A Crossmodal Multiscale Fusion Network for Semantic Segmentation of Remote Sensing Data

Xianping Ma, Xiaokang Zhang, *Member, IEEE*, and Man-On Pun, *Senior Member, IEEE*

***Abstract*—Driven by the rapid development of Earth observation sensors, semantic segmentation using multimodal fusion of remote sensing data has drawn substantial research attention in recent years. However, existing multimodal fusion methods based on convolutional neural networks cannot capture long-range dependencies across multiscale feature maps of remote sensing data in different modalities. To circumvent this problem, this work proposes a crossmodal multiscale fusion network (CMFNet) by exploiting the transformer architecture. In contrast to the conventional early, late, or hybrid fusion networks, the proposed CMFNet fuses information of different modalities at multiple scales using the cross-attention mechanism. More specifically, the CMFNet utilizes a novel crossmodal attention architecture to fuse multiscale convolutional feature maps of optical remote sensing images and digital surface model data through a crossmodal multiscale transformer (CMTrans) and a multiscale context augmented transformer (MCATrans). The CMTrans can effectively model long-range dependencies across multiscale feature maps derived from multimodal data, while the MCATrans can learn discriminative integrated representations for semantic segmentation. Extensive experiments on two large-scale fine-resolution remote sensing datasets, namely ISPRS Vaihingen and Potsdam, confirm the excellent performance of the proposed CMFNet as compared to other multimodal fusion methods.**

***Index Terms*—Combined squeeze-and-excitation (CSE), cross attention, crossmodal multiscale fusion, transformer.**

I. Introduction

SEMANTIC segmentation of remote sensing data is one of the most important tasks in geoscience research. The goal is to classify surface objects based on remote sensing data. Driven by the rapidly growing remote sensing devices and platforms, the amount of remote sensing data has grown exponentially over the past few decades, which provides the field with a wealth of multisource and multimodal data [1], [2] such as hyperspectral imagery (HSI), multispectral imagery (MSI), visible images (VIS), and light detection and ranging (LiDAR). Since each data source characterizes one modality of the surface objects, better semantic segmentation is expected by fusing different modalities derived from multiple data sources. However, multimodal data generated by different sensors exhibits vastly distinct characteristics such as heterogeneous statistical properties and noise levels across modalities [3], [4]. Thus, it remains very challenging to extract the salient features from different data sources before combining these features to obtain a better segmentation performance.

Conventional methods, such as component substitution, geostatistical analysis, and sparse representation, are hindered by their *ad hoc* complex feature extraction methods and fusion rules [5], which incurs missing details and degraded accuracy [6]. Recently, machine learning techniques have been successfully applied to a wide range of remote sensing applications, such as random forest [7], support vector machine [8], [9], and convolutional neural networks (CNNs) [10], [11]. In particular, many CNN-based methods have been proposed for semantic segmentation to fuse multisource remote sensing data [12]–[14]. For instance, the authors in [15] and [16] reported CNN-based fusion methods for infrared images and VIS, whereas EndNet was proposed in [12] to fuse HSI and LiDAR data. In particular, the fully convolutional networks (FCNs) reported in [10] is the first end-to-end CNN structure with proven effectiveness. In the FCN, an encoder extracts features by gradually decreasing the resolution of the feature map, while a decoder learns from the high level or more abstract features, and subsequently, improves the segmentation results by progressively expanding the receptive domain. By exploiting its translation equivalence and locality, the FCN demonstrated impressive performance. However, the FCN suffers from blurred edges and imprecise segmentation due to its simple upsampling operation in resolution restoration and disregard of the spatial relationship between pixels. To circumvent these problems, U-Net [11] proposed a symmetric expanding path for more accurate segmentation. However, these methods overlooked the potential benefits offered by multimodal data sources. To apply multimodal data in semantic segmentation, FuseNet was proposed to extract and fuse features from different modalities in the encoder stage [17]. By further enhancing early and late fusion in FuseNet, vFuseNet showed that early fusion can improve the learning of stronger multimodal features at the cost of poor noise susceptibility. Furthermore, late fusion based on the residual correction strategy is demonstrated to improve semantic labeling, which facilitates the recovery of critical errors on challenging pixels [18].

Manuscript received February 4, 2022; revised March 19, 2022; accepted April 2, 2022. Date of publication April 5, 2022; date of current version May 11, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 41801323, in part by China Postdoctoral Science Foundation under Grant 2020M682038, and in part by the Shenzhen Science and Technology Innovation Committee under Grant JCYJ20190813170803617. *(Corresponding authors: Man-On Pun; Xiaokang Zhang.)*

Xianping Ma and Man-On Pun are with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: xianpingma@link.cuhk.edu.cn; simonpun@cuhk.edu.cn).

Xiaokang Zhang is with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen 518172, China, and also with the School of Mathematical Sciences, University of Science and Technology of China, Hefei 230026, China (e-mail: zhangxiaokang@cuhk.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3165005

Despite their many advantages, these CNN-based fusion methods are handicapped by the problem of field of perception. It is worth noting that these feature maps generated by different convolutional layers possess various levels of resolution and characteristics. More specifically, the feature map generated by the shallow layer has a higher resolution but smaller perception field. As a result, such a feature map mainly contains detailed local features. In contrast, increasing the depth of the CNN can provide feature maps with higher level abstraction while losing detailed information. Thus, it is challenging for CNN-based methods to compensate semantic gaps among feature maps through crossmodal feature fusion using convolution operations. Furthermore, since CNNs neglect long-range dependencies across multiscale feature maps of remote sensing data in different modalities, they cannot fully exploit the inherent dependence relationships among multimodal data and learn discriminative integrated representations for semantic segmentation.

To cope with the aforementioned challenges, some pioneering attempts have been made by utilizing the transformer architecture [19]. The transformer architecture was originally developed for natural language processing (NLP) before being successfully applied to the field of computer vision [20]. Empowered by its distinct multihead self-attention blocks, the transformer architecture is capable of capturing long-range dependencies between each pair of elements in the feature map. In contrast to the convolutional layer whose perceptual field is limited by its kernel size, the perceptual field of the self-attention blocks is naturally global in theory. This remarkable characteristic enables the transformer architecture to better extract and fuse features of different abstraction levels as compared to the convolutional CNN-based networks.

Inspired by the aforementioned discussions, this work proposes a crossmodal multiscale fusion network (CMFNet) by exploiting both CNN and the transformer architecture. More specifically, the transformer architecture is utilized for multimodal fusion while its fusion output is further processed by a decoder as residual connections. Our experimental results confirm that such a joint design of fusion and residual connection can substantially enhance the segmentation performance by exploiting crossmodal multiscale data sources. The main contributions of this work are summarized as follows.

1) A novel transformer-based crossmodal network with multiscale skip fusion is first devised to reduce the decoding ambiguity by fusing complementary multimodal information. It is demonstrated that the crossmodal multiscale skip fusion can implement the residual function while achieving multimodal fusion simultaneously.
2) After establishing a cross-modal attention (CMA) layer, we propose a crossmodal multiscale transformer (CMTrans) for semantic segmentation. The proposed CMTrans can effectively model long-range dependencies across multiscale feature maps of remote sensing data in different modalities. The resulting fused features are further enhanced by multiscale context augmented transformer (MCATrans) based on a cross-scale attention mechanism to learn more discriminative and distinguishable integrated representations for semantic segmentation.
3) By combining CMTtrans and MCATrans, we propose the CMFNet to perform the crossmodal and multiscale fusion network for semantic segmentation. To our best knowledge, the proposed CMFNet is the first transformer-based architecture for the crossmodal multiscale fusion in the field of remote sensing.

In the sequel, the relevant work of semantic segmentation in remote sensing is first reviewed in Section II before Section III elaborates the proposed frame and modeling method in detail. After that, Section IV describes our experiment setup and analyses on the experimental results. Finally, the Section V concludes this article.

II. Related Work

Generally speaking, semantic segmentation is one of the most fundamental tasks in understanding an image. Driven by the rapid development of digital signal processing technology, intensive research on semantic segmentation has been carried out on digital photography, medical images, and remote sensing images [21].

A. Modality Fusion

In general, modality is defined as the form in which information is represented. For instance, image, text, voice, and video are the most common modalities under intensive investigation. Furthermore, multimodality refers to the combination of two or more modalities. To fuse information of different modalities, many multimodal fusion techniques have been developed in the literature, such as feature extraction, feature alignment, and feature fusion for various applications such as visual question answering [22]–[24], sentiment analysis [25], medical image processing [26], [27], and remote sensing [21], [28]. More recently, it was shown in MulT [29] that the cross-attention mechanism can provide a latent crossmodal adaptation that fuses crossmodal information by directly attending to low-level features in other modalities.

B. Modality Fusion in Remote Sensing Segmentation

In remote sensing, images of different resolutions and channels are called multimodality. In particular, equipped with sophisticated remote sensing acquisition technology, the remote sensing community have full access to a wide range of multimodality data such as HSI, MSI, VIS, and LiDAR. Driven by the recent advances in machine learning, many machine learning approaches have been developed to fuse multimodal data in the literature. In particular, the encoder–decoder architecture has been shown very effective in fusing multimodal data. According to where the fusion takes place, these encoder–decoder methods can be classified into three approaches, namely data-level fusion, early fusion, and late fusion. For instance, ResUNet-a fuses the RGB and depth images by directly concatenating the images in the data level before feeding the concatenated data into encoders as shown in Fig. 1(a) [13]. In contrast, FuseNet utilizes a dual-branch encoder backbone to encode the RGB-Infrared (IR) and the digital surface model (DSM) data individually

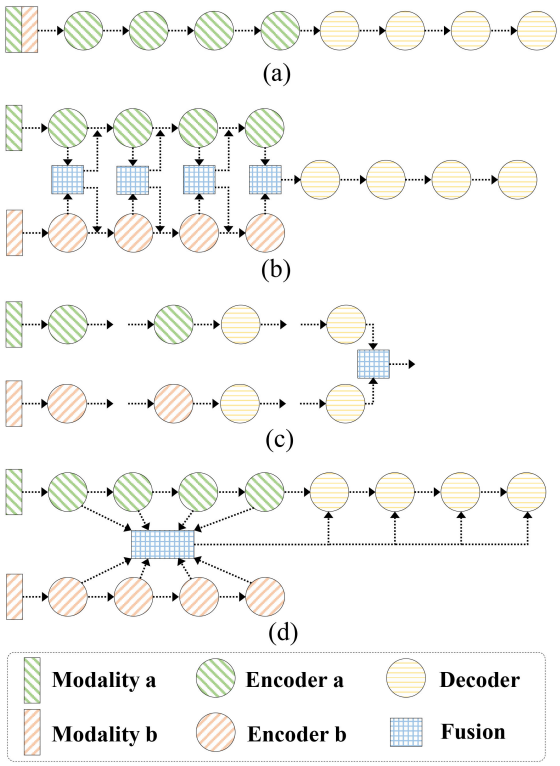


Fig. 1. Comparison of four fusion approaches. (a) Data-level fusion: To stack the data directly before encoding. (b) Early fusion: Fusion taken place in different encoder layers separately in different scales. (c) Late fusion: Fusion performed after decoding, also called the decision fusion. (d) Proposed multiscale skip fusion: To fuse multiscale data before feeding into the decoders as residual connection.

before fusing these encoded features as the input to a single decoder as shown in Fig. 1(b) [17]. This fusion approach is known as the early fusion. Finally, Fig. 1(c) shows a typical example for the late fusion, e.g., vFuseNet [18]. In the late fusion, each modality is individually encoded and decoded followed by fusion of the decoder outputs. In sharp contrast to the three existing approaches, the proposed CMFNet develops a completely different approach to fuse crossmodal features. As shown in Fig. 1(d), the proposed fusion approach first carries out crossmodal multiscale information fusion before combining the fused information with the decoder feature as a residual connection that is referred to as the multiscale skip fusion in the sequel.

C. Vision Transformer

CNN-based networks have become the dominant solution to semantic segmentation tasks. However, the major drawback of these networks is their limited local receptive fields. Recently, the transformer architecture empowered with the self-attention mechanism has been found effective in providing a global perceptual field in NLP [19]. The self-attention mechanism explores the relationship of input tokens by exploiting three weight matrices, namely query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$. As a result, the self-attention mechanism can capture long-range dependencies by recalibrating feature maps to characterize the beneficial information across channels and spatial dimensions. Inspired by the success of the transformer architecture in NLP, the seminal work [20] proposed ViT using a standard transformer encoder by converting the image into a sequence of patches. ViT is one of the first networks capable of utilizing global self-attention for image processing. After ViT, a number of transformer-based methods have been successfully developed for semantic segmentation. For instance, TransUNet [30] utilizes a transformer-based encoder and an U-Net decoder for medical image segmentation, whereas the transformer architecture was used in TransBTS for feature enhancement in [31]. In addition, DSCT [32] proposed a method extracting contextual information using the swin transformer [33] as the backbone in conjunction with a densely connected feature aggregation module designed to recover resolution and generate segmentation maps. Similar to DSCT, Swin-Unet [34] is also designed based on the swin transformer, targeting at replacing the entire network with transformer structures for medical image segmentation. However, these existing works mainly focused on applying the transformer architecture as a feature extractor or an augmented module on data of the same level or modality. Thus, it remains an open research question on how to use the transformer architecture to exploit the long-range dependence relationships across different levels and modalities.

In [35], TransFuser was proposed to fuse multilevel feature maps using multiple transformers in different levels for autonomous driving. However, such a structure of multiple transformers incurs prohibitively expensive computational complexity, which makes TransFuser impractical for applications such as semantic segmentation of high-resolution remote sensing images. To overcome the computational complexity problem, UCTransNet proposed an end-to-end network equipped with a multiscale channel-wise cross fusion transformer and recurrent neural networks [36]. It was shown that multiscale features are essential for resolving complex scale variations in medical image segmentation [36]. However, as UCTransNet only explores single-modal and low-resolution medical images, it is impractical for the multimodal high-resolution remote sensing images. Finally, most existing transformer-based segmentation methods focus on improving the encoding performance using a U-Net [11] scheme. For instance, [30], [35], [37] either simply embed the transformer into the encoder or use the transformer to fuse separate branches, which results in increased computational complexity due to the intensive computation required by the transformer.

III. Methodology

In this work, we consider cross-attention and cross-transformer multimodal data fusion for high-resolution remote sensing images. For presentational simplicity, we use dual-modal data, namely the RGB and the DSM images, to elaborate the proposed network calledCMFNet in the sequel. It should be emphasized that CMFNet can be extended to multimodal data of more than two data sources in a straightforward manner.

CMFNet first separately extracts RGB features and DSM features using two branches of VGG-16 [38]. Recalling that

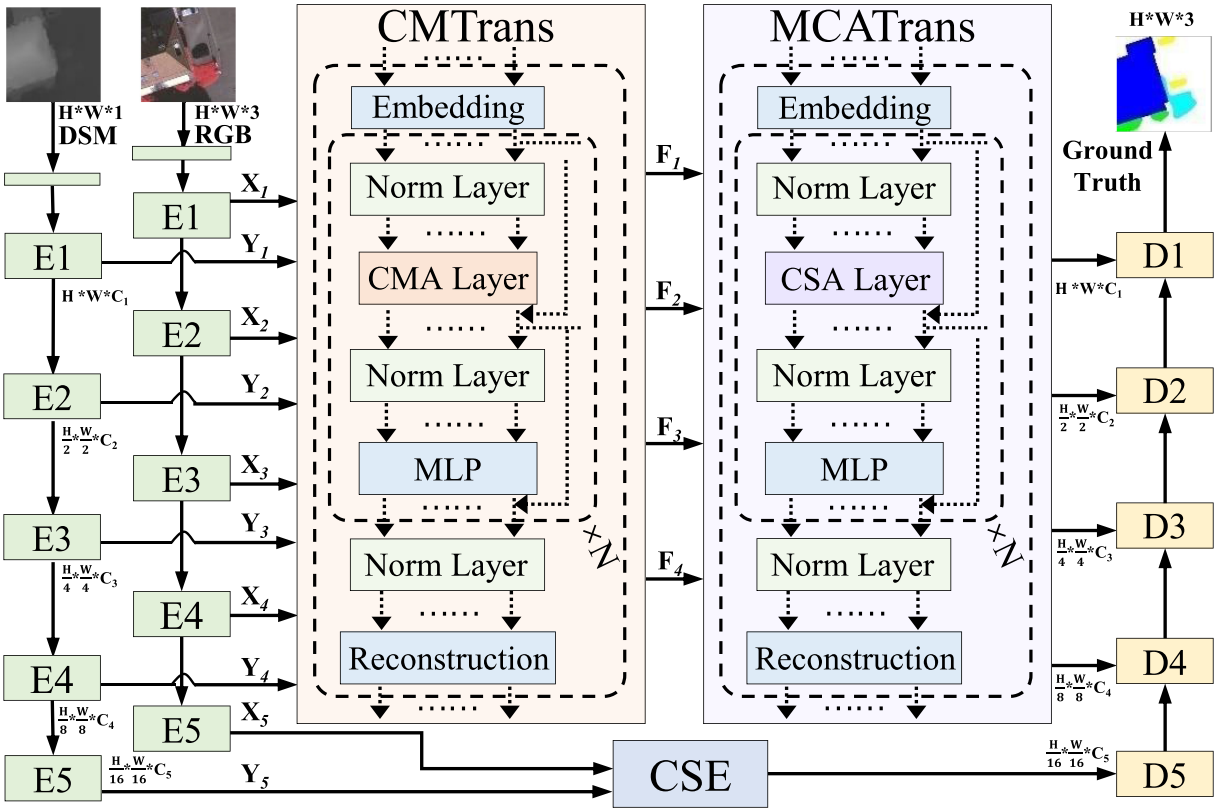


Fig. 2. Illustration of the proposed CMFNet. The proposed CMTrans fuses feature maps of different modalities by simultaneously exploiting multiscale crossmodal information. MCATrans uses the cross-scale attention mechanism to enhance the extracted features from the main modality, i.e., the RGB images. Finally, a CSE block is used to fuse the highest level features. The output of MCATrans is fed into the decoder of the corresponding scale before the pixel-level classification is performed.

VGG-16 has a five-layer encoder, we designate the first four layers to extract multiscale features from each branch before the crossmodal and multilevel features are fused by CMTrans. After that, MCATrans is proposed to refine the multiscale fusion of multimodal features and filter the representations propagated through our multiscale skip fusion strategy. Finally, the output of the fifth encoder, i.e., the features of the richest semantic information, is fused through a channel reweighting module called combined squeeze-and-excitation (CSE). This design allows CMFNet to achieve a good performance with reduced computational complexity. In the following, we will first provide an overview of the proposed CMFNet framework before elaborating on the details of each key component.

A. Network Architecture

Fig. 2 depicts the framework of the proposed network. As previously explained, we use RGB and DSM images as the two modalities of input data in the following discussions while stipulating the RGB images as our main modal and DSM data as the assisted modality. Given RGB images $X \in \mathbb{R}^{H\times W\times 3}$ and the corresponding DSM depth data $Y \in \mathbb{R}^{H\times W\times 1}$, a dual-branch encoder first extracts multilevel features from each modality. Motivated by the fact that semantic segmentation of remote sensing images requires dense pixel-wise classification, the proposed CMFNet adopts the encoder–decoder architecture reported in the literature. Furthermore, SegNet reported in [39] is chosen as the backbone of the proposed network as its output is of the same resolution as its input, which is a very convenient property to avoid performance degradation incurred by direct upsampling. In addition, the encoder in SegNet consists of classical convolutional layers from VGG-16 [38]. A dual-branch encoder architecture is proposed to process each input data source. More specifically, two VGG-16 branches are used to extract RGB features and DSM features, respectively. Assuming that the original images is of size $H \times W$, the down-sampled feature maps produced by the ith encoder layer are of size $\frac{H}{2^{i-1}} \times \frac{W}{2^{i-1}}$, where i is the layer index with $i = 1, 2, \ldots, 5$.

After the encoding process, the multilevel feature maps derived from the main modality and the assisted modality are fed into a novel network called CMTrans for crossmodal feature fusion. Details about the proposed CMTrans will be elaborated in the next section. The output of CMTrans contains a group of fused feature maps based on the main modal data, i.e., the DSM-assisted RGB image-based feature maps. These features are further enhanced by a cross-scale attention-based transformer called MCATrans whose output is fed into the decoder for upsampling and classification. More specifically, the proposed decoder layer consists of convolutional blocks, normalization (Norm) layers, RELU blocks, and upsampling block. Based on

the final output of encoder and multiscale skip fusion of the corresponding scales, the decoder first restores the complete spatial resolution while converting the encoded features into the final labels. In particular, the highest-level features produced by the fifth encoders from both branches are fused with CSE, and subsequently, fed into the first layer of decoder D5. The outputs from the MCATrans are added to the last four decoding layers for multiscale skip fusion as shown in Fig. 2. During the decoding process, the sparse feature maps are densified by the convolutional blocks, which results in higher resolution feature maps. This densification process continues until the resolution of the resulting feature maps reaches the resolution of the original input images.

B. CMTrans and MCATrans

CMTrans is a novel cross-attention-based transformer designed to fuse the multimodal features by modeling long-range dependencies across feature maps in different modalities. Specifically, cross-scale feature interactions are exploited to enhance the fusion performance. In contrast, MCATrans is designed as a cross-scale attention-based transformer to capture global contextual information and learn discriminative representations.

For the input image of size 256×256, we can obtain multiscale feature maps from two modalities with dimension $C_i \times \frac{256}{2^{i-1}} \times \frac{256}{2^{i-1}}$, where $C_i = 64 * i$ is the channel dimension at the ith scale with $i = 1, 2, 3, 4$. Tokenization is first performed by flattening the extracted features into sequences of 2-D patches using a convolutional layer of kernel and stride sizes equal to $m_i \times m_i$ with $m_i = \frac{32}{2^i}$. As a result, the 2-D feature patches of all levels are downsampled to $16 \times 16 \times C_i$ through multiscale feature embedding. After tokenization, eight tokens of two modilities are attained, including four for main modal RGB data denoted by $\mathbf{R}_i \in \mathbb{R}^{P \times C_i}$ and four for assisted modal DSM data denoted by $\mathbf{D}_i \in \mathbb{R}^{P \times C_i}$, where $P = 256$ and $i = 1, 2, 3, 4$.

1) CMTrans: The eight tokens of the two modalities are first fed into the CMA layer followed by a multilayer perceptron (MLP). We propose to employ a residual structure to reuse the features and facilitate the network training. The proposed CMA layer receives ten inputs, including eight tokens $\mathbf{R}_i$ and $\mathbf{D}_i$ as queries and two concatenated token $\mathbf{R}_\sum$ and $\mathbf{D}_\sum$ as key and value, respectively. $\mathbf{R}_\sum$ and $\mathbf{D}_\sum$ of dimension $P \times C_\sum$, where $C_\sum = \sum C_i$ are constructed by concatenating the four $\mathbf{R}_i$ and $\mathbf{D}_i$ tokens, respectively. It is worth noting that the crossmodal fusion at each scale is improved by exploiting multiscale features of other modalities.

To take full advantage of the multihead attention, we split these tokens into four nonoverlapping copies of $\frac{C_i}{4}$ channels each. For instance, $\mathbf{R}_i$ is divided into four $\tilde{\mathbf{R}}_i^{(j)} \in \mathbb{R}^{P \times \frac{C_i}{4}}$ for the queries of the ith layer after splitting, where $j = 1, 2, 3, 4$. It should be noted that key and value are also divided into four copies, and each copy contains nonoverlapping information of the same number of channels at different scales and $\tilde{\mathbf{R}}_\sum^{(j)} \in \mathbb{R}^{P \times \frac{C_\sum}{4}}$. Since the operations for all four heads are identical, we will omit the superscript $^{(j)}$ in the following discussions for the sake of notational simplicity. Finally, denoting the multilevel query, concatenated key and value for each head as $\mathbf{Q}_i, \tilde{\mathbf{Q}}_i \in \mathbb{R}^{P \times \frac{C_i}{4}}$ and $\mathbf{K}, \tilde{\mathbf{K}}, \mathbf{V}, \tilde{\mathbf{V}} \in \mathbb{R}^{P \times \frac{C_\sum}{4}}$, respectively, the update equations for these variables are given by

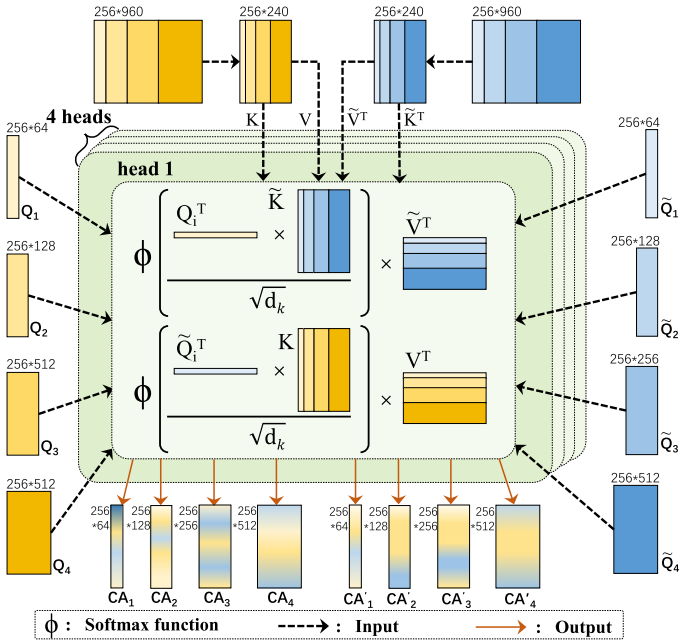


Fig. 3. Illustration of the proposed CMA layer.

$$\mathbf{Q}_i = \mathbf{R}_i W_{\mathbf{Q}_i}, \mathbf{K} = \mathbf{R}_\sum W_{\mathbf{K}}, \mathbf{V} = \mathbf{R}_\sum W_{\mathbf{V}} \tag{1}$$

$$\tilde{\mathbf{Q}}_i = \mathbf{D}_i W_{\tilde{\mathbf{Q}}_i}, \tilde{\mathbf{K}} = \mathbf{D}_\sum W_{\tilde{\mathbf{K}}}, \tilde{\mathbf{V}} = \mathbf{D}_\sum W_{\tilde{\mathbf{V}}} \tag{2}$$

where $W_{\mathbf{Q}_i}, W_{\tilde{\mathbf{Q}}_i}, W_{\mathbf{K}}, W_{\mathbf{V}}, W_{\tilde{\mathbf{K}}}$, and $W_{\tilde{\mathbf{V}}}$ are different weights. Note that different heads have different query weights, i.e., $W_{\mathbf{Q}_i}$ and $W_{\tilde{\mathbf{Q}}_i}$, but share the same key and value weights, i.e., $W_{\mathbf{K}}, W_{\mathbf{V}}, W_{\tilde{\mathbf{K}}}$, and $W_{\tilde{\mathbf{V}}}$.

The update process is illustrated in Fig. 3. With $\mathbf{Q}_i$, $\tilde{\mathbf{Q}}_i$, $\mathbf{K}$, $\mathbf{V}$, $\tilde{\mathbf{K}}$, and $\tilde{\mathbf{V}}$, the similarity matrix $\mathbf{S}_i$ is computed by the query associated with the RGB modality and the key associated with the DSM modality. Meanwhile, the similarity matrix $\mathbf{S}'_i$ is computed by the query associated with the DSM modality and the key associated with the RGB modality. Finally, the cross-attention value of main modality CA_i assisted by the DSM modality is weighted by $\mathbf{S}_i$, while the cross-attention value of assisted modality CA'_i is weighted by $\mathbf{S}'_i$. The computation of the crossmodal fusion by cross attention in the CMA layer can be expressed as

$$\begin{aligned} \mathrm{CA}_i = \mathbf{S}_i \tilde{\mathbf{V}}^\top &= \Phi\left(\frac{\mathbf{Q}_i^\top \tilde{\mathbf{K}}}{\sqrt{C_\sum}}\right) \tilde{\mathbf{V}}^\top \\ &= \Phi\left(\frac{W_{\mathbf{Q}_i}^\top \mathbf{R}_i^\top \mathbf{D}_\sum W_{\tilde{\mathbf{K}}}}{\sqrt{C_\sum}}\right) W_{\tilde{\mathbf{V}}}^\top \mathbf{D}_\sum^\top \end{aligned} \tag{3}$$

$$\mathrm{CA}'_i = \mathbf{S}'_i \mathbf{V}^\top = \Phi\left(\frac{\tilde{\mathbf{Q}}_i^\top \mathbf{K}}{\sqrt{C_\Sigma}}\right)\mathbf{V}^\top$$
$$= \Phi\left(\frac{W_{\tilde{\mathbf{Q}}_i}^\top \mathbf{D}_i^\top \mathbf{R}_\Sigma W_{\mathbf{K}}}{\sqrt{C_\Sigma}}\right) W_{\mathbf{V}}^\top \mathbf{R}_\Sigma^\top \quad (4)$$

where $\Phi(\cdot)$ and $(\cdot)^\top$ denote the softmax function and matrix transpose, respectively. Note that the similarity matrix $\mathbf{S}_i$ and $\mathbf{S}'_i$ in (3) and (4), respectively, represent the weighting matrix of all channels, while the output CA_i and $\mathrm{CA}'_i \in \mathbb{R}^{P\times C_i}$ are feature maps derived with channel weighting. It is worth noting that the attention for each point on the feature map is computed and taken into account at the global scale. After the computation of the cross attention, a simple MLP and residual operation are applied. In our experiments, cross-attention modules are stacked for N times. During this stage, CA_i and CA'_i computed in the previous round would be used for the next calculation. The resulting output can be expressed as follows:

$$\mathrm{CA}_i^n = \mathrm{CA}_i + \mathrm{MLP}\left(\mathbf{Q}_i + \mathrm{CA}_i^{n-1}\right) \quad (5)$$

$$\mathrm{CA}_i'^n = \mathrm{CA}'_i + \mathrm{MLP}\left(\tilde{\mathbf{Q}}_i + \mathrm{CA}_i'^{n-1}\right) \quad (6)$$

where the superscript $(\cdot)^n$ denotes the CMA layer index with $n = 1, 2, 3, 4$. The Norm layer shown in Fig. 2 is omitted in the aforementioned equations for notational simplicity. The procedures from (3) to (6) are repeated for four times to build a four-layer transformer. Finally, the four outputs of the main modal $\mathbf{CA}_i$ of size $P \times C_i$ are reconstructed by the reconstruction module to generate $\mathbf{F}_i$ whose size is $C_i \times \frac{256}{2^{i-1}} \times \frac{256}{2^{i-1}}$, for $i = 1, 2, 3, 4$. In short, the reconstruction is the reverse process of multiscale feature embedding. After performing the fusion at CMTrans, we obtain crossmodal features at four scales.

2) MCATrans: The output $\mathbf{F}_i$ of CMTrans is the feature fusion of four scales after the cross attention. These fusion features are then augmented to present more discriminative and distinguishable integrated information for semantic segmentation. To this end, we propose a cross-scale attention-based transformer called MCATrans to extract the global-context from the fusion feature space. MCATrans shares similar structures as CMTrans discussed previously, except that the CMA layer in CMTrans is replaced by the CSA layer as illustrated in Fig. 4. After that, the correlation information in CSA is calculated based on $\mathbf{Q}_i, \mathbf{K}$, and $\mathbf{V}$ from the main modality as follows:

$$\mathrm{A}_i = \mathbf{S}_i \mathbf{V}^\top = \Phi\left(\frac{\mathbf{Q}_i^\top \mathbf{K}}{\sqrt{\mathrm{C}_\Sigma}}\right)\mathbf{V}^\top \quad (7)$$

where $\mathbf{Q}_i, \mathbf{K}$, and $\mathbf{V}$ are computed from $\mathbf{F}_i$ using the same approach shown in (1) for $i = 1, 2, 3, 4$.

Note that the correlation of any two points on the feature map is incorporated in the computation in (7). Thus, this correlation information enables MCATrans to capture detailed local and global information through the gradient backpropagation.

After the attention computation, an MLP and residual operation are applied in a manner similar to (5), except that CA_i is replaced by A_i in (7). After that, the result is restored to the original size by the reconstruction module. Finally, the four outputs are combined with the corresponding decoder features derived from decoder layers D1 to D4 using element-wise summation, which results in the final full-resolution prediction.

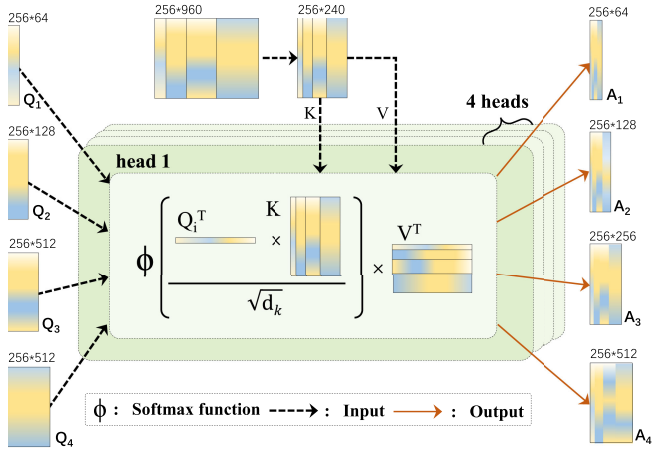


Fig. 4. Illustration of the proposed CSA layer.

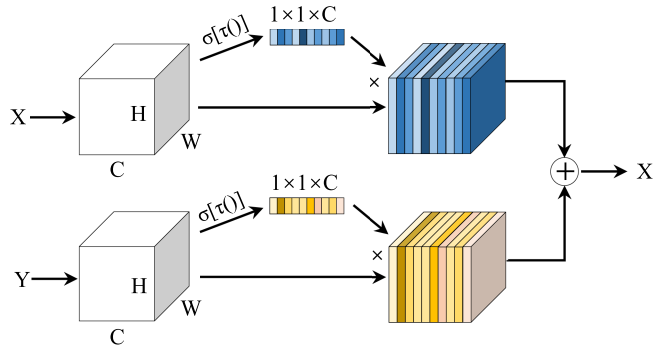


Fig. 5. Illustration of the CSE module.

C. Combined Squeeze-and-Excitation (CSE)

To fuse high-level features efficiently through enhancing the most informative channels, we propose a CSE structure consisting of two squeeze-and-excitation (SE) modules reported in [40] as shown in Fig. 5. In principle, SE can retain and enhance important feature information due to its global reception. More specifically, the first global average pooling is applied as a channel descriptor based on the channel attention mechanism followed by a 1×1 convolutional layer with the same number of channels as that of the input. After that, a Sigmoid activation function is utilized to normalize the weight vector to $[0, 1]$. Upon obtaining the weight vector, the outer products of the weight vector and the input feature maps in two branches are computed. Finally, the fused feature map X is derived by adding the results from the RGB and DSM branches. Since the CSE module is only applied to the fifth scale, its inputs include X_5 and Y_5. For notational simplicity, we will omit the subscript $_5$ and use X, Y, and C in the following discussions. The fused feature map X can be expressed as follows:

$$X = X \otimes \sigma\left(\tau(X)\right) + Y \otimes \sigma\left(\tau(Y)\right) \quad (8)$$

where $\tau(\cdot)$ denotes the operation of global pooling and 1×1 convolution, whereas $\otimes$ and $\sigma(\cdot)$ stand for the outer product operator and the Sigmoid function, respectively. Through this fusion mechanism, we can enhance the most informative features from the fifth layer of the largest number of channels. This enhancement is shown to help derive complementary information from the auxiliary modal DSM in our experiments.

D. Loss Function

The output of the last decoder layer is an image of the same resolution as the original image. For each pixel denoted by p, a prediction vector $[z_1^p, \ldots, z_k^p, \ldots, z_K^p]$ is generated where z_k^p is the probability that pixel p belongs to the kth class with $k = 1, 2, \ldots, K$ with K being the total number of classes. We denote by $\{y_k^p\}$ the ground-truth labels for pixel p with only one entry being one and others zero. We propose to train the classifier by minimizing the following loss function:

$$loss = \frac{1}{P} \sum_{p=1}^{P} \sum_{k=1}^{K} y_k^p \log \left(\frac{\exp(z_k^p)}{\sum_{l=1}^{K} \exp(z_l^p)} \right) \tag{9}$$

where P is the total number of pixels in the input image.

IV. Experiments and Discussion

A. Datasets

In this section, we use two well-known ISPRS 2-D semantic labeling datasets, namely Potsdam and Vaihingen, to validate the effectiveness of our proposed crossmodal multiscale fusion network. The Potsdam semantic labeling dataset contains 38 patches of 6000×6000 pixels, each consisting of a true orthophoto and a DSM with a ground sampling distance (GSD) of 5 cm. The Potsdam dataset provides near-infrared, red, green, and blue channels as well as DSM. In our experiments, we split the 24 patches with eroded boundary data provided by ISPRS into training and testing sets as follows.

1) Training set (18 patches): 6_10, 7_10, 2_12, 3_11, 2_10, 7_8, 5_10, 3_12, 5_12, 7_11, 7_9, 6_9, 7_7, 4_12, 6_8, 6_12, 6_7, 4_11.
2) Test set (6 patches): 2_11, 3_10, 4_10, 5_11, 6_11, 7_12.

Note that only the red, green, and blue channels are used in our experiments.

Furthermore, the Vaihingen dataset is composed of 33 images of an average size of 2494×2064 pixels and a GSD of 5 cm. However, only near-infrared, red, and green channels together with DSM are provided in the dataset. We split the 16 patches with eroded boundary data provided by ISPRS into training set with 12 patches and testing set with four patches. Specifically, ID: $1, 3, 23, 26, 7, 11, 13, 28, 17, 32, 34, 37$ are used for the training set, while ID: $5, 21, 15, 30$ the test set. Information from all three available channels is treated in a manner similar to RGB. For comparison purposes, the same training set and test set described previously are used on other existing methods.

B. Evaluation Metrics

The performance of our proposed CMFNet is evaluated in terms of the classification accuracy on the test dataset as well as standard statistical indices, including the overall accuracy (OA), the mean intersection over union (mIoU), and the F1 score (F1). These indices are defined as

$$\text{OA} = \frac{\sum_{k=1}^{P} \text{TP}_k}{\sum_{k=1}^{P} \text{TP}_k + \text{FP}_k + \text{TN}_k + \text{FN}_k} \tag{10}$$

$$\text{mIoU} = \frac{1}{P} \sum_{k=1}^{P} \frac{\text{TP}_k}{\text{TP}_k + \text{FP}_k + \text{FN}_k} \tag{11}$$

$$\text{F1} = 2 \times \frac{Q \times R}{Q + R} \tag{12}$$

where $\text{TP}_k, \text{FP}_k, \text{TN}_k$, and FN_k indicate the true positive, false positive, true negative, and false negative, respectively, for an object actually belonging to the kth class. Furthermore, Q and R are given by

$$Q = \frac{1}{P} \sum_{k=1}^{P} \frac{\text{TP}_k}{\text{TP}_k + \text{FP}_k} \tag{13}$$

$$R = \frac{1}{P} \sum_{k=1}^{P} \frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k}. \tag{14}$$

C. Experimental Setting

Our experiment platform was implemented with PyTorch on a single NVIDIA Tesla V100 GPU with 16-GB RAM. Due to the large size of the original data images, we used the sliding window to dynamically collect the training dataset. The span of the sliding window also defines the size of the overlapping area between two successive patches. During training, a suitable stride can extract more training samples and increase the performance. In contrast, a smaller step size in the testing stage enables us to average the classification results over the overlapping areas, which helps reduce boundary effects and improve the overall classification accuracy. Based on the aforementioned consideration, we used a 256-pixel stride for training and a 32-pixel stride for testing.

All models were trained using the stochastic gradient descent (SGD) algorithm with a learning rate of 0.01, momentum of 0.9, weight attenuation of 0.0005, and batch size of 10. The weights of the encoder were initialized with those from VGG-16 trained on ImageNet, while the weights of the decoder were randomly initialized as suggested in [41].

D. Performance Comparison

We benchmark the performance of our proposed method against that of seven representative deep learning methods, namely PSPNet, MAResU-Net, vFuseNet, FuseNet, ESANet, TransUNet, and SA-GATE. Designed for complex scene understanding, PSPNet [42] is an efficient multiscale pyramid scene parsing network by exploiting its characteristic global pyramid pooling features. Furthermore, MAResU-Net proposed in [43] performs semantic segmentation with a reduced computational

TABLE I
EXPERIMENTAL RESULTS ON THE VAIHINGEN DATASET

Method	CrossModal	Bui.	Tre.	Low.	Car	Imp.	OA	F1	mIoU
PSPNet	N	94.52	90.17	78.84	79.22	92.03	89.94	86.55	76.96
MAResU-Net	N	94.84	89.99	79.09	85.89	92.19	90.17	88.54	79.89
vFuseNet	Y	95.92	91.36	77.64	76.06	91.85	90.49	87.89	78.92
FuseNet	Y	96.28	90.28	78.98	81.37	91.66	90.51	87.71	78.71
ESANet	Y	95.69	90.50	77.16	85.46	91.39	90.61	88.18	79.42
TransUNet	Y	96.48	**92.77**	76.14	69.56	91.66	90.96	87.34	78.26
SA-GATE	Y	94.84	92.56	**81.29**	**87.79**	91.69	91.10	**89.81**	81.27
CMFNet	Y	**97.17**	90.82	80.37	85.47	**92.36**	**91.40**	89.48	**81.44**

Bold values are the best, while underlined values the second best.

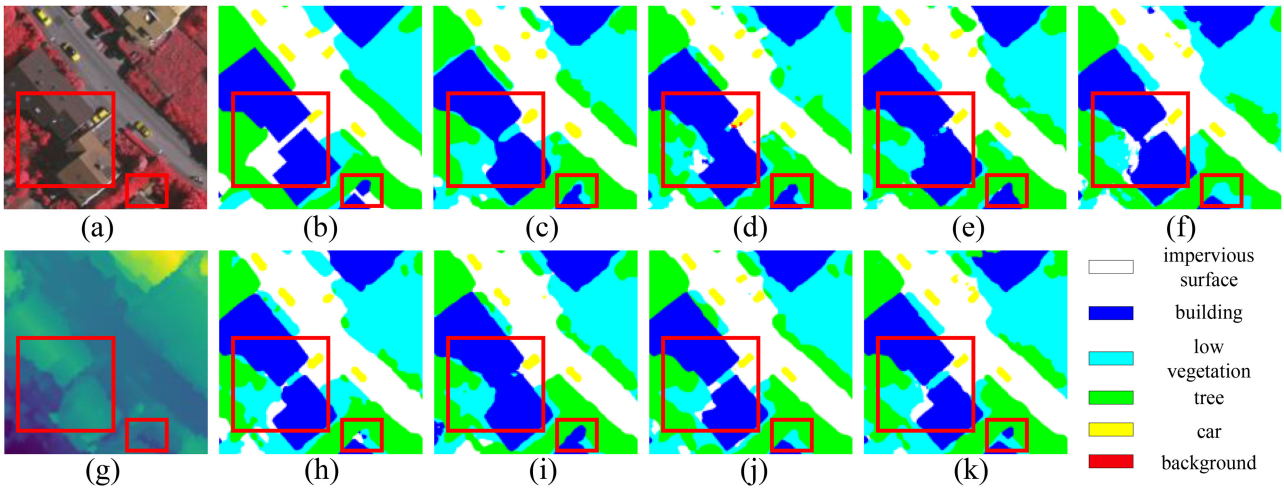


Fig. 6. Qualitative performance comparisons on the Vaihingen test set. (a) RGB images. (b) Ground truth. (c) PSPNet. (d) MAResU-Net. (e) vFuseNet. (f) FuseNet. (g) DSM. (h) ESANet. (i) TransUNet. (j) SA-GATE. (k) Proposed CMFNet.

complexity by adopting a linear attention mechanism (LAM). TransUNet [30] utilizes the transformer architecture to enhance the feature map processed by the CNN, whereas the proposed CMFNet uses the transformer architecture mainly for cross-modal fusion. SA-GATE [44] presents that depth data are generally noisy and designs separation-and-aggregation gate to fuse RGB and DSM (RGB-D) data. ESANet [45] shows an efficient RGB-D segmentation approach by two enhanced ResNet-based encoders utilizing an attention-based fusion module. In our experiments, both PSPNet and MAResU-Net only consider the main modal information, i.e., the RGB images. In contrast, other methods take into account crossmodal RGB-D data.

As listed in Table I, the proposed CMFNet improved the classification accuracy for all classes as compared to our baseline FuseNet, which confirmed that the proposed crossmodal and multiscale mechanism successfully extracted the complementary information between modalities and effectively utilized the multiscale information. Compared with existing state-of-the-art methods, the CMFNet outperformed on two classes, namely *building* and *imprevious surface*. In particular, on the Vaihingen dataset, the CMFNet provided the most significant improvement on *low vegetation* class with an increase of 1.28% as compared to the existing method MAResU-Net. Furthermore, the classification accuracy for *buildings* has been improved by 0.89%. This improvement can be explained by the fact that the RGB images fail to capture the elevation information that is available in the DSM images. As a result, elevation information contained in the DSM images, if exploited effectively, can greatly improve the quality of the features for classes of noticeable elevation values. For instance, *buildings* generally have a uniform and large DSM value, whereas the *low vegetation* a low DSM value. Thus, the improvement on these two classes are particularly impressive. In terms of the overall performance, the proposed CMFNet achieved OA of 91.40%, F1-score of 89.48%, and mIoU of 81.44%, which stands for an increase of 0.89%, 0.94%, and 1.55% as compared to the corresponding performance of FuseNet, respectively. These results confirmed that the proposed CMFNet achieved a better generalization performance. Fig. 6 shows a visualization example of the results obtained by all eight methods under consideration. The rectangle area highlights the performance difference. Clearly, it can be observed that the proposed CMFNet is able to identify complex edges of buildings with smoother results while providing an accurate classification over shadow areas, which is often very challenging for RGB image-based methods.

Experiments on the Potsdam dataset have also shown similar results. As shown in Table II, the classification accuracy rates for *buildings*, *trees*, and *impervious surfaces* were 97.63%, 87.40%, and 92.84%, respectively, which amounts to an increase of 0.15%, 2.26%, and 0.20% as compared to FuseNet. The corresponding OA, F1 score, and mIoU values were 91.16%, 92.10%, and 85.63%, respectively, which corresponds to increases of

TABLE II
EXPERIMENTAL RESULTS ON THE POTSDAM DATASET

Method	CrossModal	Bui.	Tre.	Low.	Car	Imp.	OA	F1	mIoU
SA-GATE	Y	96.54	81.18	85.35	**96.63**	90.77	87.91	90.26	82.53
PSPNet	N	97.03	83.13	85.67	88.81	90.91	88.67	88.92	80.36
ESANet	Y	97.10	85.31	87.81	94.08	92.76	89.74	91.22	84.15
MAResU-Net	N	96.82	83.97	87.70	95.88	92.19	89.82	90.86	83.61
TransUNet	Y	96.63	82.65	**89.98**	93.17	91.93	90.01	90.97	83.74
vFuseNet	Y	97.23	84.29	89.03	95.49	91.62	90.22	91.26	84.26
FuseNet	Y	97.48	85.14	87.31	96.10	92.64	90.58	91.60	84.86
CMFNet	Y	**97.63**	**87.40**	88.00	95.68	**92.84**	**91.16**	**92.10**	**85.63**

Bold values are the best while underlined values the second best.

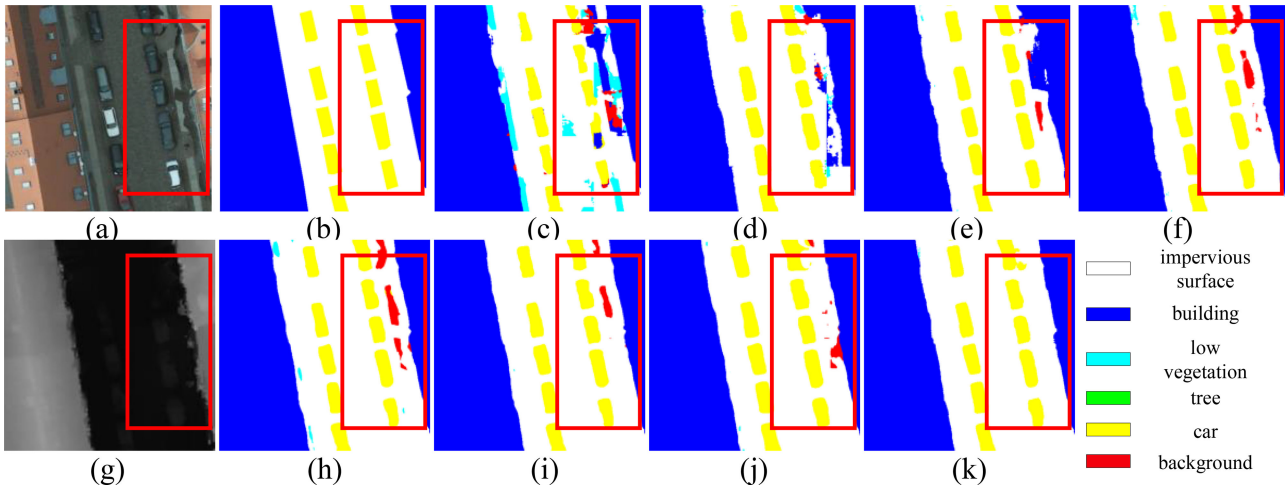


Fig. 7. Qualitative performance comparisons on the Potsdam test set. (a) RGB images. (b) Ground truth. (c) PSPNet. (d) MAResU-Net. (e) vFuseNet, (f) FuseNet. (g) DSM. (h) ESANet. (i) TransUNet. (j) SA-GATE. (k) Proposed CMFNet.

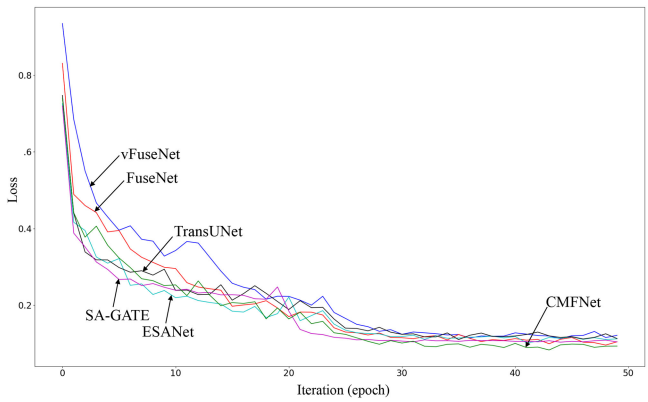


Fig. 8. Convergence behavior during training on Vaihingen.

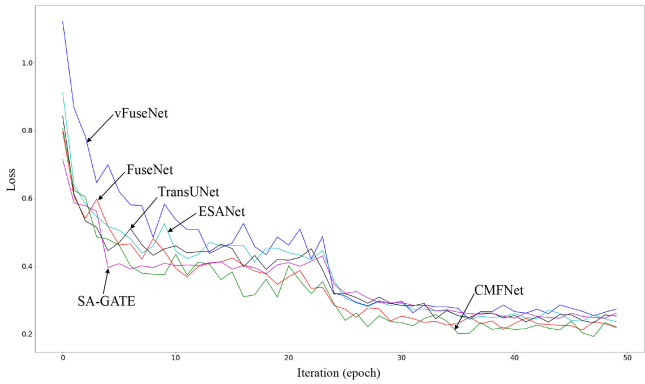


Fig. 9. Convergence behavior during training on Potsdam.

0.58%, 0.50%, and 0.77%, respectively, over FuseNet. Fig. 7 illustrates an example showing the performance comparison of all methods. In particular, it is observed that nonbuildings that were difficult to be identified using the RGB images could be much more accurately classified by the proposed crossmodal and multiscale mechanism. In sharp contrast, vFuseNet and FuseNet misclassified such nonbuilding objects as buildings and complex background.

Finally, Figs. 8 and 9 show the convergence behaviors of the six crossmodal methods among the eight methods investigated, namely, FuseNet, vFuseNet, ESANet, TransUNet, SA-GATE, and the proposed CMFNet in terms of loss defined in (9) on Vaihingen and Potsdam, respectively. Inspection of Figs. 8 and 9 suggests that the proposed CMFNet achieved comparable or even better convergence performance as compared to those existing methods. This is because that effective crossmodal and multiscale information fusion through the transformer structure helped the proposed CMFNet to more efficiently characterize surface objects by exploiting crossmodal feature fusion.

TABLE III
ABLATION STUDY ON THE VAIHINGEN DATASET

Crossmodal	Multiscale	MCATrans	OA
	✓	✓	91.00
✓		✓	91.19
✓	✓		91.30
✓	✓	✓	**91.40**

TABLE IV
SCALE ANALYSIS ON THE VAIHINGEN DATASET

Method	OA	F1	mIoU
Pure CSE	90.70	88.93	80.51
Pure Trans	90.52	88.45	79.74
Trans1-3 + CSE4-5	90.80	89.07	80.74
CMFNet	**91.40**	**89.48**	**81.44**

E. Ablation Study

To verify the effectiveness of each module in CMFNet, ablation experiments are carried out by removing certain modules. As shown in Table III, three sets of ablation experiments are designed based on CMFNet. In the first row, CMTrans is disassembled into two single-modal multiscale transformers, i.e., two MCATrans modules. This control group is used to exhibit the performance gain due to crossmodal information extraction by CMTrans. Furthermore, since the complexity of the model remains unchanged in this design, it is proved that the improvement in the CMFNet performance is not due to the increase of model parameters. In the second row, CMTrans is decomposed into four single-scale crossmodal tranformers. In this group, the crossmodal information of different scales is fused separately. The results of this experiment proves our point of view, that is, learning different scales of crossmodal information simultaneous is vital in data fusion. MCAtrans is removed in the third row to show that the global context enhanced is also helpful. Complete ablation experiments show that each module in the CMFNet has its own unique role, and we design our network based on the characteristics of crossmodal and multiscale information.

F. Scale Analysis

In this section, we compare the performance of transformer-based networks with structural variations. More specifically, we consider the following three methods. The first method uses the CSE module to fuse feature maps of all five scales before adding the resulting maps as a skip connection to the corresponding decoding layer. This method is referred to as "Pure CSE" in the sequel. In the second method, simultaneous crossmodal fusion of five scales is performed with CMTrans and MCATrans. We call this method "Pure Trans". Finally, the third method is called "Trans1-3 + CSE4-5" that uses two CSE modules to fuse high-level features, namely the fourth and fifth levels while using CMTrans and MCATrans to fuse the first three low-level features. Note that the proposed CMFNet uses CMTrans and MCATrans to perform crossmodal and multiscale fusion on the features of the first four levels, while a single CSE is employed for the benefit of reduced computational complexity. In other words, CMFNet has the "Trans1-4 + CSE5" structure. The experimental results are shown in Table IV.

Table IV shows that the synergy of the transformer and CSE is effective only if they are used properly as designed in the proposed CMFNet. For instance, choosing a proper scale for fusion can have a major impact on the performance. Specifically, as features become more abstract and globally significant after multiple layers of processing, CSE is more suitable for the fusion of these high-level features. In contrast, a nature global view derived from transformers is more suitable for simultaneous processing of multiscale information in low-level layers.

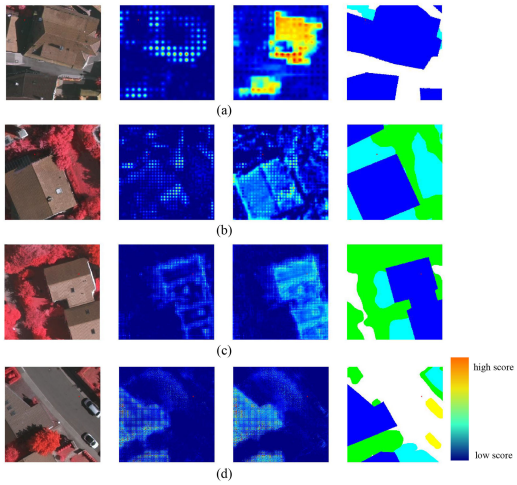


Fig. 10. Comparison of heatmaps generated before and after fusion. (a) Level-4 features. (b) Level-3 features. (c) Level-2 features. (d) Level-1 features. The first column is the remote sensing image of the validation set, whereas the second and third columns are the heatmaps before and after adding the features to their corresponding decoders, respectively. Finally, the last column is the ground truth. Note that the level-1 heatmaps have the highest resolution with many scattered pixels of high scores indicated by their bright colors.

G. Visualization Via Gradient-Based Localization

To shed light on the performance improvement achieved by the proposed CMFNet, we used Grad-CAM [46] to visually inspect the output of each decoder layer. Grad-CAM was originally developed to visualize the output of the intermediate steps leading to the final result in image classification. However, unlike the task of image classification that labels each image with one single class, semantic segmentation for remote sensing images is performed in a pixel-by-pixel manner.

To address this difference, we propose to modify the Grad-CAM method for semantic segmentation applications in order to visualize the rationale behind the classification decision of CMFNet of a given pixel. More specifically, we propose to visualize and compare the features before and after the fusion of different layers for points of interest. Fig. 10 shows the features of different levels to classify pixels as *buildings* or not. Inspection of Fig. 10 shows that the third-column illustration has higher score points, which suggests that the CMFNet could better identify pixels belonging to *buildings* by fusing multiscale and crossmodal features.

Finally, Fig. 11 compares the heatmaps from FuseNet and the proposed CMFNet. In Fig. 11, the images on the first row

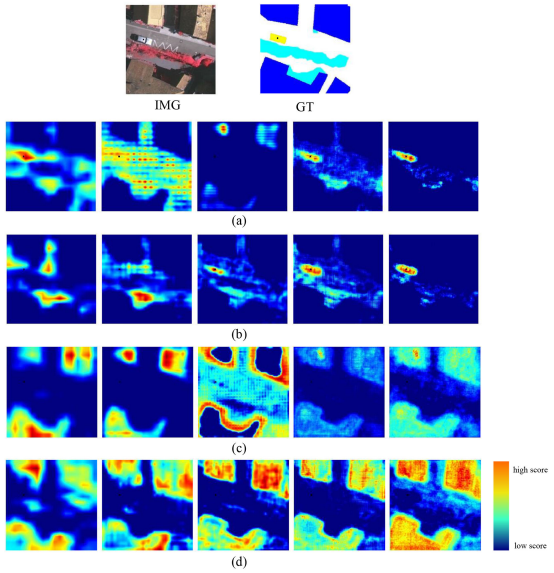


Fig. 11. (a) and (b) Heatmaps of the FuseNet encoder and CMFNet encoder. (c) and (d) Heatmaps of the FuseNet decoder and CMFNet decoder. The heatmaps in the first and second rows show the heatmaps generated by the two networks to determine if a pixel belongs to *cars* or not. In contrast, the third and fourth rows depict the heatmaps generated by the two networks to decide if a pixel belongs to *buildings* or not. Note that a higher score indicated by brighter colors suggests higher likelihood for the pixel belonging to the class under consideration.

(from left to right) are the original remote sensing image and the ground truth, respectively. Furthermore, Fig. 11(a) and (b), i.e., the first and second rows, show the heatmaps of different levels generated by the FuseNet decoders and the CMFNet decoders, respectively, when these two networks tried to determine if each pixel should be classified as *cars* or not. For reference, we marked a black dot on the car on the street. Note that pixels in brighter colors have higher scores and are more likely to be classified as *cars*. From Fig. 11(a) and (b), it is observed that the CMFNet was able to extract better features as the decoding process progressed. In contrast, Fig. 11(c) and (d) illustrates the heatmaps of different levels when FuseNet and CMFNet tried to decide if each pixel should be classified as *buildings* or not. Again, the CMFNet was able to generate more accurate features to identify those pixels belonging to buildings more rapidly. These visualization results confirmed that the proposed crossmodal multiscale fusion architecture can generate better features more effectively and efficiently, which leads to better segmentation performance.

To illustrate the capability of CMTrans and MCATrans to capture detailed local information, we examined the local prediction labels in Fig. 12. Visual inspection of Fig. 12(a), (c), (e), and (f) suggests that the CMFNet could accurately identify the building edges by comparing the results against the corresponding RGB images. In particular, the CMFNet could accurately detect *cars* covered by plants, as shown in Fig. 12(b). Furthermore, the CMFNet can predict smoother boundaries even for blurred edges in Fig. 12(d). Fig. 12 confirmed that the CMTrans and MCATrans can better identify local objects by computing point-to-point correlation on feature maps before effectively exploiting the information at the global scale.

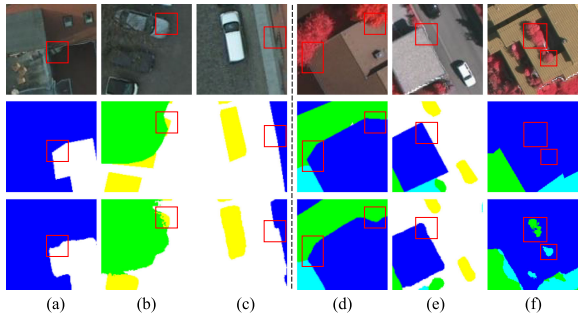


Fig. 12. From top to bottom, each row shows the RGB, ground truth, and prediction, respectively. (a)–(c) Some detailed information in the results of Potsdam. (d)–(f) Vaihingen.

V. Conclusion

In this work, a novel network architecture called CMFNet has been proposed for high-resolution semantic segmentation of remote sensing imagery by exploiting cross-attention and crossmodal transformers. In particular, CMA layers, CMTrans, and MCATrans have been developed to fuse features of different scales from multimodal data. As a result, the CMFNet can effectively model long-range dependencies across multiscale feature maps of remote sensing data in different modalities and learn discriminative integrated representations for semantic segmentation. Furthermore, a novel residual method has been established to reduce ambiguous features by connecting the output of the crossmodal and multiscale feature fusion to the decoders. Extensive simulation results on the well-known ISPRS Vaihingen and Potsdam datasets have confirmed that the proposed CMFNet can provide the most accurate segmentation performance. Finally, Grad-CAM-based visualization has been utilized to provide intuitive interpretation on the effectiveness achieved by CMFNet in feature extraction.

References

[1] P. Ghamisi *et al.*, "Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 1, pp. 6–39, Mar. 2019.

[2] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 937–949, Feb. 2017.

[3] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2018.

[4] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[5] D. Xu, Y. Wang, X. Zhang, N. Zhang, and S. Yu, "Infrared and visible image fusion using a deep unsupervised framework with perceptual loss," *IEEE Access*, vol. 8, pp. 206445–206458, 2020.

[6] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1882–1886, Dec. 2016.

[7] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, "Random forests for land cover classification," *Pattern Recognit. Lett.*, vol. 27, no. 4, pp. 294–300, 2006.

[8] X. Lu, J. Zhang, T. Li, and G. Zhang, "Synergetic classification of long-wave infrared hyperspectral and visible images," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 8, no. 7, pp. 3546–3557, Jun. 2015.

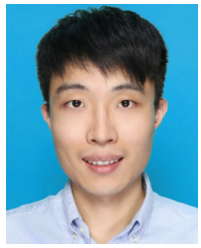
[9] L. Gao *et al.*, "Subspace-based support vector machines for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 2, pp. 349–353, Feb. 2014.

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

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2015, pp. 234–241.
[12] D. Hong, L. Gao, R. Hang, B. Zhang, and J. Chanussot, "Deep encoder-decoder networks for classification of hyperspectral and LiDAR data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, Aug. 2020.
[13] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, 2020.
[14] Y. Sun, Z. Fu, C. Sun, Y. Hu, and S. Zhang, "Deep multimodal fusion network for semantic segmentation using remote sensing image and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, Sep. 2021.
[15] H. Li, X.-J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *Proc. 24th Int. Conf. Pattern Recognit.*, 2018, pp. 2705–2710.
[16] J. Chen, X. Li, L. Luo, X. Mei, and J. Ma, "Infrared and visible image fusion based on target-enhanced multiscale transform decomposition," *Inf. Sci.*, vol. 508, pp. 64–78, 2020.
[17] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 213–228.
[18] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RCB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, 2018.
[19] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 5998–6008, 2017.
[20] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[21] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang, "Deep learning for pixel-level image fusion: Recent advances and future prospects," *Inf. Fusion*, vol. 42, pp. 158–173, 2018.
[22] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 457–468.
[23] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," *Adv. Neural Inf. Process. Syst.*, vol. 31, pp. 1571–1581, 2018.
[24] J.-H. Kim, K. W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," in *Proc. 5th Int. Conf. Learn. Representations*, 2017, pp. 1–14.
[25] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1103–1114.
[26] R. Chen *et al.*, "Developing measures of cognitive impairment in the real world from consumer-grade multimodal sensor streams," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2019, pp. 2145–2155.
[27] F. Mahmood, Z. Yang, R. Chen, D. Borders, W. Xu, and N. J. Durr, "Polyp segmentation and classification using predicted depth from monocular endoscopy," *Proc. SPIE*, vol. 10950, pp. 268–272, 2019.
[28] H. Ghassemian, "A review of remote sensing image fusion methods," *Inf. Fusion*, vol. 32, pp. 75–89, 2016.
[29] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6558–6569.
[30] J. Chen *et al.*, "TransuNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.
[31] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "TransBTS: Multimodal brain tumor segmentation using transformer," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2021, pp. 109–119.
[32] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, "A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, Jan. 2022.
[33] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
[34] H. Cao *et al.*, "Swin-Unet: Unet-like pure transformer for medical image segmentation," 2021, *arXiv:2105.05537*.
[35] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7077–7087.
[36] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "UCTransNet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer," 2021, *arXiv:2109.04335*.
[37] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and CNNs for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2021, pp. 14–24.
[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015, pp. 1–14.
[39] 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.
[40] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[41] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.
[42] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.
[43] R. Li, S. Zheng, C. Duan, J. Su, and C. Zhang, "Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, Mar. 2021.
[44] X. Chen *et al.*, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 561–577.
[45] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, and H.-M. Gross, "Efficient RGB-D semantic segmentation for indoor scene analysis," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 13525–13531.
[46] 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," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

Xianping Ma received the bachelor's degree in geographical information science from Wuhan University, Wuhan, China, in 2019. He is currently working toward the Ph.D. degree with the Chinese University of Hong Kong, Shenzhen, China.

His research interests include remote sensing and machine learning.



Xiaokang Zhang (Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, in 2018.

He is currently a Postdoctoral Researcher with the School of Science and Engineering, Chinese University of Hong Kong, Shenzhen, China. His research interests include multitemporal remote sensing image analysis and machine learning.

Man-On Pun (Senior Member, IEEE) received the B.Eng. degree in electronic engineering from the Chinese University of Hong Kong, Hong Kong, in 1996, the M.Eng. degree in computer science from University of Tsukuba, Tsukuba, Japan in 1999, and the Ph.D. degree in electrical engineering from the University of Southern California (USC), Los Angeles, CA, USA, in 2006, respectively.

He was a Postdoctoral Research Aassociate with Princeton University from 2006 to 2008. He is currently an Associate Professor with the School of Science and Engineering, Chinese University of Hong Kong, Shenzhen (CUHKSZ), Shenzhen, China. Prior to joining the CUHKSZ in 2015, he held research positions with Huawei (USA) in New Jersey, Mitsubishi Electric Research Labs (MERL) in Boston, MA, USA, and Sony in Tokyo, Japan. His research interests include artificial intelligence Internet of Things and applications of machine learning in communications and satellite remote sensing.

Prof. Pun was the recipient of best paper awards from 2006 IEEE Vehicular Technology Conference Fall, 2008 IEEE International Conference on Communications, and 2009 IEEE Conference on Computer Communications. He served as an Associate Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS in 2010–2014. He is the Founding Chair of the IEEE Joint SPS-ComSoc Chapter, Shenzhen.