MDCGA-Net: Multiscale Direction Context-Aware Network With Global Attention for Building Extraction From Remote Sensing Images

Penghui Niu, Junhua Gu, Yajuan Zhang, Ping Zhang, Taotao Cai, Wenjia Xu, and Jungong Han, *Senior Member, IEEE*

***Abstract*—Building extraction from remote sensing images (RSIs) requires exploring multiscale boundary detailed information and extracting it completely, which is challenging but indispensable. However, existing solutions tend to augment feature information solely through multiscale fusion and apply attention mechanisms to focus on feature relationships within a single layer while ignoring the multiscale information, which affects segmentation results. Therefore, enhancing the capability of the network to adaptively capture multiscale information and capture the global relationship of features remains a pivotal challenge in overcoming the aforementioned hurdles. To address the preceding challenge, we propose a Multiscale Direction Context-aware network with Global Attention (MDCGA-Net), employing a classic encoder–decoder architecture enhanced with direction information and global attention flow. Specifically, in the encoder part, the multiscale layer is used to extract contextual information from the interlayer. In addition, the multiscale direction context-aware module is adopted to adaptively acquire multiscale information. In the decoder part, we propose a global attention gate module to capture discriminative features. Furthermore, we construct an operation of attention feature flow to obtain the global relationship among the different features with long-range dependencies, which guarantees the integrity of results. Finally, we have performed comprehensive experiments on three public datasets to showcase the efficacy and efficiency of MDCGA-Net in building extraction.**

***Index Terms*—Building extraction, deep learning (DL), global attention, multiscale direction context-aware.**

Manuscript received 8 January 2024; revised 19 March 2024; accepted 7 April 2024. Date of publication 12 April 2024; date of current version 22 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62206085, in part by the Natural Resources Science and Technology Plan Project of Hebei Province under Grant 454-0601-YBN-IBBM, in part by the Innovation Capacity Improvement Plan Project of Hebei Province under Grant 22567603H, and in part by the Interdisciplinary postgraduate Training Program of Hebei University of Technology under Grant HEBUT-Y-XKJC-2022101. *(Corresponding author: Junhua Gu.)*

Penghui Niu and Yajuan Zhang are with the School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China (e-mail: qingxinqazxsw@163.com; zhangyajuan@scse.hebut.edu.cn).

Junhua Gu and Ping Zhang are with the School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China, and also with the Hebei Province Key Laboratory of Big Data Calculation, Hebei University of Technology, Tianjin 300401, China (e-mail: jhguhebut@163.com; zhangping@hebut.edu.cn).

Taotao Cai is with the Macquarie University, Sydney, NSW 2109, Australia (e-mail: taotao.cai@mq.edu.au).

Wenjia Xu is with the Hebei Prospecting Institute of Hydrogeology and Engineering Geological, Shijiazhuang 050021, China (e-mail: 562153204@qq.com).

Jungong Han is with the Department of Computer Science, University of Sheffield, S10 2TN Sheffield, U.K. (e-mail: jungonghan77@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2024.3387969

I. Introduction

BUILDING extraction from remote sensing images (RSIs) is a crucial research direction, which aims to assign corresponding labels to each pixel of buildings [1]. As one of the core elements in a basic geographic information database, accurate and immediate acquisition of building information is of great significance in remote sensing mapping applications, such as urban planning [2], land use [3], and disaster risk assessment [4]. Many works [1], [5], [6] have achieved great success in improving the accuracy of building extraction. However, due to the particular imaging conditions, RSIs have some problems, such as scale diversity and environmental interference. Therefore, accurate building extraction encounters two primary challenges: 1) the absence of detailed information across multiple scales at boundaries and 2) the inadequacy of long-range extraction.

Existing algorithms for building extraction in RSIs can be divided into two categories: 1) traditional image processing methods and 2) deep learning (DL)-based methods. Traditional image processing methods usually use manually designed feature operators (e.g., corner detection operators [7] and histograms [8]) and auxiliary features (e.g., digital surface models [9], GIS data [10], and light detection and ranging [11]) to capture building features. However, such methods are often characterized by being time-consuming and label-intensive. [12].

In contrast, the integration of DL into the RSI analysis has gained significant attention due to its end-to-end trainable property. Numerous studies have highlighted the effectiveness of fully convolutional networks (FCNs) [13], a classic semantic segmentation DL algorithm, in extracting buildings from RSIs. Following this, several methods such as SegNet [14], DeepLab [15], and U-Net [16] have been developed, leveraging the capabilities of FCNs to improve coarse-resolution segmentation results and extract multiscale feature information for building extraction. U-Net has gained popularity in recent research because of its capacity to deliver more accurate boundary details. This is achieved by preserving spatial information, a feature that has been extensively researched and documented [17], [18]. Another noteworthy development in this domain is the introduction of transformer-based networks, such as the vision transformer (ViT) [19], which have been employed for building extraction tasks. Among these, the Swin transformer stands out for its hierarchical structure with shifted windows

that facilitate cross-window connections, effectively addressing the challenges posed by multiscale and high-resolution images [20]. However, it is important to note that transformer-based algorithms tend to be computationally intensive.

To address the challenge of the absence of detailed information across multiple scales at boundaries, some researchers introduce auxiliary modules to refine the boundary information [21]. Alternatively, some other studies introduce multiscale encoder architecture [22] and the atrous spatial pyramid pooling (ASPP) [23] to obtain the multiscale contextual information. In tackling the challenge of incomplete long-range extraction, a common approach involves obtaining fused features of different resolutions. This is typically achieved by capturing relationships among long-range features through the application of attention mechanisms [24], [25], [26], [27].

While the existing works demonstrate commendable performance, they encounter three notable limitations. First, they only enhance the multiscale information by combining and expanding contextual information, ignoring the ability to adaptively extract boundary detail information by adaptive-weighting the importance of the features. In addition, there is a risk of losing location information for buildings during feature fusion and receptive field expansion. Furthermore, these methods only obtain the channel or spatial correlation of a single-feature layer through attention mechanisms, neglecting the global feature relationships of different levels with long-range dependencies.

In this article, we address existing challenges by introducing directional information and a global attention flow. The proposed model, named Multiscale Direction Context-aware network with Global Attention (MDCGA-Net), adopts a classic encoder–decoder architecture. Specifically, in the encoder part, the multiscale layer (MSL), which is a basic module of the Res2Net [28], is employed to explore the feature representation ability of the interlayer at various scales. Subsequently, to alleviate the problem of the loss of detailed information, an MDCM embedded with direction-aware and position-sensitive information is designed to adaptively obtain contextual information at multiple scales. This way enhances the ability to explore the ability to explore the boundary detailed information. Moving to the decoder, a global attention gate (GAG) module is designed to enhance feature distinctiveness across all channels using GAGs, ensuring complete building extraction. In addition, an operation of attention feature flow is constructed to guide the attention fluid from the high level of the network to the low level, which captures the global relationships among the different features with long-range dependencies. We evaluate MDCGA-Net's performance on three public datasets, demonstrating its superiority over state-of-the-art (SOTA) baselines in terms of accuracy and efficiency. The main contributions of this article are summarized as follows.

1) We propose a novel building extraction network structure, MDCGA-Net, which focuses on multiscale contextual information and global attention feature extraction to elevate segmentation results.
2) We propose the MDCM to dynamically assign weights to building boundary detail information by introducing an attention operation with directional information into multiscale contextual features.
3) We propose the GAGM with global attention flow, combining attention weights across different feature layers. This facilitates the extraction of correlations among global features.

The rest of this article is organized as follows. In Section II, we introduce several studies related to multiscale contextual information and the attention mechanism of building extraction from RSIs. In Section III, we introduce the detailed structure of the proposed method. In Section IV, we evaluate the performance of our proposed methods. Finally, Section V concludes this article.

II. Related Work

A. Multiscale Contextual Information

The ability to extract contextual information about the building of various scales, which effectively improves the boundary accuracy of segmentation results, remains challenging for building extraction. The feature map effectively enhances the representation ability of multiscale boundary information by fusing the features of different resolutions. Thus, some studies explore the description capability of multiscale contextual information of models. MAP-Net [22] proposes a multiparallel path with different resolutions to extract multiscale features, which contain multilevel spatial and semantic information. BDA-Net [29] introduces a multiscale feature fusion module, which contains three streams with different resolutions as input, to enhance the feature representation ability of buildings under various scales. Inspired by MAP-Net, Chen et al. [30] propose a transformer-based network to extract building footprints of different resolutions by the Swin transformer. Wang et al. [31] propose a novel transformer block that is composed of a series of BuildFormer Blocks to capture features of different scales.

The feature fusion of different levels with fixed receptive fields lacks essential contextual information. The atrous convolution has been demonstrated to be an effective strategy to expand the receptive field of networks, which prompts the model to extract more global information. The authors in [21] and [33] use atrous convolution with different dilation rates embedded into the model to capture contextual information about buildings of various scales. Chan et al. [34] propose a multibranch pyramid pooling to capture multiscale features to compensate for the lost information in the encoder.

However, the aforementioned methods solely capture the multiscale contextual information by fusing features of different resolutions, which results in inaccuracy boundary segmentation of buildings due to a lack of the ability to capture directional and positional information adaptively. To go one step further, we propose a multiscale direction context-aware module to adaptively obtain more spatial detailed contextual information through reweighting the importance of the multiple scales.

B. Attention Mechanism

The attention mechanism is a method to capture the discriminative features and the relationships among the global dependencies, which efficaciously solves the problem of incomplete extraction of buildings in RSIs. For example, the authors in [22] and [35] introduce channel attention to redistribute the weight

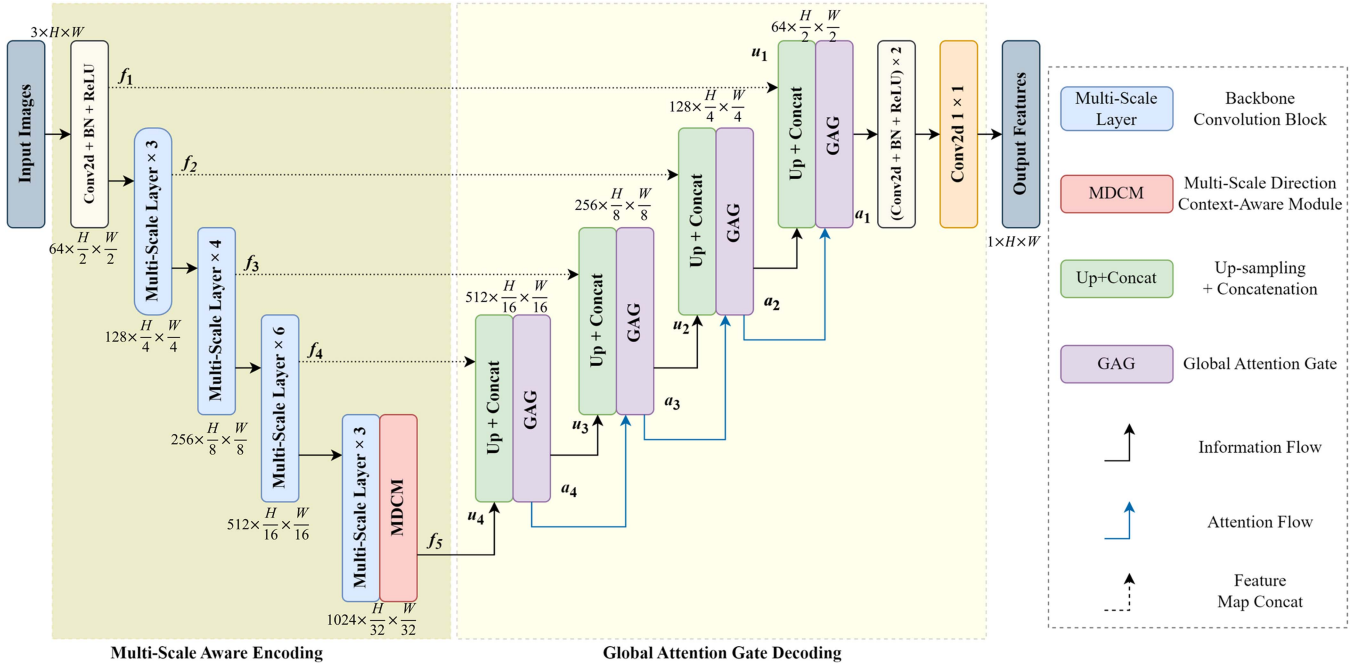

Fig. 1. Structure of the proposed MDCGA-Net, which is composed of two parts. Multiscale aware encoding and global attention decoding. The right-hand side of the figure shows the notes of some modules.

of each channel by learning feature maps and then adaptively reconstructing more detailed information about buildings. Zhou et al. [33] use channel attention as a gate for building extraction that automatically learns more local information of varying scales of buildings and suppresses irrelevant areas. However, those methods ignore the attention mechanism of spatial dimensions, which results in the spatial details loss of the feature maps. To rationally utilize low-level features, DANet [36] proposes a spatial attention fusion module embedded with the dense network, which uses the semantic information from the high level to suppress the redundant information and noises by reweighting low-level features. Some studies [29], [37] propose the new attention modules, which compose both the channel and spatial attention to explore the correlations between different feature maps of buildings. To better obtain long-range contextual feature representations, MTPA-Net [38] introduces the dual-attention mechanism, which composes a positional attention module and a channel attention module. Wang et al. [31] propose a global context path, which is composed of a window-based linear multihead self-attention to capture global dependencies.

Unfortunately, the attention mechanism in all of these methods focuses on the information of a single-feature layer, which lacks global feature relationships of different levels with long-range dependencies. For this specific constraint, we propose a GAG module that is built by a special structure of attention flow to capture the discriminative features with global correlations.

III. Proposed Method

The proposed MDCGA-Net, as shown in Fig. 1, is a universal encoder–decoder structure. In the encoder part, we use an MSL that is a basic module of the Res2Net [28] to extract essential contextual information of the interlayer. Moreover, we propose a multiscale direction context-aware module to adaptively obtain contextual information at multiple scales. A direction context-aware attention module that introduces the direction and position-sensitive module integrates coordinate information during the encoding processes of two factorized parallel 1-D features, which is introduced into the MDCM to achieve adaptive obtain the unequal-weight of the multiscale features. In the decoder part, to capture the discriminative features and the relationships among the features of long-range dependencies, a GAG module is used to guide the attention fluid from the high level to the low level. In the GAGM, the features extracted from the low level could be guided by the attention from high-level layers. In this way, we can obtain global attention and the relationships among the features of different levels. Thus, we can obtain the complete building segmentation results because the final feature map has the global relationship of features between different layers. The details of our method are as follows.

A. Multiscale Aware Encoding

As shown in Fig. 1, we abstract the overall structure of our proposed encoder as five consecutive convolutional layers, where the last four layers consist of several MSLs. Notice that we employ the same strategy to set the scale as 4 of MSL in each block as same as the literature [28] to ensure the reasoning ability and sufficient parallel computations within a single GPU of the network. The encoder takes the input image $I \in \mathbb{R}^{3\times H\times W}$ to obtain the features of different levels by passing through the five convolutional layers. More semantic feature

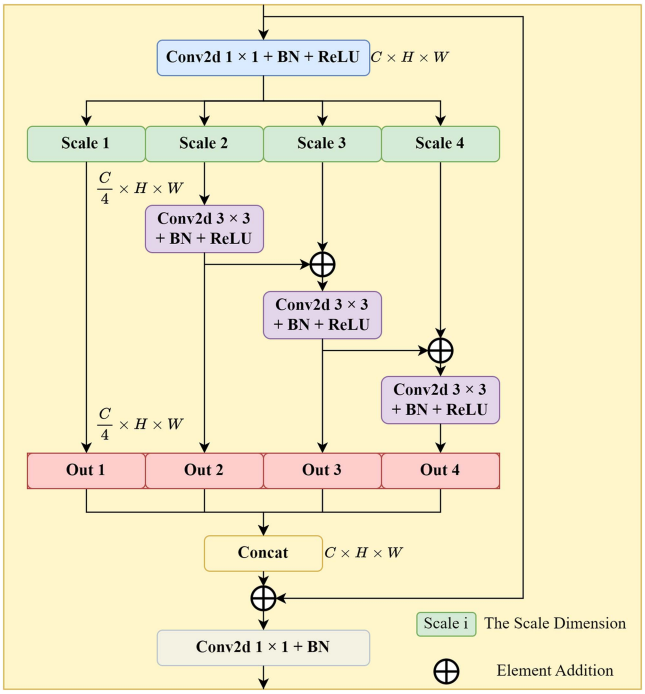

Fig. 2. Structure of the proposed MSL.

information can be obtained with each downsampling operation. We denote feature maps of each layer as $f_i(i \in [1, 2, 3, 4, 5])$ with different channels and resolution sizes. The size of f_i is $2^{i-1} \times 64 \times (H/2^i) \times (W/2^i)(i \in [1, 2, 3, 4, 5])$. To alleviate the problem of the loss of the multiscale boundary detailed information, we propose an MDCM employed behind the fifth MSL. The DCAM is introduced into the MDCM to enhance the ability to adaptively capture the detailed boundary features of the multiscale contextual information. After the encoder part, we can obtain a feature map that captures multiscale contextual features with detailed boundary information for the decoder part to get the refinement results of building segmentation from RSIs.

1) Multiscale Layer: Obtaining essential multiscale context-aware information on various scales is essential to solve the problem of losing the detailed boundary information of building extraction from RSIs. Inspired by recent work [28], in this research, we use an MSL that further explores the feature representation ability of the interlayer at various scales for building extraction. We can enlarge the receptive field by employing the MSL to solve the problem of the omission of detailed information. The MSL structure is shown in Fig. 2. In the MSL, the output feature contains a larger receptive field and both global and local information due to the subfeature by splitting the input feature map.

2) Multiscale Direction Context-Aware Module: Learning more contextual information from the scale diversity of the different buildings in RSIs plays a prominent role in building extraction. The multiple parallel atrous filters with different atrous rates are introduced into the ASPP module to capture the multiscale contextual features. As mentioned in Section II, the ASPP module has been employed in building extraction from RSIs to enhance the feature representation ability at multiple scales.

However, most existing building extraction methods that apply the ASPP module lack the proper direction-aware and position-sensitive information due to the multiscale contextual information only obtained from the fused different features. The ASPP module pays the same attention to multiscale information and the features captured from ASPP are of equal weight. The detailed information may also be lost by the repeated down-sampling operations at consecutive layers of the encoder, which results in inaccuracy boundary segmentation.

To this end, we design a multiscale direction context-aware module to improve the ability to explore multiscale detailed information adaptively. This alleviates the problem of positional information loss and effectively captures the proper boundary information of buildings with scale diversity. Differing from the convolutional ASPP module, a direction context-aware attention module (DCAM) is proposed to adaptively capture the boundary detail information by adaptive-weighting the importance of the different scale features. The DCAM pays different attention to different-scale features, which alleviates the problem of the loss of spatial information (e.g., boundary information).

The overall structure of the MDCM is illustrated on the left-hand side of Fig. 3. First, the MDCM applies five paths to extract the building information at multiple scales. Specifically, the first path has no operation to preserve the original input feature. In the middle three branches, we apply three 3×3 atrous convolutions with rates $= (3,\ 6,\ 12)$, respectively, and feed the feature maps to BN and ReLU to effectively capture multiscale information. Note that the strategy to set the rates is to ensure that the receptive fields corresponding to the atrous rates can cover buildings of all sizes in the datasets [40]. To extract global context information, we apply adaptive global average pooling on the last path. Then, the resulting feature map passes through a convolutional operation with 1×1 convolution, BN, and ReLU. After that, a bilinearly interpolate operation is applied to obtain the feature map at the original input dimension. Second, the feature maps from all the paths are then fed into DCAMs to adaptively capture the more precise direction and position information of buildings. Finally, those feature maps from all the branches are concatenated and then obtain the final feature map with rich multiscale feature information through a convolutional operation with another 1×1 convolution, BN, and ReLU.

To improve the ability to adaptively capture the multiscale contextual information, we use an attention mechanism to weight the different-scale features. The attention mechanism has been proven to be an effective method to enhance the ability of feature extraction to improve the building segmentation accuracy. Motivated by the coordinate attention (CA) blocks [41], the DCAM is introduced to guide the encoder to focus on the direction information of the feature map, which enhances the ability to adaptively obtain the detailed boundary information of the proposed method. Moreover, differing from the channel attention that only focuses on the weight distribution of different channels, the proposed DCAM also focuses on recalibrating features from the spatial dimension, which can effectively alleviate the positional information loss due to the numerous

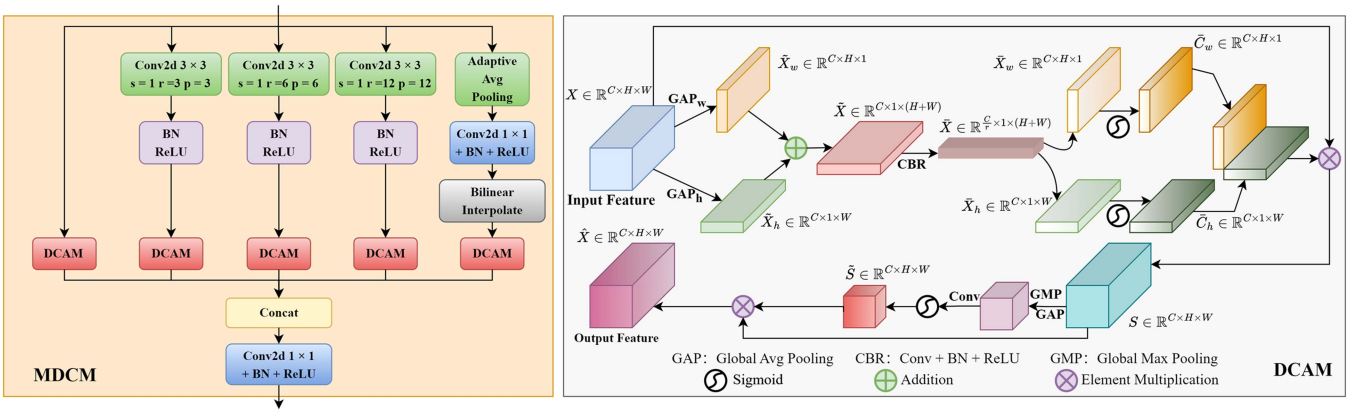

Fig. 3. Structure of the proposed MDCM. The right-hand side of the figure shows the structure of the DCAM in detail.

downsampling. The structure of the DCAM is illustrated on the right-hand side of Fig. 3, and the detailed description is as follows.

First, we apply global average pooling to the input feature map $X \in \mathbb{R}^{C\times H\times W}$ along the horizontal direction and the vertical direction, respectively, and the output feature maps are concatenated to generate a descriptor $\tilde{X} \in \mathbb{R}^{C\times 1\times (H+W)}$. This process can be formulated as

$$\tilde{X} = \left[\tilde{X}_w,\ \tilde{X}_h\right] = \left[F_{gw}(X),\ F_{gh}(X)\right] \tag{1}$$

where $[\cdot,\cdot]$ denotes the concatenation of the feature maps, $F_{gw}(\cdot)$ and $F_{gh}(\cdot)$ represent the global average pooling along the height dimension and width dimension, respectively. Subsequently, we apply a convolution operation to transform the descriptor $\tilde{X}$ to $\bar{X}$, for which the formula is as follows:

$$\bar{X} = \Phi\left(\text{conv}\left(\tilde{X}\right)\right) \tag{2}$$

where $\text{conv}(\cdot)$ denotes a 1×1 convolution operation, and $\Phi(\cdot)$ denotes the BN and ReLU, and we can obtain the feature map $\bar{X} \in \mathbb{R}^{\frac{C}{r}\times 1\times (H+W)}$ that capture the information both in the horizontal coordinate and the vertical coordinate. Note that r is the reduction rate as in the SE block to deduct dimension. After that, we split the $\bar{X}$ along the width dimension and height dimension to obtain the features with the horizontal direction and the vertical direction, respectively. Then, the features pass through two convolution operations, and we obtain the attention weights. These attention weights are guided to capture detailed boundary features by introducing directional information. The adaptive weights are obtained by focusing on the importance of different channel features. The specific formula for this process is as follows:

$$\bar{C}_w = \sigma\left(\text{conv}\left(\bar{X}_w\right)\right) \tag{3}$$

$$\bar{C}_h = \sigma\left(\text{conv}\left(\bar{X}_h\right)\right) \tag{4}$$

where $\bar{X}_w \in \mathbb{R}^{C\times H\times 1}$ and $\bar{X}_h \in \mathbb{R}^{C\times 1\times W}$ are the two separate features along width dimension and height dimension, respectively, $\text{conv}(\cdot)$ denotes a 1×1 convolution operation, and $\sigma(\cdot)$ is the sigmoid function. Finally, the feature map $S \in \mathbb{R}^{C\times H\times W}$ with direction information can be formulated as

$$S = X * \bar{C}_w * \bar{C}_h \tag{5}$$

where $*$ denotes the element multiplication. Through this process, the output feature map S adaptively captures the information both in the horizontal direction and vertical direction.

As described earlier, the attention feature map S not only has multiscale information on the important features in the channels but also captures spatial information in both the horizontal direction and the vertical direction. However, the feature maps may lack spatial position information due to the consecutive downsampling in the encoder.

To this end, the DCAM is proposed to capture precise spatial position information through a spatial selective mechanism, which adaptively focuses on the most relevant spatial position regions by a spatial feature descriptor. Specifically, given the input feature map S passing through the global max pooling and global average pooling, respectively, we obtain two feature maps that capture rich detail information. Subsequently, we concatenate two feature maps along the channel dimension to enhance the global information, and then, a convolution operation is used to reduce the dimension. Moreover, a sigmoid function is applied to generate the spatial attention descriptor $\tilde{S} \in \mathbb{R}^{C\times H\times W}$, for which the formula is as follows:

$$\tilde{S} = \sigma\left(\text{conv}\left(\left[F_{\text{GMP}}(S), F_{\text{GAP}}(S)\right]\right)\right) \tag{6}$$

where $F_{\text{GMP}}(\cdot)$ and $F_{\text{GAP}}(\cdot)$ are the global max pooling and global average pooling, respectively, $[\cdot,\cdot]$ denotes the concatenation of the channel, conv denotes a 3×3 convolution, and $\sigma(\cdot)$ is the sigmoid function. Finally, the output feature map $\hat{X}$ of MDCAM is obtained as shown in the following formula:

$$\hat{X} = S * \tilde{S} \tag{7}$$

where $*$ denotes the element multiplication.

B. GAG Decoding

The decoder of most existing methods is designed to fuse the same scale feature maps of different levels through upsampling operations and skip connections. However, there still have some

challenges with the traditional decoders. 1) The fused feature map with the only same scale levels lacks the discriminative features due to the simple skip connection. 2) The upsampling operations ignore the relationships among the global feature relationships of different levels with long-range dependencies, resulting in the discontinuity of information and incomplete extraction. In fact, the attention mechanism of different feature layers can capture the discriminative features of the current layer, which also implies the selection strategy for the importance of different features.

To this end, we propose a decoder with a GAG module to enhance the proposed the ability of the method to capture the discriminative features. Moreover, we build a special structure of attention flow by using high-level attention as guidance to improve the generation of low-level attention, which obtain the global correlations of the different level features with long-range dependencies.

Our proposed decoder, as shown in Fig. 1, consists of four layers. The feature map of the last layer f_5 will restore the feature scale as the fourth layer by passing through an upsampling operation, and then, the f_5 is concatenated with the feature map f_4 from the encoder to get the feature map u_4. The similar processing of the next layer in the decoder will be repeated and we can obtain the fused features $u_i (i \in [1, 2, 3, 4])$. Note that the upsampling operation is composed of two convolutional operations (the kernel size is 3) and bilinear interpolation with the scale factor is set to 2. Different from the decoders of the traditional methods, we apply a GAGM to increase the feature distinction across all channels by GAGs. Specifically, an operation of attention feature flow is also constructed to guide the attention fluid from the high level to the low level, which captures the global relationships among the different features with long-range dependencies. Then, we can obtain feature map a_i from each layer.

First, the feature map f_5 that is passing through an upsampling operation is concatenated with feature map f_4 to obtain the feature map u_4, for which the formula is as follows:

$$u_4 = \text{conv}\left(\left[F_{\text{up}}\left(f_5\right), f_4\right]\right) \tag{8}$$

where $[\cdot, \cdot]$ denotes the concatenation of the feature maps and $F_{\text{up}}(\cdot)$ denotes the upsampling operation. Subsequently, we apply a GAG module, which consists of an efficient channel attention mechanism [42] to capture the discriminative information of the feature map by redistributing the weights of f_4 among the channels. Then, the attention feature map a_4 is obtained as shown in the following formula:

$$a_4 = f_4 * g_4 = f_4 * \sigma\left(\text{conv1}D_k\left(F_{\text{GAP}}\left(u_4\right)\right)\right) \tag{9}$$

where $F_{\text{GAP}}(\cdot)$ denotes the global average pooling, $\text{conv1}D_{1k}(\cdot)$ denotes the 1-D convolution (the kernel size is 5), σ denotes the sigmoid function, and $*$ is the element multiplication. Note that $g_4 \in \mathbb{R}^{C\times1\times1}$ is the channel weight descriptor. The GAG module is used to appropriately capture cross-channel interaction, and then, we obtain the attention feature map with more discriminative information.

Second, as depicted in Fig. 1, to ensure the feature map of each subsequent layer captures the discriminative features of the previous layer, we take the attention feature map $a_{i+1} (i \in [1, 2, 3])$ as the input of the next layer and fuse it with the feature map f_i from the corresponding layer of the encoder. Then, the attention feature map $a_i (i \in [1, 2, 3])$ of the current layer with discriminative information through the GAG module is obtained. In particular, inspired by the work in [43], we apply an attention fluid operation to aggregate global relationships among the different features with long-range dependencies. To update the attention descriptor $g_i (i \in [1, 2, 3])$, the descriptor $g_{i+1} (i \in [1, 2, 3])$ of the previous layer is used as an attention flow to enhance the descriptors. Specifically, as shown in Fig. 4, we not only take the output attention map a_{i+1} of the previous layer fused with f_i as the input feature map but also flow the attention descriptor g_{i+1} into the next layer. After the same attention operation as in the process of obtaining g_4 above, an attention descriptor $\hat{g}_i$ of the current layer is generated. Then, the descriptor g_{i+1} is concatenated with $\hat{g}_i$ and a sigmoid function is used to produce the final attention descriptor g_i. The specific formula for this process is as follows:

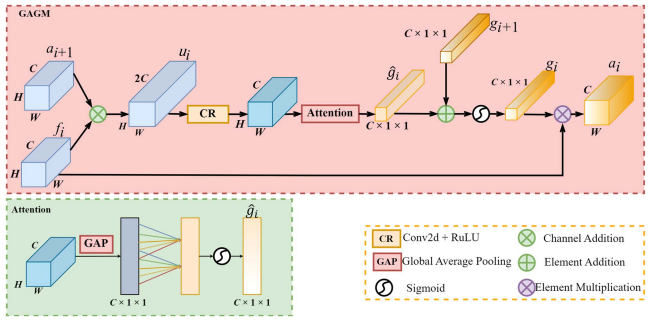

Fig. 4. Structure of the proposed GAGM. The figure shows the structure of the attention in detail.

$$u_l = \text{conv}\left(\left[F_{\text{up}}\left(a_{l+1}\right), f_l\right]\right) \tag{10}$$

$$g_l = \sigma\left(\text{conv1}D_k\left(F_{\text{GAP}}\left(u_l\right)\right)\right) \tag{11}$$

$$a_l = f_l * \sigma\left(\{g_l, g_{l+1}\}\right) \tag{12}$$

where $\{\cdot, \cdot\}$ denotes the element addition. The attention fluid operation is designed to produce a GAG of each layer in the decoder, which can obtain the relationships among the features of long-range dependencies.

Finally, the final feature map a_1 of the last layer undergoes an upsampling operation and two convolutional operations (the kernel size is 3) to restore to the same size as the original input images. Then, a 1×1 convolution is used to adjust the number of the channel to make the final prediction.

C. Loss Function

Data imbalance is a basic problem in the task of building extraction due to the difference in the pixels between the buildings and the background, which affects the accuracy of the building extraction in RSIs. Therefore, in this study, a joint loss is used to accelerate the network convergence and optimize our proposed network during the training. The joint loss L can be

formulated as

$$L = L_f + L_d \tag{13}$$

where L_f denotes the focal loss and L_d denotes the dice loss.

Focal loss, proposed in [44], can balance the weights of positive and negative samples by applying a modulating term to the cross entropy loss to solve the problem of imbalanced foreground and background. We use the focal loss as one component of the loss function to optimize the gradient descent and yield more detailed feature representations of buildings.

Dice loss was proposed to deal with situations where there is an imbalance between the foreground and background pixels by optimizing the Dice coefficient between the prediction results and the ground truth and was indicated for binary segmentation tasks [45]. We use dice loss, as a part of the joint loss, to balance the distribution of buildings and background by mining more foreground regions. The specific formulations of the joint loss are as follows:

$$\begin{aligned} L_f = & -\frac{1}{N}\sum_{i}^{N}(\alpha(1-P_i)^{\gamma} \times \hat{P}_i \log(P_i) \\ & + (1-\alpha)P_i^{\gamma} \times (1-\hat{P}_i)\log(1-P_i)) \end{aligned} \tag{14}$$

$$L_d = 1 - \frac{2 \times \sum_i^N P_i \hat{P}_i + \epsilon}{\sum_N^i (P_i^2 + \hat{P}_i^2) + \epsilon} \tag{15}$$

where P_i and $\hat{P}_i$ denote the predicted probability value of the pixel of sample i and the ground truth of the pixel of sample i, respectively. N denotes the total number of pixels of the input image. α and γ are hyperparameters, which can adjust the weights of the positive and negative samples. Their values are set to 0.25 and 2, respectively, similar to that in [44]. ϵ denotes the smoothing coefficient, which can prevent the denominator from being zero. It is set to 10^{-5}.

IV. Experimental Results and Discussions

In this section, we introduce the experimental setup and measure the performance of our proposed *MDCGA-Net* compared with SOTA baselines.

A. Dataset

To evaluate the effectiveness of our proposed method, three publicly building datasets are adopted comprehensively for the experiments, including the Massachusetts building dataset [46], the WHU building dataset [47], and the Inria aerial image labeling dataset [48].

The Massachusetts building dataset comprises 151 aerial images of the Boston area. The size of each image in the dataset is 1500 × 1500 pixels, and the spatial resolution of the image is 1 m. The dataset mainly covers 2.25 km^2 of urban and suburban areas, including residential areas, commercial areas, and industrial areas. In this dataset, 137 images are used for training, another 10 images are used for testing, and 4 images are used for validation. We followed the suggested partition of the original dataset released.

The WHU building dataset is widely used in building extraction tasks, and it is composed of 8188 aerial images with a resolution of 0.3 m. The dataset contains 220 000 buildings covering 450 km^2 in Christchurch, New Zealand. The size of each image in the dataset is 512 × 512 pixels, and the spatial resolution of the image is 0.3 m. Following the original partition of the dataset, 4736 images are used for training, 1036 images are used for validation, and 2416 images are used for testing. The dataset is labeled accurately, and all images come from a single region, which can better examine the building extraction ability of our proposed method.

The Inria aerial image labeling dataset is composed of 360 aerial images from five cities: Austin, Chicago, Kitsap, Tyrol, and Vienna. The size of each image of the dataset is 5000 × 5000 pixels, and the spatial resolution of the image is 0.3 m. In the official dataset, 180 images are labeled for the experiment. We followed the original partition, the first five images of every city for testing, and the remaining images for training and validation. Compared with the earlier two datasets, the background of the images in the Inria dataset is more complex, and the buildings in the images have greater scale diversity. Thus, it can better verify the effectiveness of the proposed modules.

B. Evaluation Metrics

To quantitatively evaluate the performance of our method, several common pixel-level metrics, which are widely used in building extraction tasks, are employed. These metrics include precision (P), recall (R), intersection over union (IoU), and F1-score (F1). The IoU indicates the intersection between building prediction results and the ground truth, which is an important metric to determine the segmentation results. The F1 is a metric that integrates P and R and maximizes their relationship to balance them. The specific formulations are as follows:

$$P = \frac{\text{TP}}{\text{TP+FP}} \tag{16}$$

$$R = \frac{\text{TP}}{\text{TP+FN}} \tag{17}$$

$$\text{IoU} = \frac{P_p \cap P_g}{P_p \cup P_g} \tag{18}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{19}$$

where TP, TN, FP, and FN denote the true positive, the true negative, the false positive, and the false negative, respectively. P_p and P_g denote the pixels of the predicted result and the ground truth, respectively.

C. Implementation Details

The code of our method is built on the PyTorch framework based on Python 3.9 programming language. We train the proposed method on a computer with Ubuntu 18.04, Intel i7-7800X CPU, and two NVIDIA GeForce RTX 2080Ti GPUs with graphic memory of 11 GB. During the training, we used the Adam optimizer with an initial learning rate of 10^{-4}, 0.9 momenta, 10^{-5} weight decay, and bath size 8. We set the learning

TABLE I
QUANTITATIVE COMPARISON WITH DIFFERENT METHODS ON THE MASSACHUSETTS BUILDING DATASET

Method	IoU (%)	P (%)	R (%)	F1 (%)
U-Net [16]	67.61	79.13	82.29	80.68
SegNet [14]	69.02	85.00	78.59	81.67
DeeplabV3+ [40]	76.80	87.35	86.40	86.88
BiSeNetV2 [49]	73.42	86.49	81.03	83.67
MAP-Net [22]	71.51	86.84	80.20	83.39
BOMSC-Net [33]	74.71	86.64	83.68	85.13
ED-DDL [18]	76.00	**89.58**	83.50	86.43
C3Net [23]	70.74	82.45	83.27	82.86
BuildFormer [31]	75.74	87.52	84.90	86.19
MDCGA-Net	**76.31**	86.86	**85.01**	**86.72**

TABLE II
QUANTITATIVE COMPARISON WITH DIFFERENT METHODS ON THE WHU BUILDING DATASET

Method	IoU (%)	P (%)	R (%)	F1 (%)
U-Net [16]	85.51	91.86	92.53	92.19
SegNet [14]	86.12	92.73	92.35	92.54
DeeplabV3+ [40]	89.61	94.68	94.36	94.52
BiSeNetV2 [49]	86.62	93.45	91.92	92.47
RDMSC-Net [17]	86.47	92.45	91.62	92.03
MAP-Net [22]	90.86	95.62	94.81	95.21
BOMSC-Net [33]	90.15	95.14	94.50	94.80
CGSANet [21]	91.55	95.11	**96.07**	95.59
LRAD-Net [50]	88.89	94.21	93.74	93.86
BuildFormer [31]	91.44	**95.65**	95.40	95.53
MDCGA-Net	**91.73**	95.61	95.79	**95.70**

TABLE III
QUANTITATIVE COMPARISON WITH DIFFERENT METHODS ON THE INRIA AERIAL IMAGE LABELING DATASET

Method	IoU (%)	P (%)	R (%)	F1 (%)
U-Net [16]	70.77	85.20	80.71	82.90
SegNet [14]	73.78	86.02	83.84	84.92
DeeplabV3+ [40]	76.81	87.38	86.43	86.90
BiSeNetV2 [49]	76.68	85.58	87.74	86.33
RDMSC-Net [17]	80.47	90.15	85.42	87.72
MAP-Net [22]	76.15	87.19	85.75	86.46
BOMSC-Net [33]	78.18	87.93	87.58	87.75
CGSANet [21]	80.90	90.22	88.68	89.94
ED-DDL [18]	81.00	**91.35**	**89.62**	**90.48**
C3Net [23]	76.21	86.94	85.95	86.42
LRAD-Net [50]	79.82	89.39	88.79	88.37
BuildFormer [31]	81.44	90.75	88.81	89.77
MDCGA-Net	**81.73**	90.76	89.15	89.95

rate to decay every 10 epochs during the training phase, and the minimum learning rate is 10^{-6}. Due to the limitation of computing resources, all images in the experiments are cropped to 512 × 512 pixels as the input. In addition, we implement several data augmentation strategies such as random horizontal and vertical flipping to increase the diversity of input images. The input images are randomly rotated with angles of 0°, 90°, 180°, and 270° and flipped horizontally and vertically for 0.3 probability. The MDCGA-Net is trained from scratch for 100 epochs on three datasets, and the training time was 2.3 h, 9.9 h, and 33.6 h, respectively.

D. Results and Discussions

To evaluate the performance of the proposed method, several SOTA methods are employed for comparison on MDCGA-Net. Those methods include the classic semantic segmentation networks (i.e., U-Net [16], SegNet [14], DeeplabV3+ [40], and BiSeNetV2 [49]) and the recent building extraction methods (i.e., RDMSC-Net [17], MAP-Net [22], BOMSC-Net [33], CGSANet [21], ED-DDL [18], C3Net [23], LRAD-Net [50], and BuildFormer [31]). All the experimental results are generated by open-source codes or provided by the authors. Note that the implementation settings of our method are consistent with the experimental details of all reproduced networks, including data augmentation strategies.

1) Quantitative Comparison: Similar to the work in [22], we use the common pixel-level metrics, such as P, R, IOU, and F1, as the evaluation of the proposed methods. The results of the quantitative evaluation of our proposed MDCGA-Net and the comparison methods on the Massachusetts building dataset, the WHU building dataset, and the Inria aerial image labeling dataset are shown in Tables I– III, respectively. Each row in these tables is the evaluation results of each method, and each column is the evaluation metrics. The highest records are marked in bold.

The performance of our proposed model MDCGA-Net is better than all the comparative methods on the three datasets. Specifically, compared with the recent SOTA method BuildFormer, the IoU of MDCGA-Net shows improvements of 0.57%, 0.29%, and 0.29% on the Massachusetts building dataset, the WHU building dataset, and the Inria aerial image labeling dataset, respectively. The F1 of our method is improved by 0.53%, 0.19%, and 0.18% on the three datasets, respectively, over BuildFormer. Compared with the classic segmentation networks, the recent building extraction methods show a great improvement in the quantitative evaluation of the three datasets. Among these methods, CGSANet, BOMSC-Net, and C3Net outperform the classic networks because these methods employ auxiliary modules (e.g., the boundary refinement module) to improve the ability to capture detailed features of building boundaries. The MAP-Net and BuildFormer methods use a multiple-path framework to extract multiscale features of diverse-scale buildings. The IoU and F1 of BiSeNetV2 and LRAD-Net are lower than other comparison methods since these two models are lightweight networks. The superior performance of ED-DDL since that ED-DDL captures rich multiscale contextual information by using dense blocks and restores the information through deconvolution layers. However, these methods solely capture the multiscale contextual information by busing features of different resolutions and lack global feature relationships of different levels with long-range dependencies. The MDCGA-Net exhibits a superior performance owning to the proposed MDCM can improve the ability to adaptively obtain more spatial detailed features and the GAGM can enhance the ability to capture the discriminative features with global correlations.

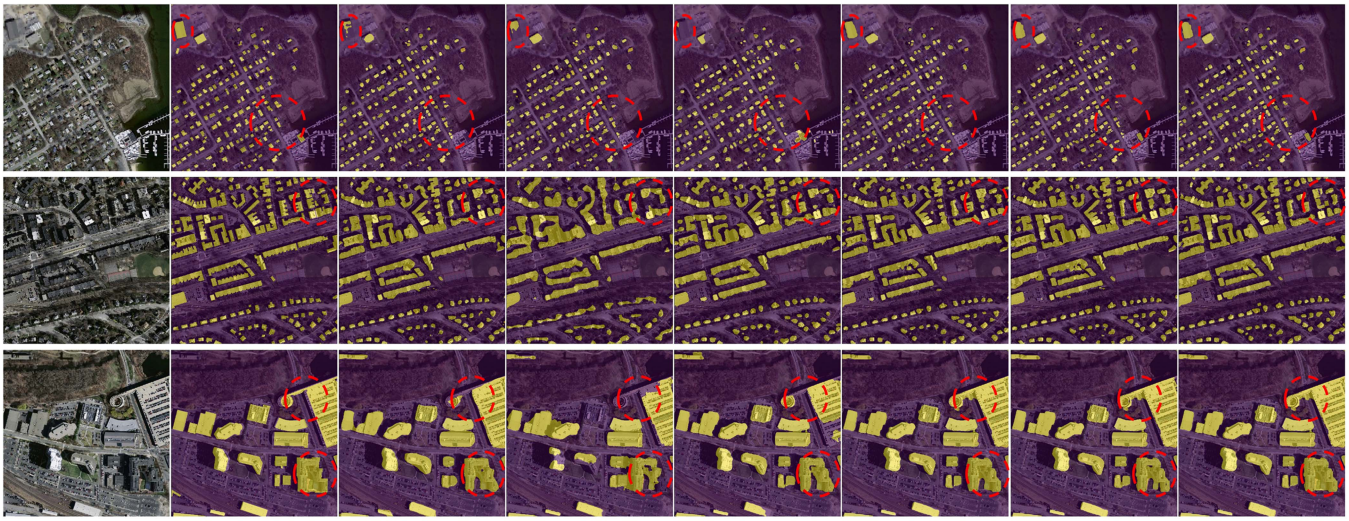

Fig. 5. Segmentation results on the Massachusetts building dataset. (a) Original images. (b) Ground truth. (c) U-Net. (d) SegNet. (e) Deeplabv3+. (f) MAP-Net. (g) BuildFormer. (h) MDCGA-Net. Among these methods, panels (c)–(e) show the classic semantic segmentation algorithms, and panels (f)–(g) show the recent similar enhancement algorithms of building extraction. (h) Our method. The key comparison areas are marked with red dashed circles.

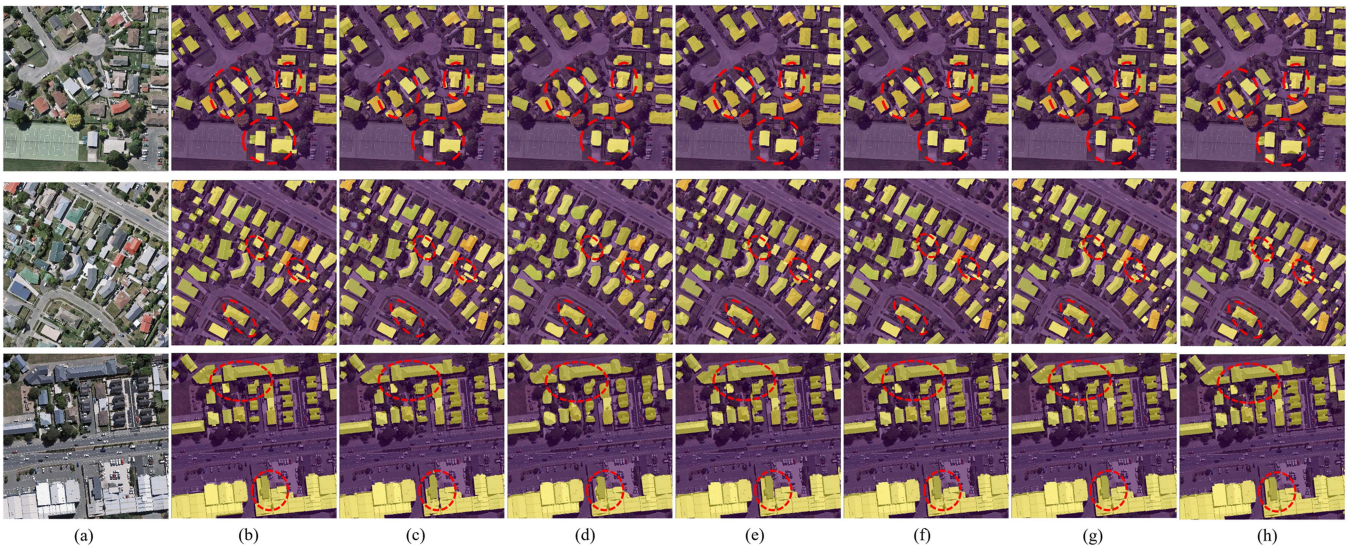

Fig. 6. Segmentation results on the WHU building dataset. (a) Original images. (b) Ground truth. (c) U-Net. (d) SegNet. (e) Deeplabv3+. (f) MAP-Net. (g) BuildFormer. (h) MDCGA-Net. Among these methods, panels (c)–(e) show the classic semantic segmentation algorithms, and panels (f)–(g) show the recent similar enhancement algorithms of building extraction. (h) Our method. The key comparison areas are marked with red dashed circles.

2) Qualitative Comparison: To further demonstrate the effectiveness of our method, we analyze the results of the comparison methods including classic semantic segmentation networks and recent building extraction methods from a qualitative perspective. The building extraction examples from each dataset are shown in Figs. 5– 7. We select some representative samples including buildings with complex shapes, shadow occlusions, and long-range coverage in every row in figures. Note that there are five rows (I–V) in Fig. 7 because the Inria aerial image labeling dataset contains the aerial images from five cities, including Austin (USA), Chicago (USA), Kitsap (USA), Tyrol (Austria), and Vienna (Austria). The first column (a) and the second column (b) of every figure are the original images and the corresponding ground truth, and columns (c)–(i) are the segmentation results from U-Net, SegNet, Deeplabv3+, MAP-Net, BuildFormer, and our proposed method, respectively.

Compared with other methods, the proposed MDCGA-Net exhibits a more advanced extraction performance for different scale buildings. As shown in the second row of Fig. 5 and the first row of Figs. 6 and 7, the segmentation results of U-Net, SegNet, and Deeplabv3+ omit more building boundaries. The MAP-Net and BuildFormer introduce the multiparallel path with different resolutions to extract multiscale features. These methods greatly alleviate the omissions of the boundary in the extraction results.

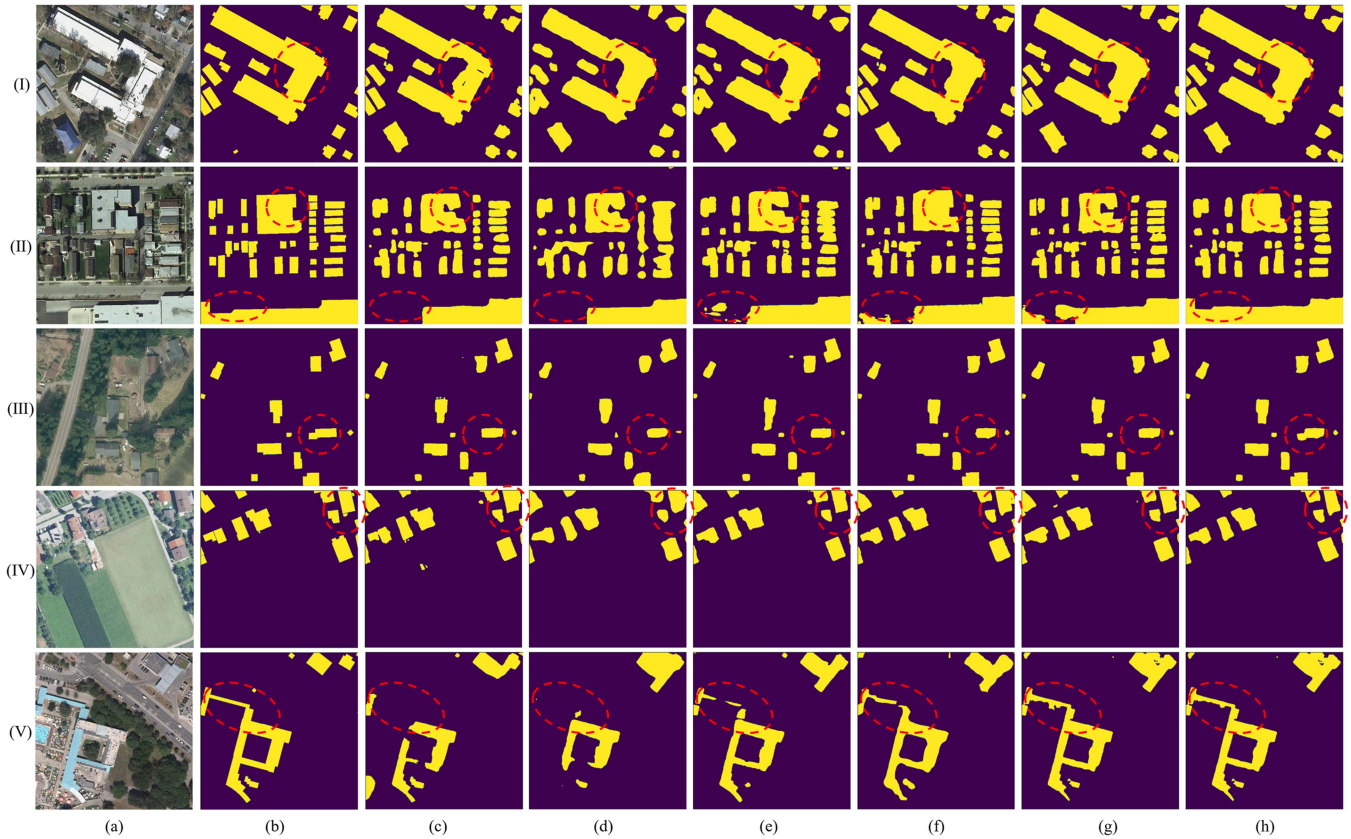

Fig. 7. Segmentation results on the Inria aerial image labeling dataset. (a) Original images. (b) Ground truth. (c) U-Net. (d) SegNet. (e) Deeplabv3+. (f) MAP-Net. (g) BuildFormer. (h) MDCGA-Net. (I) Austin. (II) Chicago. (III) Kitsap. (IV) Tyrol. (V) Vienna. Among these methods, panels (c)–(e) show the classic semantic segmentation algorithms, and panels (f)–(h) show the recent similar enhancement algorithms of building extraction. (i) Our method. The key comparison areas are marked with red dashed circles.

However, the ability to distinguish detailed information is still insufficient, resulting in a few missing parts of the building in the results. In contrast, the MDCGA-Net achieves the best extraction performance owing to its ability to adaptively capture multiscale contextual features by introducing the MDCM with directional information. In addition, the first row of Fig. 5 and the third row of Fig. 7 show that some small buildings or parts of buildings were not detected by the comparison methods due to spatial information loss. However, MDCGA-Net can detect more small buildings in the visual results because the proposed modules in this study can enhance the ability to capture discriminative spatial features. Moreover, from the third row of Figs. 5 and 6, especially in Fig. 7, we can see that the traditional semantic segmentation algorithms cannot completely extract large-scale buildings. The recent building extraction methods improve the segmentation integrity because those methods introduce multiscale information and attention mechanism. In contrast, the extract results of our method can guarantee structural integrity, which benefits from the proposed GAGM that enhances the ability to obtain the global relationship between different level features. In summary, our method MDCGA-Net outperforms the other compared methods to solve the problems of the loss of multiscale boundary detailed information and incomplete extraction.

TABLE IV
COMPLEXITY OF DIFFERENT COMPARATIVE METHODS ON THE WHU BUILDING DATASET

Method	Parameters (M)	FPS	IoU (%)
U-Net [16]	43.93	29.38	85.51
SegNet [14]	28.30	32.64	86.12
DeeplabV3+ [40]	54.71	21.70	89.61
BiSeNetV2 [49]	5.19	84.60	86.62
MAP-Net [22]	24.00	-	90.86
BOMSC-Net [33]	129.32	-	90.15
CGSANet [21]	43.03	11.85	91.55
LRAD-Net [50]	7.30	-	88.89
BuildFormer [31]	40.52	17.18	91.44
MDCGA-Net	46.70	28.76	91.73

3) Complexity Comparison: To evaluate the computation complexity of our method, we compare the parameter, frame per second (FPS), and IoU of related methods on the WHU Building dataset. As shown in Table IV, DeeplabV3+ has the highest performance but with the highest complexity among the three classic semantic segmentation networks. Compared with Deeplabv3+, the complexity of our method has reduced with

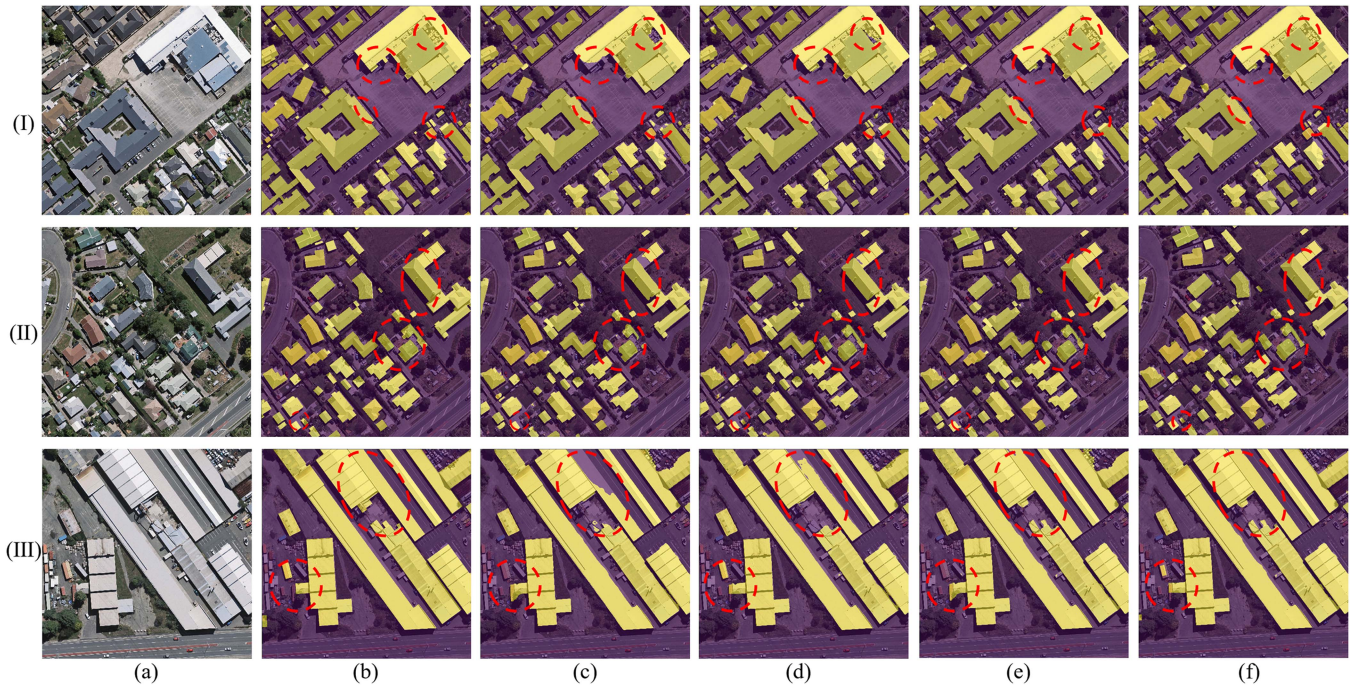

Fig. 8. Segmentation results on ablation experiments. The results are shown in (a) original images, (b) ground truth, (c) baseline, (d) baseline with DCAM, (e) baseline with GAGM, and (f) MDCGA-Net, respectively.

higher IoU since the MDCM reduces the scale of convolutional kernels and introduces the DCAM with low computational complexity. The BOMSC-Net has the most parameters among other SOTA methods due to the introduction of graph reasoning blocks. The IoU and FPS of our MDCGA-Net outperform other methods since our method lacks the reasoning time of the auxiliary modules. It is worth noting that although the outperform of the lightweight networks BiSeNetV2 and LRAD-Net are better than other algorithms in model size and FPS, IoU shows that the effectiveness of our proposed method establishes better speed-accuracy tradeoff.

E. Ablation Study

To further verify the influence of the different modules of our proposed MDCGA-Net, we also conduct an ablation study on the WHU building dataset. To be mentioned, compared to U-Net, we employ MSL, ASPP, DCAM, and GAGM within the MDCGA-Net. Note that the MDCM and the GAGM are designed in this article to overcome the existing problems. Our proposed MDCM aims to mitigate the limitations of the ASPP by introducing the DCAM for adaptive feature aggregation. In addition, the GAGM is crafted to discern correlations among global features. Therefore, we conduct different experiments to verify every proposed module and focus on the effectiveness of the two designed modules. First, we selected U-Net with the MSL and the ASPP for the baseline, to verify the effectiveness of the MSL and the ASPP by comparing it with U-Net under the ResNet-50 backbone. Second, we add the DCAM and the GAGM to the baseline separately to verify the validity of each module. Finally, the experiment to verify the effectiveness of our proposed method MDCGA-Net.

To prove the proposed modules of the MDCGA-Net that can solve the problem of the loss of boundary detailed information and the incompleteness of long-range extraction, we select some representative segmentation results, as shown in Fig. 8. The buildings in these selected images include complex shapes, shadow occlusions, and long-range coverage as shown in the first row to the third row (I–III) of Fig. 8. The first column (a) and the second column (b) of Fig. 8 are the original images and the corresponding ground truth, respectively. The third column to the sixth column of Fig. 8(c)–(f) shows the experimental results of the baseline, baseline with DCAM, baseline with GAGM, and our proposed MDCGA-Net, respectively.

The third column of Fig. 8 shows that the results of the baseline present boundary loss and incomplete detection. The comparison between columns (c) and (d) shows the effect of the proposed DCAM on boundary detection. As shown in the first row, the missing part of the boundary in the detection result of the baseline network is improved in Fig. 8(d). From the second row, we can see that our proposed model can improve the extraction effect of the buildings with shadow occlusions. In addition, as shown in the fifth column and the third row, the baseline with GAGM improves the extraction integrity of large-scale buildings because the GAGM can obtain the global relationships between the features of different levels. From Fig. 8, we can see that the results of our proposed method are the closest to the ground truth by comparing the segmentation results of the sixth column and the third to fifth columns, which demonstrates our MDCGA-Net can effectively improve the boundary extraction and extract completely in the building segmentation task.

Fig. 9. Segmentation results on ablation experiments on the Massachusetts building dataset and the Inria aerial image labeling dataset. The results are shown in (a) original images, (b) ground truth, (c) baseline, (d) baseline with DCAM, (e) baseline with GAGM, and (f) MDCGA-Net, respectively. (I) and (II) Massachusetts. (III) and (IV) Inria.

We also select some results of the Massachusetts building dataset and the Inria aerial image labeling dataset, as shown in Fig. 9. We select some buildings with complicated boundary shapes to demonstrate our proposed modules can solve the problems of detailed boundary loss and incompleteness of long-rang extraction. The samples of the Massachusetts building dataset are on the first row and the second row of Fig. 9. The results of the Inria aerial image labeling dataset are shown in the last two rows. From Fig. 9, we can see that the results of the third column present boundary loss and incomplete detection. The results of column (d) and column (f) show that the proposed DCAM can effectively solve the problem of boundary information loss. For example, the red circles in column (c) of the first row in Fig. 9 show that the results of the baseline have the misclassification of boundary information. In contrast, the fourth column and sixth column show that our method can effectively solve this problem. In addition, the third column of the fourth row shows that the baseline method has the omissions of the boundary information of the buildings. The column (f) shows that our method can also solve the problem of boundary information loss. From the second and third rows, we can see that the baseline with GAGM improves the extraction integrity of the complicated building. For example, the red circles above column (c) of the second and the third rows show the incomplete segmentation result of the baseline. The fifth and the sixth columns show our method can effectively alleviate this problem. The ablation study on these three datasets demonstrates our MDCGA-Net alleviates the problems of boundary information loss and incomplete segmentation in building extraction from RSIs.

TABLE V
ABLATION EXPERIMENTAL RESULTS ON THE WHU BUILDING DATASET

Method	IoU (%)	P (%)	R (%)	F1 (%)
U-Net (ResNet-50)	89.81	94.53	94.71	94.62
U-Net + MSL	90.33	94.84	95.01	94.92
U-Net + ASPP	90.02	94.62	94.84	94.73
Baseline (U-Net + MSL +ASPP)	90.54	95.01	95.21	95.11
Baseline + MDCM	90.98	95.16	95.43	95.29
Baseline + GAGM	90.91	95.28	95.37	95.32
MDCGA-Net (Ours)	91.73	95.61	95.79	95.70

We present a quantitative evaluation as shown in Table V. We can see that the IoU and the F1-score obtained with each module used in MDCGA-Net demonstrate the effectiveness of our method in the WHU building dataset. The integration of MSL into U-Net brought about a 0.52% uplift in IoU and a 0.30% increase in F1-score. ASPP augmented the IoU by 0.21% and F1-score by 0.11%. Table VII shows that the baseline method obtains a 90.54% in the IoU and 95.11% in the F1-score, which improves the performance of U-Net under the ResNet-50 backbone by 0.73% and 0.49%, respectively. This is because the

TABLE VI
ABLATION EXPERIMENTAL RESULTS ON THE THREE DATASETS

Method	Massachusetts		WHU Building		Inria	
	IoU (%)	F1 (%)	IoU (%)	F1 (%)	IoU (%)	F1 (%)
U-Net (ResNet-50)	70.68	81.94	89.81	94.62	76.28	86.46
U-Net + MSL	71.92	82.32	90.33	94.92	76.93	86.97
U-Net + ASPP	71.77	82.28	90.02	94.73	77.13	87.01
Baseline (U-Net + MSL + ASPP)	73.01	82.67	90.54	95.01	77.78	87.52
Baseline + MDCM	74.76	84.67	90.98	95.29	79.50	88.47
Baseline + GAGM	74.55	84.72	90.91	95.32	79.16	88.62
MDCGA-Net (Ours)	76.31	86.72	91.73	95.70	81.73	89.95

TABLE VII
ABLATION EXPERIMENTAL RESULTS WITH DIFFERENT NUMBERS OF THE DOWNSAMPLING LAYER

Layer (N)	IoU (%)	P (%)	R (%)	F1 (%)
3	87.61	92.37	93.04	92.70
4	90.42	95.01	95.23	95.12
5	91.73	95.61	95.79	95.72
6	91.71	95.57	95.71	95.64

feature representation ability of building extraction at various scales can be improved by the MSL and the ASPP. By adding the DCAM to the baseline, the method performance achieves a 0.44% improvement in the IoU and 0.18% improvement in the F1-score, which demonstrates that the MDCM effectively alleviates the positional information loss and integrates direction information of building boundaries. However, the P-value of the baseline with the DCAM method is not as high as the method of baseline with GAGM. This indicates that the baseline with the MDCM only improves the ability to adaptively capture the multiscale contextual information but ignores the global relationship between features of different levels, which leads to failure to guarantee the building extraction completely. Moreover, the proposed GAGM improves the IoU by 0.37% and the F1-score by 0.21% from the ablation results. This indicates that GAGM can enhance the ability to capture the global information of the discriminative features with long-range dependencies. As a result, our method achieves the best results of 91.73% IoU and 95.72% F1-score on the WHU building dataset owing to the proposed modules.

In addition, we conduct the ablation experiments on the Massachusetts building dataset and the Inria aerial image labeling dataset, as shown in Table VI. We can see that the IoU and the F1-score have improved with the proposed modules on the Massachusetts building dataset, the WHU building dataset, and the Inria aerial image labeling dataset. Specifically, adding the MSL brought 1.24%, 0.52%, and 0.65% improvement in the IoU and 0.38%, 0.30%, and 0.51% improvement in the F1-score on the three datasets, respectively. The ASPP improves the IoU by 1.09%, 0.21%, and 0.85% compared to the U-Net. The baseline methods improve the performance of U-Net under the ResNet-50 backbone by 2.33%, 0.73%, and 1.50% in the three datasets, respectively. Adding the DCAM brought 1.75%, 0.44%, and 1.72% improvement in the IoU and 2.00%, 0.28%, and 0.95% improvement in the F1-score on the three datasets, respectively. This indicates that the DCAM can improve the ability to extract the multiscale boundary information of the buildings. Moreover, the GAGM improves the IoU by 1.54%, 0.37%, and 1.38% compared to the baseline method. In particular, the F1-scores of the baseline with the GAGM are higher than the method of baseline, which increases the F1 by 2.05%, 0.31%, and 1.10%, respectively. This demonstrates that GAGM can enhance the ability to capture the global relationship between features of different levels.

To further verify the effectiveness of our proposed modules in building extraction, we present the visualization results. We extract the visualization results of the output feature map $f_i (i \in [1, 2, 3, 4, 5])$ of each layer in the encoder and the feature map $\hat{f}_5$ of the fifth layer without the MDCM, as shown in Fig. 10. From the results of f_1 to f_4, we can see that the location information in f_1 is more prominent, but the semantic and context information may be insufficient. This shows that the features learned by the model are more sufficient with the network layer deeper. In particular, from $\hat{f}_5$ and f_5, we can see that f_5 pays more attention to building boundary details, which indicates that the proposed MDCM effectively improves the ability to explore the boundary detailed information. To verify the influence of the GAGM, we extract the visualization results of the fused feature maps u_1 and the feature map a_1 of the decoder. In addition, we remove the GAGM to obtain the final feature map $\hat{a}_1$. Those visualization results are shown in Fig. 11. It can be seen that the feature map a_1 captures more discriminative features than u_1. Moreover, the comparison results between a_1 and $\hat{a}_1$ show that the GAGM can effectively obtain the global relationship of different features, which improves the integrity of building extraction.

Moreover, to evaluate the performance with different numbers of the downsampling layer (N), we also conduct the ablation study on the depth of the U-Net-like structure in MDCGA-Net, and the results as shown in Table VII. We set $N = 3$, 4, 5, and 6, where $N = 5$ is our method. Table VII shows that the IoU and the F1 are increasing with the N increases, and the IoU and the F1 of our method are highest. However, when $N = 6$, the IoU and F1 both have decreased since the boundary detailed information is lost more during too many downsampling layers. The experimental results show that we selected the proper number of downsampling layers to achieve the best building segmentation effect.

F. Limitations

Despite our method having achieved excellent results in the task of building extraction from RSIs, our method still has some limitations worthy of further improvement. The first disadvantage is that the experiments lack the verification of more real RSI data. Since our method only achieves superior performance on public datasets, it limits the generalization of our algorithm. Another disadvantage of our method is restricted to mobile hardware. Since our method adopts the deep network, the size of the network is still higher than the lightweight network in the area of mobile. Therefore, in the future, we will focus on the generalization and lightweight of the model.

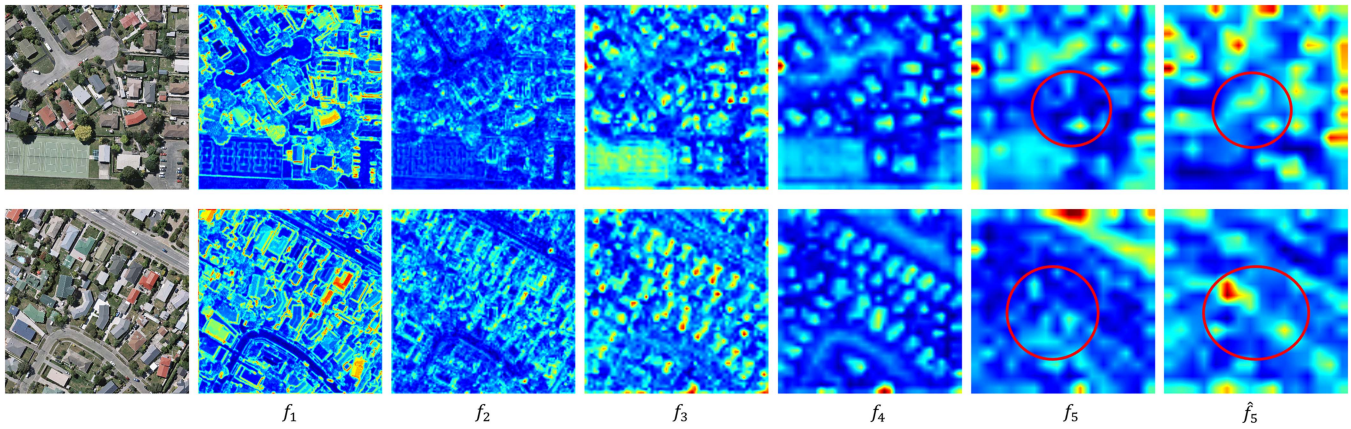

Fig. 10. Visualization results of feature map $f_i (i \in [1, 2, 3, 4, 5])$ and feature map $\hat{f}_5$ for the proposed MDCM.

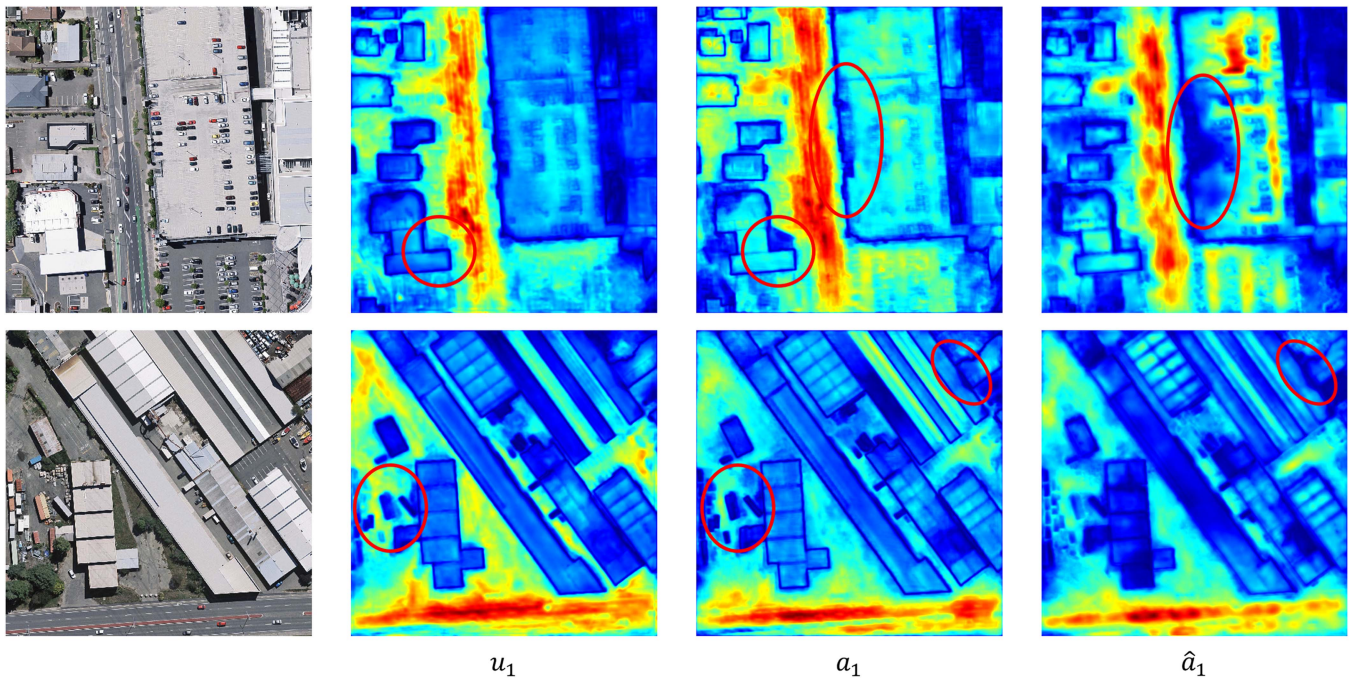

Fig. 11. Visualization results of feature map u_1, a_1, and $\hat{a}_1$ for the proposed GAGM.

V. Conclusion

In this article, we propose MDCGA-Net for building extraction from RSIs requiring boundary detailed information and complete extraction. We improve the segmentation results by introducing the direction information and global attention flow. We propose a multiscale aware encoding and GAG decoding to adaptively obtain multiscale contextual feature information and capture the relationship of global features, respectively. The quantitative and qualitative experimental analyses on the three benchmark datasets demonstrate that our model obtains significant performance and outperforms the classic semantic segmentation algorithms and the recent SOTA building extraction approaches. Moreover, the visualization results of the ablation study demonstrate the effectiveness of the MDCM and GAGM modules of the proposed MDCGA-Net. In future work, we will extend our proposed MDCGA-Net method to more classification tasks (e.g., road extraction) other than the building extraction task, to achieve automatic interpretation of RSIs.

References

[1] W. Li, C. He, J. Fang, and H. Fu, "Semantic segmentation based building extraction method using multi-source GIS map datasets and satellite imagery," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 238–241.

[2] C. Zhang et al., "Joint deep learning for land cover and land use classification," *Remote Sens. Environ.*, vol. 221, pp. 173–187, 2019.

[3] M. Li, A. Stein, and W. Bijker, "Urban land use extraction from very high resolution remote sensing images by Bayesian network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 3334–3337.

[4] I. Demir et al., "DeepGlobe 2018: A challenge to parse the earth through satellite images," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis.Pattern Recognit. Workshops*, 2018, pp. 172–181.

[5] M. Chen et al., "DR-Net: An improved network for building extraction from high resolution remote sensing image," *Remote. Sens.*, vol. 13, no. 2, 2021, Art. no. 294.

[6] W. Boonpook, Y. Tan, and B. Xu, "Deep learning-based multi-feature semantic segmentation in building extraction from images of UAV photogrammetry," *Int. J. Remote Sens.*, vol. 42, no. 1, pp. 1–19, 2021.

[7] M. Cote and P. Saeedi, "Automatic rooftop extraction in nadir aerial imagery of suburban regions using corners and variational level set evolution," *IEEE Trans. Geosci. Remote. Sens.*, vol. 51, no. 1, pp. 313–328, Jan. 2013.

[8] J. Wang, X. Yang, X. Qin, X. Ye, and Q. Qin, "An efficient approach for automatic rectangular building extraction from very high resolution optical satellite imagery," *IEEE Geosci. Remote. Sens. Lett.*, vol. 12, no. 3, pp. 487–491, Mar. 2015.

[9] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[10] Z. Guo and S. Du, "Mining parameter information for building extraction and change detection with very high-resolution imagery and GIS data," *GIScience Remote Sens.*, vol. 54, no. 1, pp. 38–63, 2017.

[11] M. Awrangjeb, C. Zhang, and C. S. Fraser, "Automatic extraction of building roofs using Lidar data and multispectral imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 83, pp. 1–18, 2013.

[12] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral-spatial hyperspectral image classification: An overview and new guidelines," *IEEE Trans. Geosci. Remote. Sens.*, vol. 56, no. 3, pp. 1579–1597, Mar. 2018.

[13] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[14] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[15] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

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

[17] W. Yu, B. Liu, H. Liu, and G. Gou, "Recurrent residual deformable conv unit and multi-head with channel self-attention based on U-Net for building extraction from remote sensing images," *Remote. Sens.*, vol. 15, no. 20, 2023, Art. no. 5048.

[18] S. D. Khan, L. Alarabi, and S. Basalamah, "An encoder-decoder deep learning framework for building footprints extraction from aerial imagery," *Arabian J. Sci. Eng.*, vol. 48, no. 2, pp. 1273–1284, 2023.

[19] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representation*, 2021, pp. 3–7.

[20] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE. Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[21] S. Chen, W. Shi, M. Zhou, M. Zhang, and Z. Xuan, "CGSANet: A contour-guided and local structure-aware encoder-decoder network for accurate building extraction from very high-resolution remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote. Sens.*, vol. 15, pp. 1526–1542, 2022.

[22] 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.

[23] M. Gong et al., "Context-content collaborative network for building extraction from high-resolution imagery," *Knowl. Based Syst.*, vol. 263, 2023, Art. no. 110283.

[24] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

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

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

[27] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[28] S. Gao, M. Cheng, K. Zhao, X. Zhang, M. Yang, and P. H. S. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[29] Y. Shen et al., "BDANet: Multiscale convolutional neural network with cross-directional attention for building damage assessment from satellite images," *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, 2022, Art. no. 5402114.

[30] X. Chen, C. Qiu, W. Guo, A. Yu, X. Tong, and M. Schmitt, "Multiscale feature learning by transformer for building extraction from satellite images," *IEEE Geosci. Remote. Sens. Lett.*, vol. 19, 2022, Art. no. 2503605.

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

[32] Y. Wang, Q. Zhao, Y. Wu, W. Tian, and G. Zhang, "SCA-Net: Multiscale contextual information network for building extraction based on high-resolution remote sensing images," *Remote. Sens.*, vol. 15, no. 18, 2023, Art. no. 4466.

[33] Y. Zhou et al., "BOMSC-Net: Boundary optimization and multi-scale context awareness based building extraction from high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, 2022, Art. no. 5618617.

[34] S. Chan, Y. Wang, Y. Lei, X. Cheng, Z. Chen, and W. Wu, "Asymmetric cascade fusion network for building extraction," *IEEE Trans. Geosci. Remote. Sens.*, vol. 61, 2023, Art. no. 2004218.

[35] H. Jung, H. Choi, and M. Kang, "Boundary enhancement semantic segmentation for building extraction from remote sensed image," *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, 2022, Art. no. 5215512.

[36] H. Yang, P. Wu, X. Yao, Y. Wu, B. Wang, and Y. Xu, "Building extraction in very high resolution imagery by dense-attention networks," *Remote. Sens.*, vol. 10, no. 11, 2018, Art. no. 1768.

[37] X. Pan et al., "Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms," *Remote. Sens.*, vol. 11, no. 8, 2019, Art. no. 917.

[38] H. Guo, Q. Shi, B. Du, L. Zhang, D. Wang, and H. Ding, "Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images," *IEEE Trans. Geosci. Remote. Sens.*, vol. 59, no. 5, pp. 4287–4306, May 2021.

[39] 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.

[40] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[41] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13713–13722.

[42] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11531–11539.

[43] Q. Zhang et al., "Dense attention fluid network for salient object detection in optical remote sensing images," *IEEE Trans. Image Process.*, vol. 30, pp. 1305–1317, 2021.

[44] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[45] F. Milletari, N. Navab, and S. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. Int. Conf. 3D Vis.*, 2016, pp. 565–571.

[46] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, Dept. Comput. Sci. Univ. Toronto, Toronto, ON, Canada, 2013.

[47] 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.

[48] 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.

[49] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3051–3068, 2021.

[50] J. Liu, H. Huang, H. Sun, Z. Wu, and R. Luo, "LRAD-Net: An improved lightweight network for building extraction from remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote. Sens.*, vol. 16, pp. 675–687, 2023.

Penghui Niu is currently working toward the Ph.D. degree in control science and engineering with the School of Artificial Intelligence, Hebei University of Technology, Tianjin, China.

His research interests include deep learning and intelligent interpretation of remote sensing images.

Junhua Gu received the B.S. degree in mathematics from Shanghai Jiaotong University, Shanghai, China, in 1988, and the M.S. degree in computer science and the Ph.D. degree in electrical engineering from the Hebei University of Technology, Tianjin, China, in 1993 and 1997, respectively.

He is currently a Professor with the School of Artificial Intelligence, Hebei University of Technology. He has authored or coauthored more than 70 papers. His current research interests include big data, intelligent control, and intelligent transportation systems.

Dr. Gu is the Hebei new century "333 Talent Project" second-level suitable person.

Yajuan Zhang received the B.S. degree in computer science and technology from Inner Mongolia Agricultural University, Inner Mongolia, China, in 2007, and the M.S. degree in control science and engineering from the Hebei University of Technology, Tianjin, China, in 2010.

She is currently a Senior Experimentalist with the School of Artificial Intelligence, Hebei University of Technology. Her current research interests include image processing and data mining.

Ping Zhang received the M.S. degree in software engineering from the College of Software, Jilin University, Changchun, China, in 2018, and the Ph.D. degree in computer science from Jilin University, Changchun, in 2021.

She is currently a Lecturer with the School of Artificial Intelligence, Hebei University of Technology, Tianjin, China. Her research interests include feature selection and machine learning.

Taotao Cai received the Ph.D. degree in computer science from Deakin University, Melbourne, VIC, Australia, in 2020.

He is currently a Postdoctoral Research Fellow with the School of Computing, Macquarie University, Sydney, NSW, Australia. Before moving to Macquarie University, he was an Associate Research Fellow with the School of Info Technology, Deakin University. His research interests include graph data processing, social network analytic and computing, and data mining.

Wenjia Xu received the M.S. degree in vegetation ecology and remote sensing from the Institute of Geographic Sciences and Resources, Chinese Academy of Sciences, Beijing, China, in 2008.

She is currently a Senior Engineer with the Hebei Institute of Hydrological Engineering Geological Exploration (Hebei Remote Sensing Center), Shijiazhuang, China. Her research interests include remote sensing big data.

Ms. Xu is the Hebei new century "333 Talent Project" second-level suitable person. In 2019, she was rated as an expert in youth posts in Hebei Province. The project she presided over has won two-second prizes and two third prizes of provincial and ministerial science and technology progress awards.

Jungong Han (Senior Member, IEEE) received the Ph.D. degree in telecommunication and information systems from Xidian University, Xi'an, China, in 2004.

He is currently the Chair Professor of Computer Vision with the Department of Computer Science, The University of Sheffield, Sheffield, U.K. He also holds an Honorary Professorship with The University of Warwick, Coventry, U.K. His research interests include computer vision, artificial intelligence, and machine learning.

Dr. Han is a Fellow of the International Association of Pattern Recognition. He is an Associate Editor for several prestigious journals, such as IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and *Pattern Recognition*.