomes more pronounced in the Fourier spectrum of feature maps in deep stages. The bright areas are smaller in size, and there are fewer stripes that are concentrated. These characteristics indicate typical low-frequency features in the spectrum.

Besides, the heatmaps of other methods have also been shown for comparison on two typical test images, each featuring a larger building surrounded by numerous smaller ones (Figs. 13 and 14),

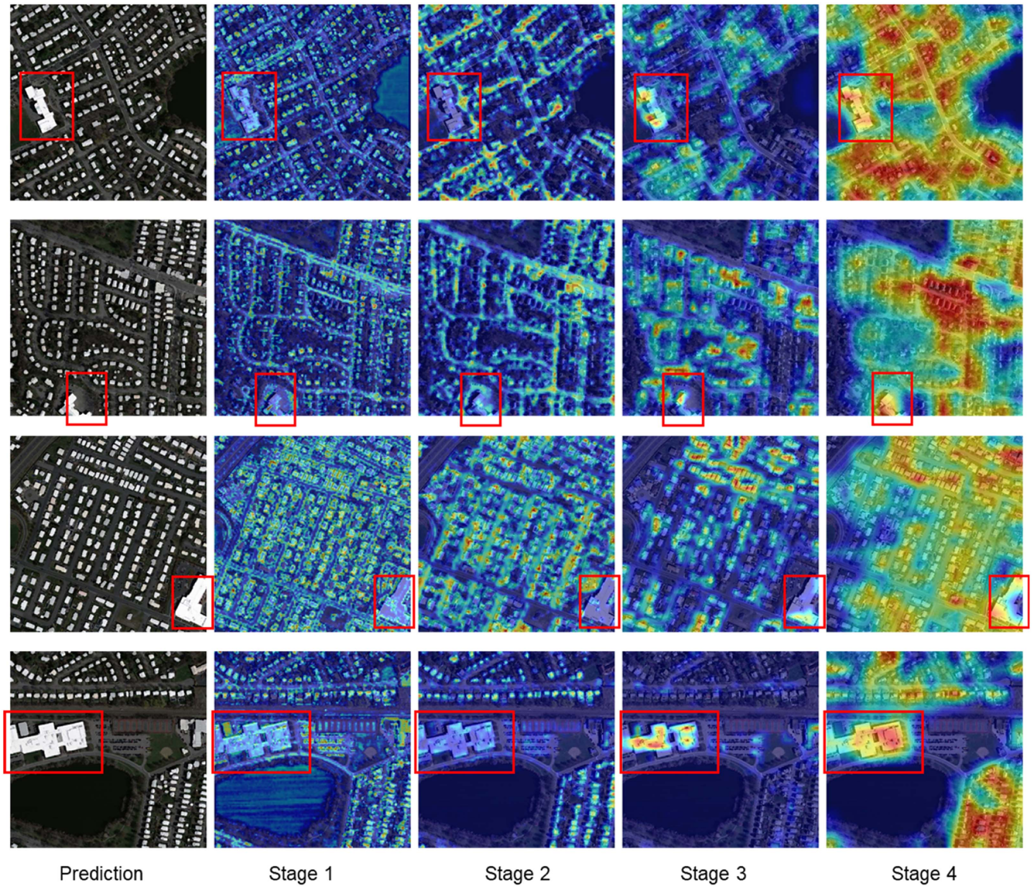


Fig. 11. Visualization of the feature maps across the four stages of CICF-net on the Massachusetts Building Dataset. The first column demonstrates the overlay of prediction results (buildings depicted in white) onto the original image (serving as the base map). The colors in the feature maps from red to blue represent perception from strong to weak.

which emphasize the advantages of CICF-Net in extracting and combining high- and low-frequency features.

We select Deeplabv3+ (pure convolutional method), and Swin Transformer (pure Transformer method) for comparison. Deeplabv3+ exhibits limited receptive fields. Although it gradually focuses on larger buildings, it fails to achieve the expansive receptive field for this size of images as the other three methods including Transformers. Swin Transformer presents more fragmented and irregular features in the focus of shallow layers compared to CICF-Net, which lacks the fine positioning of high-frequency features. It shows some instability and abruptness, potentially leading to the loss of features that have been attended to in the shallow layers, possibly associated with its shifted window operation. The two methods struggle to extract and integrate both high- and low-frequency features. It can also be seen from their stage-by-stage Fourier spectrograms that DeeplabV3+ focuses more on high-frequency features at all four stages compared to CICF-Net, while Swin Transformer extracts more low-frequency features. The heatmaps and Fourier spectrograms of CICF-Net confirm that the four stages gradually transition from emphasizing high- to low-frequency information, realizing the stage-by-stage gradual fusion of high and low-frequency features, which contributes to accurate building extraction.

IV. Discussion

Our CICF-Net excels in building extraction tasks from high-spatial resolution remote sensing images as demonstrated by comparison with existing SOTA methods. By incorporating key modules such as the iConvFormer Block, the cascaded fusion strategy, and GCUPerNet, we successfully capitalize on the strengths of high- and low-frequency features at different depths, as well as achieving efficient multilevel spatial interactions among buildings and between buildings and backgrounds. Consequently, CICF-Net effectively addresses the challenges of missing-detected small buildings and discontinuous large buildings. The visualization results on both datasets show accurate and smooth predicted boundaries, thereby confirming the superior performance of our network. The promising results validate the potential for practical application of our method for building extraction from high-spatial resolution remote sensing images.

We have made an initial attempt to explore the combination of high- and low-frequency branches in the iConvFormer Block, but we solely use ConvNeXt and sparse ViT as the pairing for the high- and low-frequency branches. With the rapid development of deep learning and artificial intelligence, there have been numerous CNN and Transformer networks, offering countless

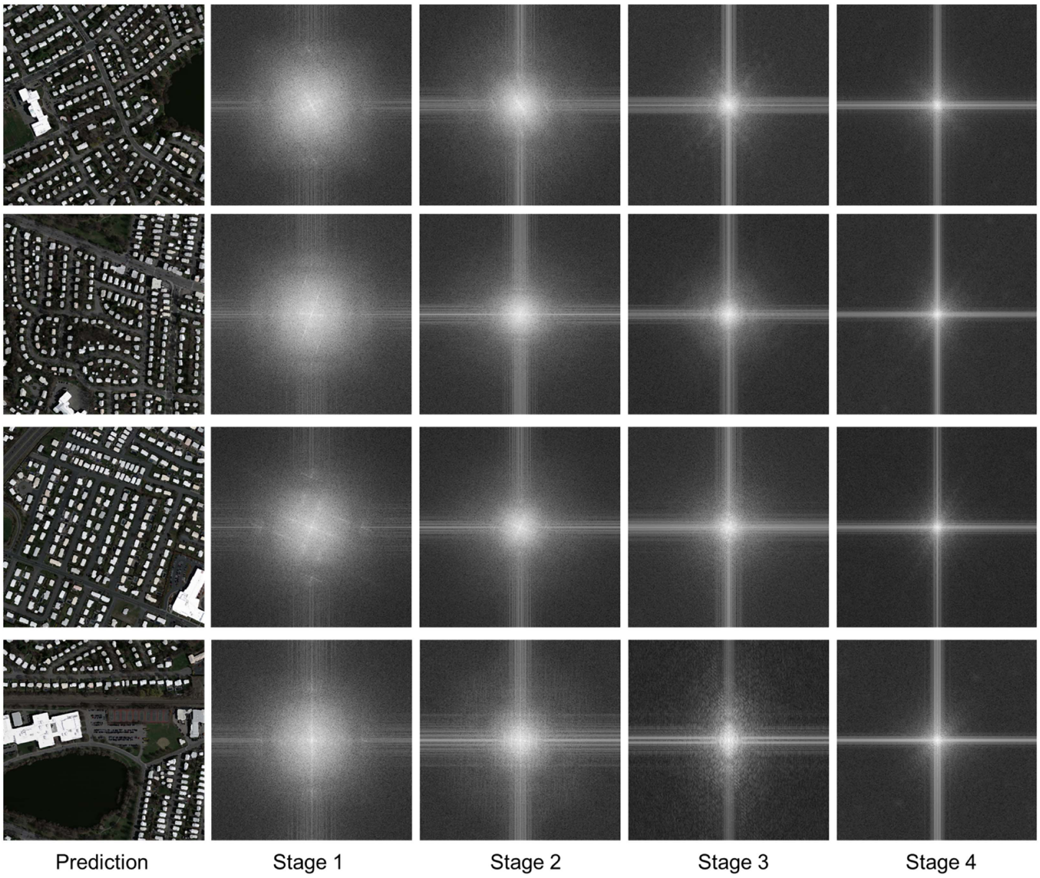


Fig. 12. Fourier spectrum of the feature maps at each stage across the four stages of CICF-net on the Massachusetts Building Dataset. The first column demonstrates the overlay of prediction results (buildings depicted in white) onto the original image (serving as the base map).

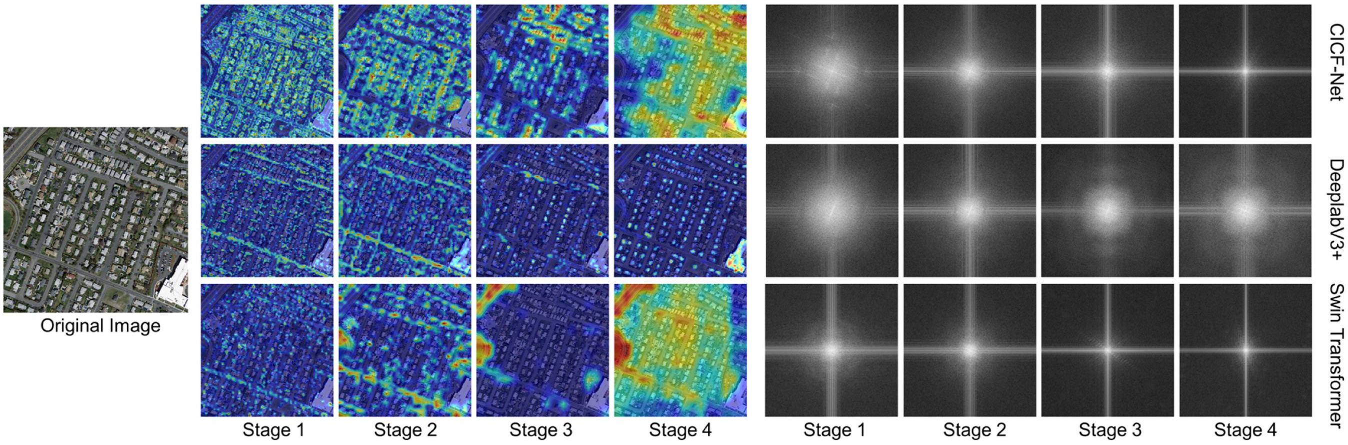


Fig. 13. Visualization of the feature maps and their Fourier spectrograms across the four stages of different methods on the Massachusetts Building Dataset. The colors in the figure from red to blue represent perception from strong to weak.

possible pairs for feature extraction. However, due to limited computational resources, we have not extensively explored other combination pairs of high- and low-frequency branches in this study. In the future, we anticipate that more optimal combination pairs can be discovered and integrated into our framework, enabling improved building extraction performance.

Besides, we can further explore the generalization ability of our model. We have applied the model to our proprietary satellite dataset. Our dataset is derived from the SuperView-1 satellite, which fuses panchromatic band with a resolution of 0.5 m and multispectral bands with a resolution of 2 m. We have applied CICF-Net trained on the Massachusetts Building Dataset to

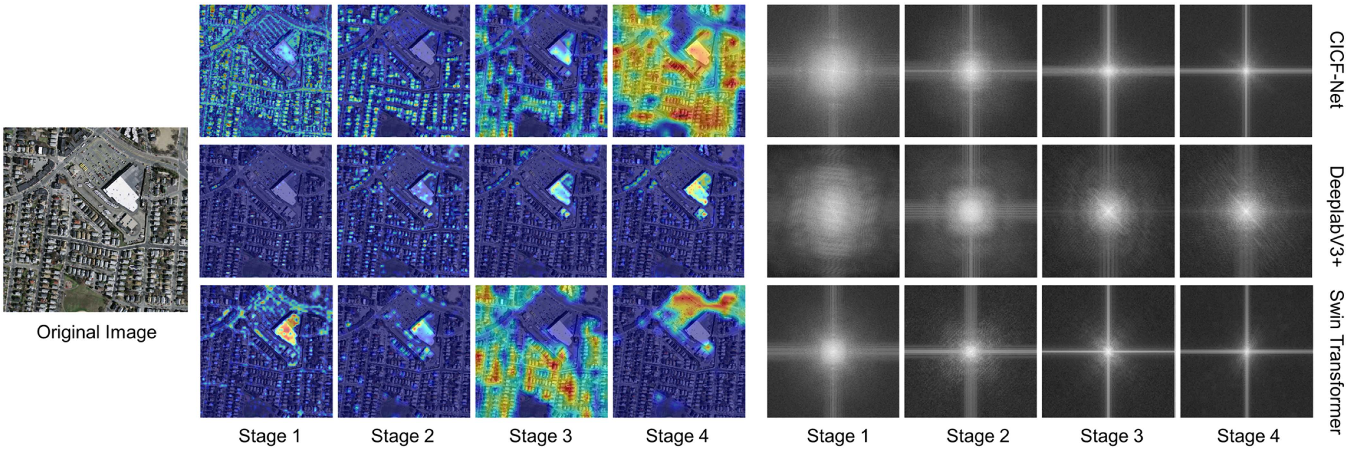


Fig. 14. Visualization of the feature maps and their Fourier spectrograms across the four stages of different methods on the Massachusetts Building Dataset. The colors in the figure from red to blue represent perception from strong to weak.

these unlabeled satellite images, which yields preliminary building segmentation results. This verifies that our model exhibits a notable degree of generalization, enabling its application in a wide range of practical scenarios. However, this part of the experimental results is not currently suitable for inclusion in this article due to the confidentiality of the data. It is possible to explore the applicability of our model across data of varying quality or types during the process. Therefore, it could provide further insights to combine our idea with existing relevant studies in spectral variability and data noise [84], [85].

V. Conclusion

In this study, we propose a novel network CICF-Net for building extraction from high-spatial resolution remote sensing images. We focus on the challenges encountered in building extraction by optimizing both the encoder and decoder parts of the network. In the encoder, we design the iConvFormer Block and a cascaded strategy to efficiently fuse multiscale high- and low-frequency features. In the decoder part, we propose the GCUperNet, which can efficiently extract multilevel spatial interactions. Through the design and integration of these modules, we have effectively alleviated the issues of omissions of small buildings and discontinuities in large buildings, enabling precise and straight segmentation boundaries for building extraction. Experimental results demonstrate that the proposed method exhibits outstanding performance, achieving the best result when compared to other SOTA methods. This solidly proves the effectiveness and superiority of CICF-Net, which couples high- and low-frequency features for building extraction.

References

[1] M. M. Rathore, A. Ahmad, A. Paul, and S. Rho, “Urban planning and building smart cities based on the Internet of Things using big data analytics,” *Comput. Netw.*, vol. 101, pp. 63–80, Jun. 2016.

[2] Z. Chen et al., “EGDE-Net: A building change detection method for high-resolution remote sensing imagery based on edge guidance and differential enhancement,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 191, pp. 203–222, Sep. 2022.

[3] M. Bouziani, K. Goïta, and D.-C. He, “Automatic change detection of buildings in urban environment from very high spatial resolution images using existing geodatabase and prior knowledge,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 65, no. 1, pp. 143–153, Jan. 2010.

[4] B. Chai and P. Li, “An ensemble method for monitoring land cover changes in urban areas using dense Landsat time series data,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 195, pp. 29–42, Jan. 2023.

[5] X. Wang et al., “Double U-Net (W-Net): A change detection network with two heads for remote sensing imagery,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 122, Aug. 2023, Art. no. 103456.

[6] L. Dong and J. Shan, “A comprehensive review of earthquake-induced building damage detection with remote sensing techniques,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 84, pp. 85–99, Oct. 2013.

[7] Y. Qing et al., “Operational earthquake-induced building damage assessment using CNN-based direct remote sensing change detection on superpixel level,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, Aug. 2022, Art. no. 102899.

[8] B. Adriano et al., “Learning from multimodal and multitemporal earth observation data for building damage mapping,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 175, pp. 132–143, May 2021.

[9] P. Xiao, X. Feng, R. An, and S. Zhao, “Segmentation of multispectral high-resolution satellite imagery using log Gabor filters,” *Int. J. Remote Sens.*, vol. 31, no. 6, pp. 1427–1439, Mar. 2010.

[10] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, “Unified perceptual parsing for scene understanding,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 432–448.

[11] W. Sweldens, “The lifting scheme: A construction of second generation wavelets,” *SIAM J. Math. Anal.*, vol. 29, no. 2, pp. 511–546, Mar. 1998.

[12] F. W. Campbell and J. G. Robson, “Application of Fourier analysis to the visibility of gratings,” *J. Physiol.*, vol. 197, no. 3, pp. 551–566, 1968.

[13] J. Bai, L. Yuan, S.-T. Xia, S. Yan, Z. Li, and W. Liu, “Improving vision transformers by revisiting high-frequency components,” in *Proc. Eur. Conf. Comput. Vis.*, 2022, vol. 13684, pp. 1–18.

[14] F. Karsli and O. Kahya, “Detecting the buildings from airborne laser scanner data by using Fourier transform,” *Exp. Techn.*, vol. 36, no. 1, pp. 5–17, 2012.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[16] A. Vaswani et al., “Attention is all you need,” in *Proc. 31st Conf. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 5998–6008.

[17] J. Li et al., “Next-ViT: Next generation vision transformer for efficient deployment in realistic industrial scenarios,” 2022, *arXiv:2207.05501*.

[18] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, “MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6169–6181, Jul. 2021.

[19] Y. Shi, Q. Li, and X. X. Zhu, “Building segmentation through a gated graph convolutional neural network with deep structured feature embedding,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 184–197, Jan. 2020.

[20] W. Kang, Y. Xiang, F. Wang, and H. You, "EU-Net: An efficient fully convolutional network for building extraction from optical remote sensing images," *Remote Sens.*, vol. 11, no. 23, Nov. 2019, Art. no. 2813.

[21] Y. Xie et al., "Refined extraction of building outlines from high-resolution remote sensing imagery based on a multifeature convolutional neural network and morphological filtering," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1842–1855, 2020.

[22] J. Cai and Y. Chen, "MHA-Net: Multipath hybrid attention network for building footprint extraction from high-resolution remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5807–5817, 2021.

[23] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using CNN and regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2178–2189, Mar. 2020.

[24] H. Guo, B. Du, L. Zhang, and X. Su, "A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 183, pp. 240–252, Jan. 2022.

[25] J. Chen et al., "Memory-contrastive unsupervised domain adaptation for building extraction of high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5605615.

[26] S. Chen, Y. Ogawa, C. Zhao, and Y. Sekimoto, "Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmentation approach," *ISPRS J. Photogrammetry Remote Sens.*, vol. 195, pp. 129–152, Jan. 2023.

[27] W. Qiu, L. Gu, F. Gao, and T. Jiang, "Building extraction from very high-resolution remote sensing images using refine-UNet," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 6002905.

[28] W. Li, K. Sun, H. Zhao, W. Li, J. Wei, and S. Gao, "Extracting buildings from high-resolution remote sensing images by deep ConvNets equipped with structural-cue-guided feature alignment," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 113, Sep. 2022, Art. no. 102970.

[29] J. Chen, Y. Jiang, L. Luo, and W. Gong, "ASF-Net: Adaptive screening feature network for building footprint extraction from remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4706413.

[30] Y. Zhu, Z. Liang, J. Yan, G. Chen, and X. Wang, "E-D-Net: Automatic building extraction from high-resolution aerial images with boundary information," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4595–4606, 2021.

[31] H. Huang, Y. Chen, and R. Wang, "A lightweight network for building extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5614812.

[32] H. Hosseinpour, F. Samadzadegan, and F. D. Javan, "CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 184, pp. 96–115, Feb. 2022.

[33] S. Yao, L. Li, G. Cheng, and B. Zhang, "Analyzing long-term high-rise building areas changes using deep learning and multisource satellite images," *Remote Sens.*, vol. 15, no. 9, May 2023, Art. no. 2427.

[34] W. Li et al., "Joint semantic–geometric learning for polygonal building segmentation from high-resolution remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 201, pp. 26–37, Jul. 2023.

[35] X. Wang, K. Tan, P. Du, C. Pan, and J. Ding, "A unified multiscale learning framework for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4508319.

[36] Y. Li, D. Hong, C. Li, J. Yao, and J. Chanussot, "HD-Net: High-resolution decoupled network for building footprint extraction via deeply supervised body and boundary decomposition," *ISPRS J. Photogrammetry Remote Sens.*, vol. 209, pp. 51–65, Mar. 2024.

[37] X. Wang et al., "A high-resolution feature difference attention network for the application of building change detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, Aug. 2022, Art. no. 102950.

[38] W. Ding, X. Li, G. Li, and Y. Wei, "Global relational reasoning with spatial temporal graph interaction networks for skeleton-based action recognition," *Signal Process. Image Commun.*, vol. 83, Apr. 2020, Art. no. 115776.

[39] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng, "Strip pooling: Rethinking spatial pooling for scene parsing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4002–4011.

[40] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "CoTr: Efficiently bridging CNN and transformer for 3D medical image segmentation," in *Proc. Med. Image Comput. Comput. Assist. Intervention*, 2021, pp. 171–180.

[41] K. Li et al., "UniFormer: Unifying convolution and self-attention for visual recognition," 2022, *arXiv:2201.09450*.

[42] H. Wu et al., "CvT: Introducing convolutions to vision transformers," 2021, *arXiv:2103.15808*.

[43] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7242–7252.

[44] L. Wang et al., "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 190, pp. 196–214, Aug. 2022.

[45] L. Ding et al., "Looking outside the window: Wide-context transformer for the semantic segmentation of high-resolution remote sensing images," 2022, *arXiv:2106.15754*.

[46] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure transformer network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5224713.

[47] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2021, *arXiv:2010.11929*.

[48] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," 2021, *arXiv:2103.14030*.

[49] H. Bao, L. Dong, and F. Wei, "BEiT: BERT pre-training of image transformers," 2021, *arXiv:2106.08254*.

[50] W. Li et al., "SepViT: Separable vision transformer," 2022, *arXiv:2203.15380*.

[51] Y. Liu et al., "A survey of visual transformers," 2021, *arXiv:2111.06091*.

[52] D. Hong et al., "SpectralGPT: Spectral remote sensing foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2024.3362475.

[53] D. Muhtar, Z. Li, F. Gu, X. Zhang, and P. Xiao, "LHRS-Bot: Empowering remote sensing with VGI-enhanced large multimodal language model," 2024, *arXiv: 2402.02544*.

[54] J. Pan et al., "EdgeViTs: Competing light-weight CNNs on mobile devices with vision transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 294–311.

[55] C. Si, W. Yu, P. Zhou, Y. Zhou, X. Wang, and S. Yan, "Inception transformer," 2022, *arXiv:2212.03035*.

[56] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin Transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408715.

[57] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, "Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408820.

[58] K. Chen, Z. Zou, and Z. Shi, "Building extraction from remote sensing images with sparse token transformers," *Remote Sens.*, vol. 13, no. 21, Nov. 2021, Art. no. 4441.

[59] 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.

[60] L. Xu, Y. Li, J. Xu, Y. Zhang, and L. Guo, "BCTNet: Bi-branch cross-fusion transformer for building footprint extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4402014.

[61] L. Wang, S. Fang, X. Meng, and R. Li, "Building extraction with vision transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5625711.

[62] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2117–2125.

[63] M. Hu, Y. Li, L. Fang, and S. Wang, "A^2-FPN: Attention aggregation based feature pyramid network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15338–15347.

[64] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7036–7045.

[65] G. Zhang, Z. Li, C. Tang, J. Li, and X. Hu, "CEDNet: A cascade encoder-decoder network for dense prediction," 2023, *arXiv:2302.06052*.

[66] R. Hang, P. Yang, F. Zhou, and Q. Liu, "Multiscale progressive segmentation network for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5412012.

[67] Y. Rao, W. Zhao, Y. Tang, J. Zhou, S.-N. Lim, and J. Lu, "HorNet: Efficient high-order spatial interactions with recursive gated convolutions," 2022, *arXiv:2207.14284*.

[68] D. Muhtar, X. Zhang, P. Xiao, Z. Li, and F. Gu, "CMID: A unified self-supervised learning framework for remote sensing image understanding," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5607817.

[69] Y. Long et al., "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-AID," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4205–4230, 2021.

[70] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15979–15988.

[71] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11966–11976.

[72] M. A. Islam, S. Jia, and N. D. B. Bruce, "How much position information do convolutional neural networks encode?" 2020, *arXiv:2001.08248*.

[73] X. Chu et al., "Conditional positional encodings for vision transformers," 2021, *arXiv:2102.10882*.

[74] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," 2017, *arXiv: 1612.01105*.

[75] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li, "Dice loss for data-imbalanced NLP Tasks," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 465–476.

[76] M. Berman, A. R. Triki, and M. B. Blaschko, "The Lovasz-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4413–4421.

[77] A. Bokhovkin and E. Burnaev, "Boundary loss for remote sensing imagery semantic segmentation," 2019, *arXiv:1905.07852*.

[78] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[79] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 3226–3229.

[80] X. Xie, P. Zhou, H. Li, Z. Lin, and S. Yan, "Adan: Adaptive Nesterov momentum algorithm for faster optimizing deep models," 2023, *arXiv:2208.06677*.

[81] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.

[82] 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, vol. 11211, pp. 833–851.

[83] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.

[84] Z. Li, X. Zhang, and P. Xiao, "One model is enough: Toward multiclass weakly supervised remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4503513.

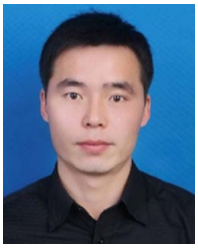
[85] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

Xinyang Chen received the B.S. degree in geographic information science from Jilin University, Changchun, China, in 2021. She is currently working toward the M.S. degree in cartography and geographical information system with Nanjing University, Nanjing, China.

Her research interests include semantic segmentation and deep learning for remote sensing.

Pengfeng Xiao (Senior Member, IEEE) was born in Hunan, China, in 1979. He received the B.M. degree in land resource management from Hunan Normal University, Changsha, China, in 2002, and the Ph.D. degree in cartography and geographical information system from Nanjing University, Nanjing, China, in 2007.

From 2007 to 2009, he was a Lecturer with the School of Geography and Ocean Science, Nanjing University, where he was an Associate Professor from 2010 to 2018. Since 2019, he has been a Professor with Nanjing University. From 2011 to 2012, he was a Visiting Scholar with the Department of Geography, University of Giessen, Germany, and from 2014 to 2015, with the Department of Environmental Science, Policy, and Management, University of California at Berkeley, USA. He has authored 4 books and more than 100 articles. His current research interests include high-resolution remote sensing image analysis, remote sensing of snow cover, and land use and land cover change.



Xueliang Zhang (Senior Member, IEEE) received the B.S. degree in geographical information systems and the Ph.D. degree in remote sensing of resources and environment from Nanjing University, Nanjing, China, in 2010 and 2015, respectively.

From 2014 to 2015, he was a Visiting Student with the Informatics Institute, University of Missouri, Columbia, MO, USA. From 2016 to 2018, he was an Associate Researcher with the Department of Geographic Information Science, Nanjing University, where he is currently an Associate Professor with the Department of Geographic Information Science. His research interests include high-resolution remote sensing image analysis, semantic segmentation, and deep learning for remote sensing.

Dilxat Muhtar received the B.S. degree in geographic information science in 2022 from Nanjing University, Nanjing, China, where he is currently working toward the M.S. degree in cartography and geographic information system.

His research interests include self-supervised and transfer learning for remote sensing.

Luhan Wang received the B.S. degree in geographical information science in 2021 from Nanjing University, Nanjing, China, where she is currently working toward the M.S. degree in cartography and geographic information system.

Her research interests include semantic segmentation, object detection, unsupervised domain adaptation, and deep learning for remote sensing.