LMF-Net: A Learnable Multimodal Fusion Network for Semantic Segmentation of Remote Sensing Data

Jihao Li, *Member, IEEE*, Wenkai Zhang, *Member, IEEE*, Weihang Zhang, *Student Member, IEEE*, Ruixue Zhou, *Member, IEEE*, Chongyang Li, *Member, IEEE*, Boyuan Tong, *Student Member, IEEE*, Xian Sun, *Senior Member, IEEE*, and Kun Fu, *Member, IEEE*

***Abstract*—Semantic segmentation of remote sensing images has produced a significant effect on many applications, such as land cover, land use, and smoke detection. With the ever-growing remote sensing data, fusing multimodal data from different sensors is a feasible and effective scheme for semantic segmentation task. Deep learning technology has prominently promoted the development of semantic segmentation. However, the majority of current approaches commonly focus more on feature mixing and construct relatively complex architectures. The further mining for cross-modal features is comparatively insufficient in heterogeneous data fusion. In addition, complex structures also lead to relatively heavy computation burden. Therefore, in this article, we propose an end-to-end learnable multimodal fusion network (LMF-Net) for remote sensing semantic segmentation. Concretely, we first develop a multiscale pooling fusion module by leveraging pooling operator. It provides key-value pairs with multimodal complementary information in a parameter-free manner and assigns them to self-attention (SA) layers of different modal branches. Then, to further harness the cross-modal collaborative embeddings/features, we elaborate two learnable fusion modules, learnable embedding fusion and learnable feature fusion. They are able to dynamically adjust the collaborative relationships of different modal embeddings and features in a learnable approach, respectively. Experiments on two well-established benchmark datasets reveal that our LMF-Net possesses superior segmentation behavior and strong generalization capability. In terms of computation complexity, it achieves competitive performance as well. Ultimately, the contribution of each component involved in LMF-Net is evaluated and discussed in detail.**

***Index Terms*—Deep learning (DL), learnable fusion, multimodal data, remote sensing, semantic segmentation.**

Received 26 November 2024; revised 29 December 2024; accepted 4 January 2025. Date of publication 8 January 2025; date of current version 24 January 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62331027 and Grant 62425115, and in part by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant XDA0360300. *(Corresponding author: Wenkai Zhang.)*

Jihao Li and Ruixue Zhou are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, and also with the Key Laboratory of Network Information System Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China (e-mail: lijihao17@mails.ucas.ac.cn; zhourx@aircas.ac.cn).

Wenkai Zhang, Weihang Zhang, Chongyang Li, Boyuan Tong, Xian Sun, and Kun Fu are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, also with the Key Laboratory of Network Information System Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, also with the University of Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: zhangwk@aircas.ac.cn; zhangweihang21@mails.ucas.ac.cn; lichongyang22@mails.ucas.ac.cn; tongboyuan22@mails.ucas.ac.cn; sunxian@aircas.ac.cn; kunfuiecas@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2025.3527213

I. Introduction

With the constant breakthrough of sensor technology, the acquisition of remote sensing data is becoming increasingly convenient, leading to a kind of exponential growth of remote sensing data. Moreover, the types of remote sensing data are also becoming more diverse, including optical data, LiDAR data, synthetic aperture radar (SAR) data, etc. The massive amount of remote sensing data and abundant data type have provided a strong data support for the application of remote sensing interpretation in the fields of ecological protection [1], [2], national defense construction [3], [4], etc. The industrial application based on remote sensing technology is also developing toward diversification [5], [6], [7].

Semantic segmentation, which aims to identify the semantic label of each pixel via an automated model, is a critical one among various remote sensing interpretation tasks. Due to the greatly significant effect on land cover and land use [8], [9], smoke detection [10], [11], water body extraction [12], [13], urban management [14], [15], [16], etc., semantic segmentation has gained widespread attention. Drawing support from deep learning (DL) technology, the performance of semantic segmentation has made a considerable advancement.

However, these DL approaches are generally applicable to single-modal remote sensing data, e.g., optical images. While single-modal data are vulnerable to obstruction, noise, lighting, etc., causing the semantic representation capability to be relatively limited. As illustrated in Fig. 1(a), it is difficult to discriminate whether the red object in the box is *low vegetation* or *tree* only through this optical image, because these two categories have specially similar colors. And the shadow of this object is obscured by a nearby building, making it impossible to speculate the height through its shadow. While the digital surface model (DSM) image presents that this object has an obvious height above the ground. Accordingly, its category can be determined as *tree*. Similarly, in Fig. 1(b), it is unable to distinguish the category of the object in box just merely relying on the elevation information of DSM. Yet, the color and texture features in optical image can clearly reflect that it belongs to the category of *building*.

Evidently, different modal remote sensing data have their own unique characteristics, such as color information of optical data and elevation information of DSM data. In the context of increasingly abundant remote sensing data, incorporating multimodal data is a feasible and effective solution for remote

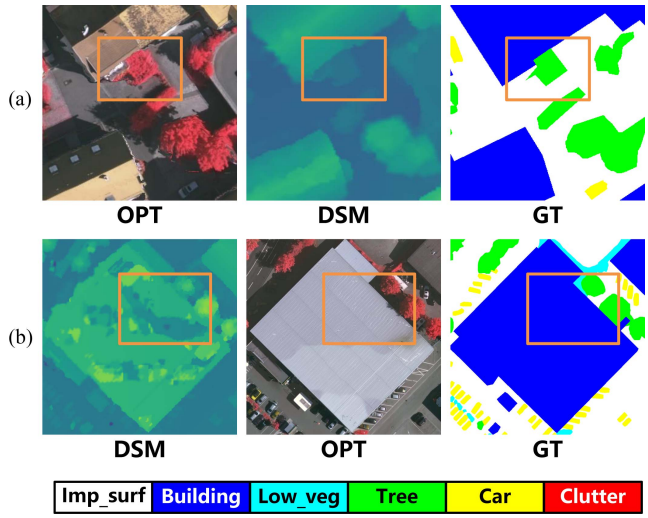

Fig. 1. Limitations of single-modal remote sensing data. **OPT**, **DSM**, and **GT** means optical image, DSM image, and ground truth, respectively. Best viewed in color.

sensing semantic segmentation task. The comprehensive utilization of these multimodal data can contribute to exert the effect of complementary information. Therefore, many scholars have conducted explorations in this research direction. Fan et al. [17] built a hierarchical fusion structure by leveraging cross-layer features. Zheng et al. [18] employed category-prior knowledge to guide multimodal fusion under the situation of sample imbalance. Besides, MoCG [19] and FTransUNet [20] are also typical representative networks for utilizing cross-modal attention mechanism. Nevertheless, the majority of current approaches commonly concern more on feature mixing and construct relatively complex architectures. The cross-modal features are underutilized in heterogeneous data fusion, and the information exchange between fused collaborative features and modal-specific features is comparatively insufficient. In addition, complex structures also result in a relatively high computation burden, posing challenges for their practical application.

For this purpose, in this article, we propose a novel learnable multimodal fusion network (LMF-Net) for semantic segmentation task to address the problems analyzed previously. To be concrete, we first design a multiscale pooling fusion (MSPF) module to assign key-value pairs with complementary information in a parameter-free mean to SA layers of different modal branches. Then, two learnable fusion modules are developed, i.e., learnable embedding fusion (LEF) and learnable feature fusion (LFF). They fully utilize cross-modal information and are able to dynamically adjust the collaborative relationships of different modal embeddings and features in a learnable manner. Finally, we integrate MSPF, LEF, and LFF into an encoder–decoder segmentation architecture and establish the end-to-end LMF-Net. The achieved performances on two challenging multimodal semantic segmentation datasets, ISPRS Vaihingen [21] and ISPRS Potsdam [21], confirm the effectiveness and practicability of the proposed LMF-Net.

Our main contributions of this article can be succinctly summarized in the following.

1) This article designs an MSPF module to provide key-value pairs with complementary information in a parameter-free approach by leveraging pooling operator.
2) This article exploits an LEF module to dynamically adjust the collaborative relationships of cross-modal 1-D embeddings in a learnable manner.
3) This article develops an LFF module to realize learnable collaboration of cross-modal 2-D features in heterogeneous data fusion.
4) This article conducts extensive experiments and detailed analyses on two well-established benchmarks to evaluate the segmentation capability and computation efficiency of the LMF-Net.

The rest of this article is organized as follows. Section II provides a brief review of some relevant achievements from single-modal semantic segmentation to multimodal semantic segmentation. Section III offers a detailed description of the proposed LMF-Net framework. In Section IV, we implement sufficient experiments on two challenging datasets to verify the performance and generalization capability of our LMF-Net pipeline. In this Section, the experimental settings and analyses are fully presented as well. Then, a depth discussion about the proposed LMF-Net architecture is given in Section V. Finally, Section VI conclude this article.

II. Related Work

In this section, we briefly review several remarkable research achievements based on DL technology in single-modal segmentation and multimodal segmentation in series.

A. Single-Modal Segmentation in Natural Scene

CNN significantly promotes the progress of visual recognition technology. Based on the CNN architecture, Long et al. [22] proposed a fully convolutional network for dense pixel prediction. It is the first CNN segmentation model which makes end-to-end training available. Chen et al. [23], [24] established a significant structure by leveraging atrous convolution and its derivative atrous spatial pyramid pooling module, called DeepLab. Owing to the outstanding behavior, DeepLab series occupy an important position in semantic segmentation task. In the same period, Zhao et al. [25] introduced a type of pyramid pooling module (PPM) to capture the contextual information and multiscale information of images, named PSPNet. It substantially reduces the segmentation errors and effectively improves the accuracy of semantic understanding. In addition, U-Net [26] is also a kind of typical encoder–decoder network. It has been widely adopted in medical image interpretation and has had profound inspiration for subsequent work [27], [28], [29].

With the proposal of SETR [30], semantic segmentation steps into the era of transformer [31]. Based on the successful application of ViT [32] in vision tasks, SETR constructs a *ViT+decoder* framework and converts image semantic segmentation to a sequence-to-sequence prediction task. Unlike SETR whose decoder is a CNN architecture, segmenter [33] utilizes a mask transformer to decode semantic features. It also enables the patches and class embeddings to be processed jointly. Xie

et al. [34] elaborately designed a lightweight model, dubbed SegFormer. It simplifies SA operation and replaces positional encoding with a 3×3 depthwise (DW) convolution [35]. Experiments imply that SegFormer not only achieves satisfactory results, but also alleviates the problem of high computation overhead. Empowered with the extraordinary capability, transformer architecture has gradually become a popular scheme in semantic segmentation task and has enlightened many following studies [36], [37], [38], [39], [40].

B. Single-Modal Segmentation in Remote Sensing

Along with the improvement of Earth observation technology, the number of remote sensing images manifests an explosive growth. Aiming to fulfill the demand for refined classification of massive-scale data, many researchers have also extended DL to remote sensing semantic segmentation. ResUNet-a [41] is a representative work of introducing the thought of computer vision methodology into remote sensing interpretation. It integrates the design concepts of residual connection [42], atrous convolution [23], [24], PPM [25], etc., and its effectiveness is demonstrated through a series of quantitative and qualitative experiments. Ma et al. [43] proposed an efficient segmentation network, called FactSeg that principally focuses on small objects. In the respect of its network structure, foreground activation is applied to activate small-scale objects while to suppress background. From the perspective of optimization, the small object mining approach is used to tackle the sample imbalance issue. In order to improve the segmentation results of fine-structured geographic entities, Deng et al. [44] leveraged attention mechanism to develop a CCANet. It constructs a kind of class information constraint to acquire the explicit long-range dependency and achieves impressive performance in extensive experiments. Furthermore, motivated by giving consideration to both local information and long-range contextual dependency, Wu et al. [7] presented CMTFNet that combines the local feature extraction capability of CNN with the global feature expression capability of transformer. It shows a strong competitiveness compared with many existing popular models. Moreover, [45], [46], [47], [48] are also preeminent methods with significant influence on remote sensing semantic segmentation for specific objects, such as photovoltaic panels, roads, and buildings.

C. Multimodal Segmentation in Remote Sensing

Since the imaging mechanism is diverse, remote sensing data exhibit various modalities, such as optical images, hyperspectral images, SAR data, DSMs, and so on. Currently, there is a broad consensus that different modal data have complementary information [18], [49], [50], [51], [52]. Hence, the field of multimodal interpretation is attracting an increasing number of researchers. Wu et al. [53] exploited a novel cross-modal interaction framework which combines a cross fusion module and a multimodal features aggregation module to aggregate optical feature and SAR feature. It shows relatively obvious superiority in urban impervious surface segmentation task. Nonetheless, in terms of the fusion for the optical image and the SAR image, on the whole, data limitation is a major obstacle to the development of this research direction. To this end, Li et al. [54] and Ren [52] separately constructed the WHU-OPT-SAR dataset and a fusion dataset covering Xi'an, Dongying, and Pohang, and thence, introduce MCANet and DDHRNet on the basis of attention thought. Their work produces a notable advancement for this area. The authors in [55] and [56] developed a CNN-based and a transformer-based fusion manner for joint hyperspectral image and LiDAR data segmentation, respectively. Both of them are able to improve the extraction capability for multimodal features and their generalization performances have also been validated on various benchmarks. The authors in [18] and [57] built cleverly designed fusion models by employing CNN architecture to realize the complementarity between texture information and elevation characteristics. Besides, CMFNet [58] exerts the advantage of transformer architecture to learn long-range dependencies of fused features, contributing to capture more discriminative expressions in optical and DSM data. It reduces the decoding ambiguity to a certain extent as well. In order to excavate informative cues from low-quality feature maps of DSM and capture multiscale adjacent contextual features, Ma et al. [59] presented ABHNet. The collaboration of detail and deep features brings a definite enhancement on small-scale object segmentation.

Although many breakthroughs have been made in multimodal remote sensing semantic segmentation, we argue that there is still much potential space for further exploitation of cross-modal collaborative features. Motivated by this viewpoint, we present a novel LMF-Net for the fusion of multimodal remote sensing data. And to verify the availability of this framework, we also conduct a series of experiments and compare it with many state-of-the-art methods.

III. Methodology

In this section, we provide a comprehensive introduction of the proposed LMF-Net. Specifically, Section III-A overviews the architecture of the entire LMF-Net. It is carefully elaborated for the diversity of remote sensing data type. The core ingredients of the LMF-Net, including MSPF, LEF, and LFF are exhaustively illustrated in Sections III-B–III-D.

A. Network Architectiure

As shown in Fig. 2, the proposed LMF-Net is an end-to-end model. It adopts a kind of encoder–decoder architecture in semantic segmentation based on a two-stream SegFormer [34] model. LMF-Net takes different modal data as input. For the sake of clarity, we utilize modality1 and modality2 for explanation in the following sections. And in our experiments, the two input types are optical remote sensing images and DSM data, respectively.

According to Fig. 2, we set four stages in encoder to gradually reduce the resolution of the inputs. Except for the first stage which uses four times downsampling operation, all others use two times downsampling operation. That means, assuming that the height and the width of the inputs are H and W. After feature extraction of each stage, their size will become $H/4 \times W/4$,

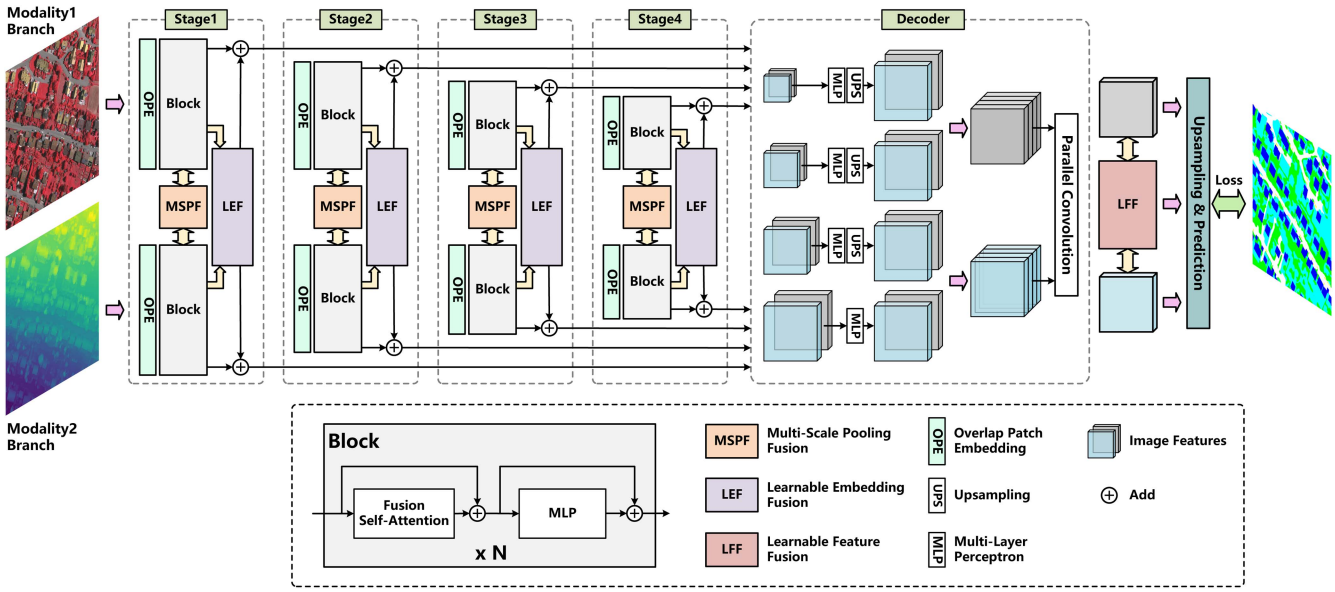

Fig. 2. Overview of the proposed LMF-Net. It is a two-stream architecture. The left half is a four-stage hierarchical transformer encoder. The right half is a decoder that integrates multilevel features. Best viewed in color.

$H/8 \times W/8$, $H/16 \times W/16$, and $H/32 \times W/32$, respectively. In each stage, overlap patch embedding (OPE) is first to serialize the features of image form. Compared to nonoverlap pattern, patches with overlap are more conducive to preserve the continuity of local structure. After the OPE, we set a transformer block which is composed of $N\times$ repeated fusion self-attention (FSA) and multilayer perceptron (MLP). The workflows and parameter settings of OPE and block are consistent with those of the SegFormer [34] model. To mix the information from modality1 and modality2 in FSA, we introduce a multiscale cross-modal fusion module by leveraging pooling operator, named MSPF. Besides, we further explore a learnable fusion scheme for 1-D image embeddings of the two modalities at the output of each stage, which can be termed as LEF. Details about the presented MSPF and LEF will be described in the subsequent sections.

Correspondingly, in order to reconstruct the input data, we collect the output features from each stage and individually perform MLP and upsampling (UPS) operation. Notice that all features are temporarily restored to $H/4 \times W/4$. As a consequence, for features in stage1, only MLP is performed, but UPS is excluded. Next, we concatenate these reconstructed features of different resolutions from the two modal branches, and then compress their channels through a parallel convolution unit. After that, we exploit another learnable fusion scheme, called LFF, which is specifically designed for 2-D image features. In LFF, modality1 branch, modality2 branch and the fusion branch all contribute to the final prediction. Ultimately, after a four times UPS operation, the full-resolution segmentation results are obtained. And parameters of the proposed LMF-Net will be optimized toward the direction in which the objective function decreases. For the network inference, the output format is the same as that of the network training.

B. MSPF

Due to its superiorities, for instance, satisfactory effect and simple architecture, transformer framework has been widely recognized in many fields. However, MLP-like structure requires a significant amount of computation resources. For images, particularly for remote sensing images with extremely large size, the extra-long sequence length will pose a prohibitive cost for the training of the model. For this reason, many scholars have also explored solutions represented by pooling operator to reduce the computation complexity of MLP-like transformer [60], [61], [62], [63]. Inspired by these valuable studies, we introduce this parameter-free method into the task of remote sensing multimodal interpretation as well and present MSPF to reduce the computation overhead. Based on the aforementioned pioneering work, MSPF further integrates multiscale heterogeneous remote sensing features and assigns these fusion features to SA unit of different modal branches. The workflow of the proposed MSPF module is given in Fig. 3.

We suppose that X and Y denote the features of modality1 and modality2, respectively. Note that both X and Y are in the form of 2-D image. Similar to [63], we also adopt pyramid pooling approach to capture implicit contextual information. It can be expressed as follows:

$$F_i^X = \text{pool}_i(X),\ i = 1, 2, \ldots, N \tag{1}$$

$$F_i^Y = \text{pool}_i(Y),\ i = 1, 2, \ldots, N \tag{2}$$

where pool_i indicates pooling operators with different kernel sizes. F_i^X and F_i^Y separately represent the features of modality1 and modality2 after pooling. While considering the information fusion, we integrate pooling features from heterogeneous modalities at the same resolution level and send them into a shared

operation. That is

$$S_i = \text{sharedOP}(\text{concat}([F_i^X, F_i^Y])),\ i = 1, 2, \ldots, N. \quad (3)$$

In actual implementation, to avoid causing more computation burden, the shared operation is only a simple convolution operator. And it is noteworthy that features at different levels all share the same operation. Then, the pixelwise addition is conducted between the integrated features and different modal features of the same level. Subsequently, we flatten these multiscale features and concatenate them together

$$O^X = \text{concat}([S_1 + F_1^X, S_2 + F_2^X, \ldots, S_N + F_N^X]_{\text{fl}}) \quad (4)$$

$$O^Y = \text{concat}([S_1 + F_1^Y, S_2 + F_2^Y, \ldots, S_N + F_N^Y]_{\text{fl}}) \quad (5)$$

where fl means flattening operation. O^X and O^Y are the output of the MSPF module.

With regard to the sequence length of X and O^X (here, we only take the case of one modality as an example), we formalize them as follows:

$$L(X_{\text{fl}}) = hw \quad (6)$$

$$L(O^X) = \frac{hw}{r_1^2} + \frac{hw}{r_2^2} + \cdots + \frac{hw}{r_N^2} \quad (7)$$

where h and w are the height and width of the features. r_1–r_N all stand for pooling ratios and in most practices, they are power series of integer 2. Therefore, we have

$$\begin{aligned} L(O^X) &= hw \cdot \left(\frac{1}{r_1^2} + \frac{1}{r_2^2} + \cdots + \frac{1}{r_N^2}\right) \\ &= hw \cdot \left(\frac{1}{(2^1)^2} + \frac{1}{(2^2)^2} + \frac{1}{(2^3)^2} + \cdots + \frac{1}{(2^N)^2}\right) \\ &= hw \cdot \left(\frac{1}{3}\left(1 - \frac{1}{4^N}\right)\right) \\ &< hw \\ &= L(X_{\text{fl}}). \end{aligned} \quad (8)$$

Obviously, compared with X_{fl} and Y_{fl}, the sequence length of O^X and O^Y is relatively short owing to the effect of pooling. More importantly, the intermodal heterogeneous information and the intramodal contextual information are both integrated in O^X and O^Y. Thereby, regarding the calculation of key-value vectors in SA of different modal branches, we replace the directly flattened X and Y with compressed O^X and O^Y, respectively. To be concrete, the three core elements, query Q, key K, and value V can be computed as follows:

$$Q^X = X_{\text{fl}} W_{XQ},\ K^X = O^X W_{XK},\ V^X = O^X W_{XV} \quad (9)$$

$$Q^Y = Y_{\text{fl}} W_{YQ},\ K^Y = O^Y W_{YK},\ V^Y = O^Y W_{YV} \quad (10)$$

where $W_{XQ}, W_{XK}, W_{XV}, W_{YQ}, W_{YK}, W_{YV}$ are all the linear projection matrices. Therefrom, FSA can be written as follows:

$$\text{FSA}(X) = \text{softmax}\left(\frac{Q^X \times (K^X)^T}{\sqrt{d_{\text{head}}}}\right) \times V^X \quad (11)$$

$$\text{FSA}(Y) = \text{softmax}\left(\frac{Q^Y \times (K^Y)^T}{\sqrt{d_{\text{head}}}}\right) \times V^Y \quad (12)$$

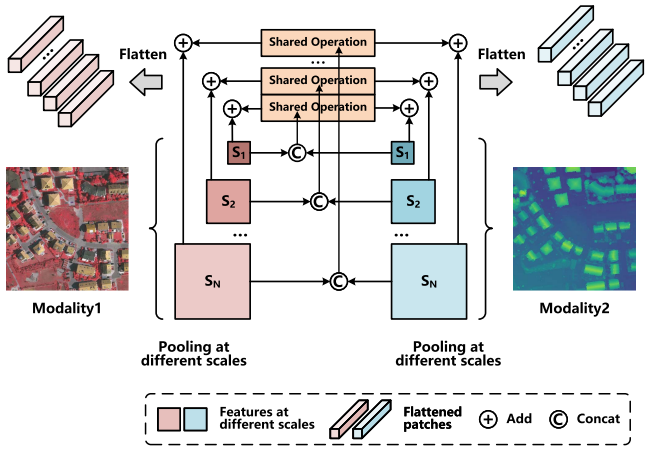

Fig. 3. Schematic diagram of MSPF. It takes 2-D image features, X and Y, as input, and outputs 1-D flattened vectors, O^X and O^Y. These 1-D vectors are then sent to an SA layer. Features and vectors of different branches are colored differently. Best viewed in color.

where the superscript T is the signal of matrix transposition.

C. Learnable Embedding Fusion

Aiming to utilize the complementary effect of different modal data in the form of 1-D image embeddings, we establish an LEF module at the output layer of each stage. Fig. 4 displays the structure of the proposed LEF in detail.

The LEF module consists of an embedding learning component and an embedding fusion component. For the embedding learning component, it takes the 1-D image embedding of modality1 and modality2 as input. These embeddings are first concatenated together. Then, we apply an average pooling operation to aggregate global information along the dimension of embedding quantity and obtain a channel-maintained aggregation embedding. This process can be formulated as follows:

$$\text{AE} = \text{avg\_pool}(\text{concat}([\text{IE}^X, \text{IE}^Y])) \quad (13)$$

where IE^X and IE^Y are the input embedding of modality1 and modality2, AE is the aggregation embedding. Considering that the embeddings of different modalities have differentiated contributions to the final prediction, we thus leverage an embedding learner following a softmax function to learn a kind of embedding mask

$$\text{EM} = \text{softmax}(\text{learner}_{\text{emb}}(\text{AE})) \quad (14)$$

where EM is the embedding mask. The $\text{learner}_{\text{emb}}$ in our subsequent implementation is a two-layer MLP with a nonlinear mapping. In (14), the tensor slicing operation before softmax is omitted for simplicity. The generated mask is able to adaptively adjust the importance of different modal embeddings in modality merging from the perspective of channel.

In the embedding fusion component, the embedding mask acts as a kind of weight coefficient. It multiplies with the embeddings of two modal branches. Next, we add the two adjusted branches together and feed the branch sum into a linear mapping unit to

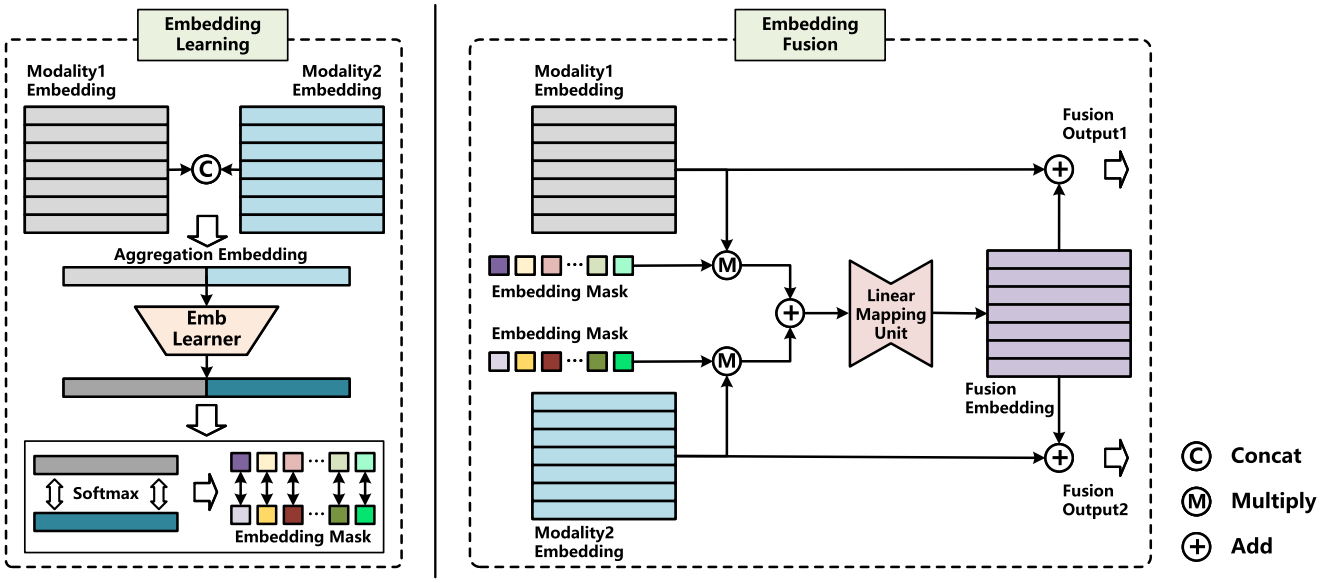

Fig. 4. Structure of LEF. **Emb learner** means embedding learner. Rectangles all denote 1-D embeddings. The color of the rectangle indicates which modality it belongs to. Best viewed in color.

obtain a fusion embedding FE

$$\mathrm{FE} = \mathrm{LMU}(\mathrm{EM}^X \cdot \mathrm{IE}^X + \mathrm{EM}^Y \cdot \mathrm{IE}^Y) \tag{15}$$

where $\cdot$ stands for tensor multiplication with broadcast mechanism. The linear mapping unit LMU in our experiments consists of a linear layer with channel reduction, an activation function GELU() and a linear layer with channel expansion. After that, we assign the fusion information to modality1 and modality2 through elementwise addition

$$\mathrm{OE}^X = \mathrm{LN}(E^X + \mathrm{FE}) \tag{16}$$

$$\mathrm{OE}^Y = \mathrm{LN}(E^Y + \mathrm{FE}) \tag{17}$$

where OE^X and OE^Y are the outputs of the LEF module, LN is layer normalization [64]. This shortcut connection not only preserves intramodal characteristics, but also incorporates intermodal commonalities.

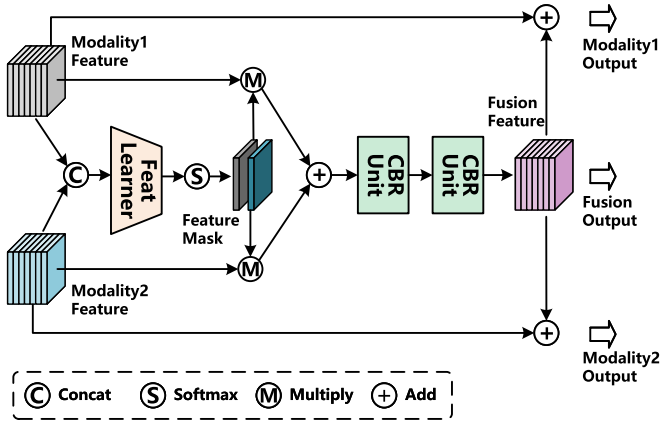

Fig. 5. Illustration of LFF. **Feat learner** means feature learner. In LFF, 2-D features of different modalities are represented by the cuboids of different colors. Best viewed in color.

D. Learnable Feature Fusion

LEF mainly focuses on learnable fusion for 1-D embeddings. Naturally, we further build another learnable fusion approach, LFF, which realizes information collaboration in the form of 2-D image features. The fusion process is depicted in Fig. 5.

In LFF, we first concatenate the upsamling reconstruction features of modality1 IF^X and modality2 IF^Y. Then, we input the combined features into a feature learner and obtain a mask that describes spatial importance in information fusion through a softmax function

$$\mathrm{FM} = \mathrm{softmax}(\mathrm{learner}_{\mathrm{feat}}(\mathrm{concat}([\mathrm{IF}^X, \mathrm{IF}^Y]))) \tag{18}$$

where FM is the feature mask. The feature learner $\mathrm{learner}_{\mathrm{feat}}$ in general can be any operators that maintains the image feature size. In our LFF structure, it is a pointwise (PW) convolution. The feature mask can be regarded as a gating factor that aggregates information from two modalities. Different from LEF, we set two consecutive convolution-BN-ReLU (CBR) units in the LFF module to capture a reconstructed fusion feature FF

$$\mathrm{FF} = \mathrm{CBR}(\mathrm{CBR}(\mathrm{FM}^X \cdot \mathrm{IF}^X + \mathrm{FM}^Y \cdot \mathrm{IF}^Y)) \tag{19}$$

where CBR is the initial letter of convolution, batch normalization [65], and ReLU. The former CBR unit is based on 3×3 DW convolution while the latter is PW convolution. Next, following a similar thought of LEF, we still assign FF to modality1 and modality2 through pixelwise addition, i.e.,

$$\mathrm{OF}^X = F^X + \mathrm{FF} \tag{20}$$

$$\mathrm{OF}^Y = F^Y + \mathrm{FF}. \tag{21}$$

Yet, the final output of the LFF also includes FF, rather than just OF^X and OF^Y. And FF participates in generating the final classification probability as well.

IV. Experiments

In this section, we perform several experiments on two extensively used remote sensing semantic segmentation datasets to confirm the effectiveness and generalization of the proposed LMF-Net. Experimental settings and corresponding results are also explicitly elucidated.

A. Datasets

We evaluate the proposed multimodal fusion method on two extensively used benchmark datasets, i.e., the International Society for Photogrammetry and Remote Sensing (ISPRS) Vaihingen [21] and Potsdam [21] dataset. A brief introduction of these two datasets is listed as follows.

1) Vaihingen: The Vaihingen[1] dataset published by ISPRS was used as the official dataset of the ISPRS 2-D Semantic Labeling Contest. It was captured over the small village of Vaihingen in Germany. This area is characterized by many detached buildings and small multistory buildings. This dataset is partitioned into 33 large-scale true orthophoto (TOP) tiles with a spatial resolution of 9cm. The average size of these TOP tiles is 2064 $\times$ 2494 pixels. Among them, 11 tiles are used for training purpose, 5 other tiles are taken as the validation set, and the remaining 17 tiles are utilized to evaluate the model's behavior. In Vaihingen, each tile is formatted as a 8-bit TIF file which is composed of three types of bands corresponding to near infrared, red and green wavelengths. The DSM defined on the same grid as TOP tile is encoded as a 32-bit single channel gray-scale TIF file, and its value indicates the DSM height. All pixels in this dataset are categorized with respect to six predefined categories, represented by *impervious surfaces* (white), *building* (blue), *low vegetation* (cyan), *tree* (green), *car* (yellow), and *clutter/background* (red). Several TOP tile samples, corresponding DSMs and corresponding annotations in the Vaihingen dataset are shown in Fig. 6. The proportion of different categories in the Vaihingen dataset is described in Fig. 8(a). This demonstrates that the occurrence of these classes is highly imbalanced, making the dataset very challenging.

2) Potsdam: Similar to the Vaihingen dataset, the Potsdam[2] dataset was also one of the ISPRS contest datasets. While, this dataset is relatively larger in volume. It consists of 38 TOP tiles with a ground sample distance of 5 cm. The width and height of each tile are both 6000 pixels. The study area in this dataset is characterized by a typical historic city with narrow streets, large building blocks and dense settlement structures. Fig. 7 displays some samples of TOP tile, DSM and ground truth of the same area. Concretely, each TOP tile and DSM are defined on the same grid and they are also formatted as TIF files. Besides, this datasest has been annotated manually into six most common land cover categories as well, which is consistent with the Vaihingen dataset, i.e., *impervious surfaces* (white), *building* (blue), *low vegetation* (cyan), *tree* (green), *car* (yellow), and *clutter/background* (red). Nevertheless, there are four types of bands for each tile in total in the Potsdam dataset, namely, near infrared, red, green, and blue wavelengths. As regards the partition of the Potsdam dataset, there are 18 tiles, six tiles, and 14 tiles in training part, validation part, and test part, respectively. The pixel distribution of different categories in this dataset is shown in Fig. 8(b). The number of pixels of category *car* only accounts for 1.69% of the total dataset, indicating that there is also a problem of relatively imbalanced distribution of categories as well.

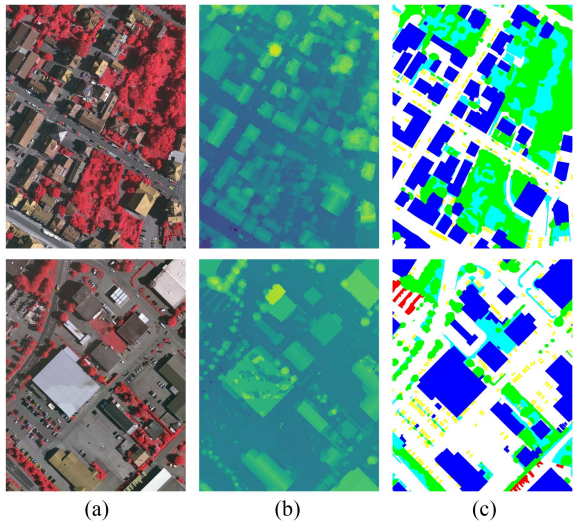

Fig. 6. Several example tiles, corresponding DSMs and corresponding annotations in the Vaihingen dataset. Legend—White: Impervious surfaces; Blue: Buildings; Cyan: Low vegetation; Green: Trees; Yellow: Cars; Red: Clutter/background. (a) IRRG. (b) DSM. (c) GT.

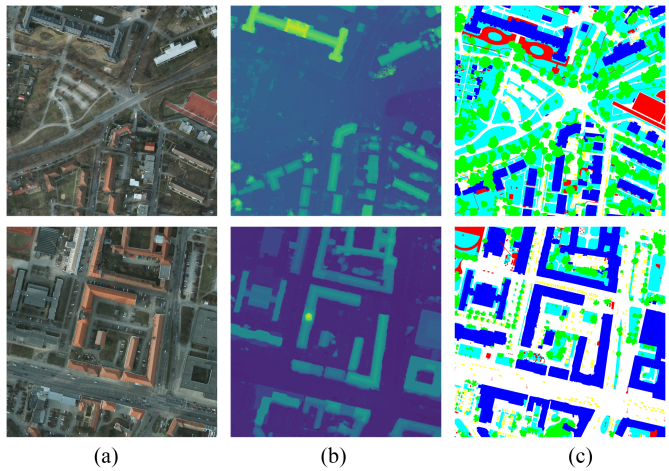

Fig. 7. Several example tiles, corresponding DSMs and corresponding annotations in the Potsdam dataset. Legend—White: Impervious surfaces; Blue: Buildings; Cyan: Low vegetation; Green: Trees; Yellow: Cars; Red: Clutter/background. (a) RGB. (b) DSM. (c) GT.

[1][Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx

[2][Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx

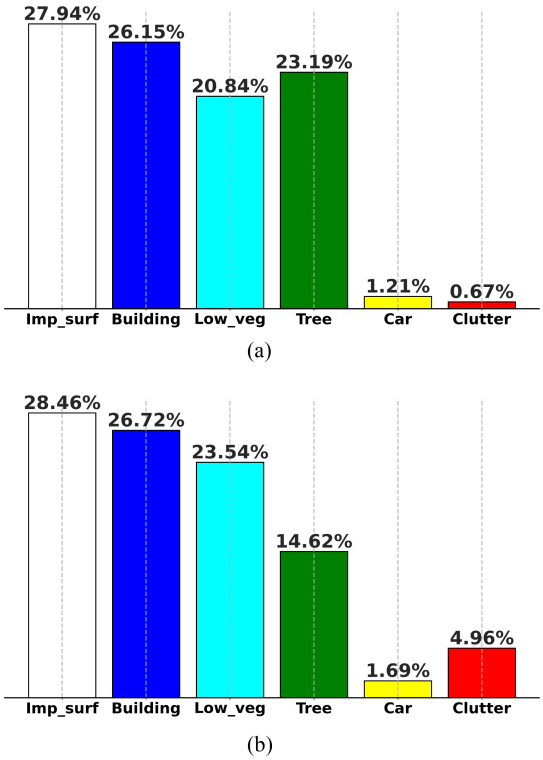

Fig. 8. Percentage of pixels for different categories in the Vaihingen and the Potsdam dataset. (a) Vaihingen. (b) Potsdam.

B. Evaluation Metrics

Similar to most work on multimodal semantic segmentation task, we adopt three extensively-used evaluation metrics represented by overall accuracy (OA), mean intersection-over-union (mIoU), and mean F_1-score (mF_1) to evaluate the performance of different approaches.

OA can be written as follows:

$$\text{OA} = \frac{N_{\text{correct}}}{N_{\text{all}}} \tag{22}$$

where N_{correct} and N_{all} separately characterize the quantity of properly discriminated pixels and the total number of pixels. Normally, a higher OA means a better segmentation behavior and vice versa.

For the purpose of calculating mF_1 indicator, we first concentrate on the performance of a certain category $\mathcal{C}$. True positives, abbreviated as TP, indicates that the pixels should belong to category $\mathcal{C}$ and are also correctly identified. False negatives (FN) means that the pixels are labeled as $\mathcal{C}$, but are recognized as any of the other categories. As for the false positives (FP), its definition is that the annotations of those pixels fall into other categories and yet they are classified as category $\mathcal{C}$. On the basis of the acquired TP, FN, and FP, IoU value of a certain category can be determined by the formula

$$\text{IoU} = \frac{\text{TP}}{\text{TP+FP+FN}}. \tag{23}$$

Furthermore, we can calculate precision and recall per category as

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{24}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}. \tag{25}$$

Then, the F_1-score of category $\mathcal{C}$ can be obtained through

$$F_\beta = (1+\beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision+Recall}}, \quad \beta = 1. \tag{26}$$

Ultimately, by averaging IoU and F_1-score across all of the categories, we have

$$\text{mIoU} = \frac{1}{C_N} \sum_{i=1}^{C_N} \text{IoU}_i \tag{27}$$

$$\text{m}F_1 = \frac{1}{C_N} \sum_{i=1}^{C_N} F_{1_i} \tag{28}$$

where C_N expresses the number of categories. Similar to OA, a higher mIoU or mF_1 implies that the network can probably achieve a more satisfactory performance.

In addition, apropos of the computation cost, we utilize two indicators to describe the behavior of different networks, i.e., the quantity of optimizable parameters and floating point operations (FLOPs).

Parameters of linear layers are related to the input and output neurons

$$\text{Param(Linear)} = (L_{\text{in}} + 1) * L_{\text{out}} \tag{29}$$

where L_{in} and L_{out} represent the number of input neurons and output neurons, respectively. While parameters of convolution layers can be described as

$$\text{Param(Conv)} = (K_W * K_H * C_{\text{in}} + 1) * C_{\text{out}} \tag{30}$$

where K_W and K_H separately express the width and height of convolution kernel. C_{in} and C_{out} denote input channels and output channels, respectively. In particular, for PW convolution, K_W and K_H are both 1. Nonetheless, for DW convolution, due to the situation that a convolutional kernel is only responsible for the operation of one channel, the formula for parameter quantity becomes

$$\text{Param(DW Conv)} = K_W * K_H * C_{\text{in}} + C_{\text{out}}. \tag{31}$$

It is worth noting that the computation process for parameters aforementioned all involves bias.

Accordingly, we can obtain the FLOPs for linear layers and convolution layers

$$\text{FLOPs(Linear)} = 2L_{\text{in}} * L_{\text{out}} \tag{32}$$

$$\text{FLOPs(Conv)} = 2C_{\text{in}} * K_W * K_H * W_{\text{out}} * H_{\text{out}} * C_{\text{out}} \tag{33}$$

$$\text{FLOPs(DW Conv)} = 2C_{\text{in}} * K_W * K_H * W_{\text{out}} * H_{\text{out}}. \tag{34}$$

In a similar vein, the computation for FLOPs value all contains bias as well.

TABLE I
QUANTITATIVE RESULTS ACHIEVED FOR THE VAIHINGEN DATASET

Method	F_1(%)					OA(%)	mF_1(%)	mIoU(%)
	Imp_surf	Building	Low_veg	Tree	Car			
DSMFNet [66]	90.25	91.59	78.39	87.49	77.18	-	84.98	74.36
TreeUNet [67]	92.50	94.90	83.60	89.60	85.90	90.40	89.30	80.92
SA-GATE [68]	92.95	96.05	84.45	89.91	87.34	91.06	90.14	82.30
CF-Net [69]	91.40	95.10	80.30	88.80	89.10	89.30	88.94	80.42
LA-Net [70]	92.40	94.90	82.90	88.90	81.30	89.80	88.08	79.09
C3Net [71]	93.00	96.10	**85.40**	90.30	85.40	**91.30**	90.04	82.15
CaFE [18]	91.19	93.89	79.10	89.10	79.04	-	86.46	76.68
CMFNet [58]	92.37	94.94	83.34	89.03	83.04	90.05	88.54	79.77
CIMFNet [72]	92.13	96.07	80.26	88.26	**90.91**	-	89.52	81.44
MFNet [73]	92.75	95.61	84.21	89.26	88.82	90.63	90.13	82.26
CEGFNet [57]	92.27	95.91	81.39	88.59	87.89	-	89.21	80.86
TokenFusion [49]	92.85	**96.27**	85.18	90.27	85.08	91.25	89.93	81.99
CEN [74]	92.83	95.99	85.37	**90.35**	80.16	91.13	88.94	80.53
PACSCNet [17]	92.92	96.11	84.08	89.61	89.11	90.92	90.37	82.67
CMX [50]	93.07	95.98	84.69	90.16	87.45	91.15	90.27	82.50
FTransUNet [20]	92.69	96.23	84.14	89.74	84.81	90.89	89.52	81.35
LMF-Net	**93.08**	96.12	84.71	90.02	89.41	91.19	**90.67**	**83.15**

The first column lists the name of different multi-modal semantic segmentation approaches. The next five columns show the $F1$ value for each category in Vaihingen. And three kinds of overall performance indicators are recorded in the remaining columns. Bold font is the maximum value in its column.

C. Implementation Details

We train the proposed multimodal fusion framework LMF-Net in an end-to-end manner. Owing to the GPU memory limitation, we crop the original large-size remote sensing image through sliding window technique. The resolution of every input slice is set to 512×512 pixels. One branch's input modality is optical image and the other branch is raw 32-bit floating point DSM image in all experiments. Aiming to deal with the problem of class-imbalanced distribution, we adopt weighted cross entropy loss to characterize the gap between the network output and the ground truth. The losses of different branches are first calculated separately, and then averaged. And they are equally important. During the training procedure, pretrained weights on ImageNet [75] are loaded and AdamW [76] is employed to update the gradients of model parameters. The batch size is set to 6. In addition, the learning rate value is initialized to 2×10^{-4}, and then it is declined according to a "poly" learning rate schedule with warmup strategy [77], where the power factor is set to 0.9. The decay coefficient and drop path rate are separately set to 0.05 and 0.1 to prevent overfitting phenomenon.

D. Experimental Results

1) Performance on Vaihingen: The segmentation results of different neural networks achieved for the Vaihingen dataset are provided in Table I. Here, all of the models are trained on multimodal remote sensing data, namely, paired optical and DSM images. From these statistical records, we can see that the proposed LMF-Net has relatively satisfactory behavior on each category. The obtained F_1 scores are superior to the majority of approaches. Besides, from the perspective of comprehensive capability, LMF-Net is only slightly inferior to C3Net [71] in OA, yet the mF_1 and mIoU indicator outperform all comparison methods, reaching 90.67% and 83.15%, respectively. This fully demonstrates that our fusion framework is available and effective. It is able to learn more distinguishable expressions by properly leveraging multimodal remote sensing data.

In order to intuitively assess the segmentation effect, we further give the qualitative results in Fig. 9. Here, we select several representative and high-performance methods covering natural scene segmentation and remote sensing scene segmentation for comparison. It can be observed that the visualizations produced by LMF-Net are relatively close to the semantic annotations. In row1, row3, and row4, each *car* instance is independently distinguished through LMF-Net and there are quite few pixels which are misidentified from *low vegetation* to *tree*. Moreover, as displayed in row2 and row3, LMF-Net generates more smooth boundaries at the edges of *building* and *impervious surfaces*, as well as *low vegetation* and *impervious surfaces*, compared with other models.

2) Performance on Potsdam: We next evaluate our LMF-Net and other extraordinary methods on the Potsdam dataset. The comparison results are recorded in Table II. It can be found that our proposed LMF-Net also possesses powerful competitiveness on this larger and more challenging dataset. The OA, mF_1, and mIoU achieve 91.40%, 92.84%, and 86.83%, respectively. These three indicators are all better than the comparison models listed in this table. While, regarding the categories with relatively few samples and scattered distribution, for instance, *low vegetation* and *car*, our proposed LMF-Net obtains more prominent behavior and realizes the highest F_1 score, separately reaching 88.33% and 96.16%. This suggests that the LMF-Net is able to capture more generic expressions from multimodal remote sensing data and to generalize to different remote sensing scenes.

Similarly, the visualization results of these representative networks on Potsdam are also presented in Fig. 10. We can notice that the LMF-Net achieves more accurate predictions and relatively sound effects for detail information. In row1, LMF-Net obtains a better recognition result for the scene where vegetation is distributed on both sides of a narrow road. As shown

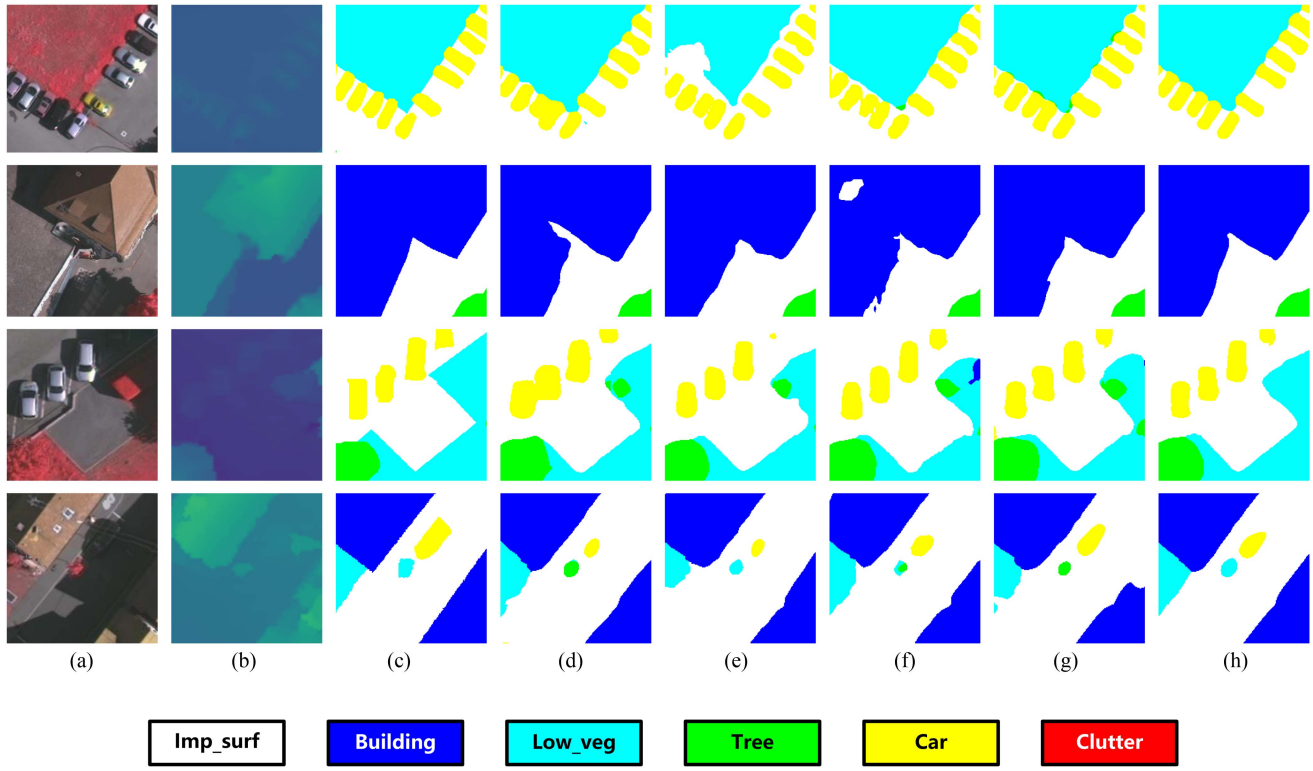

Fig. 9. Visualization of results achieved for the Vaihingen dataset. The label legend of each category can be referenced to the bottom of this figure. (a) Optical images. (b) DSM images. (c) Ground truth. (d) Results of SA-GATE [68]. (e) Results of TokenFusion [49]. (f) Results of PACSCNet [17]. (g) Results of FTransUNet [20]. (h) Results of the proposed LMF-Net. Best viewed in color.

TABLE II
QUANTITATIVE RESULTS ACHIEVED FOR THE POTSDAM DATASET

Method	$F_1(\%)$					OA(%)	mF_1(%)	mIoU(%)
	Imp_surf	Building	Low_veg	Tree	Car			
V-FuseNet [78]	90.91	93.23	84.30	86.23	90.56	-	89.05	80.41
DSMFNet [66]	91.64	92.24	84.66	86.80	89.02	-	88.87	80.09
TreeUNet [67]	93.10	**97.30**	86.80	87.10	94.10	90.60	91.68	84.90
SA-GATE [68]	93.41	97.14	87.31	88.50	95.47	91.10	92.36	86.05
CaFE [18]	91.23	95.01	85.12	**89.83**	92.17	-	90.67	83.10
CMFNet [58]	91.19	93.91	86.14	86.25	95.15	88.48	90.53	82.91
CIMFNet [72]	90.58	94.95	84.97	84.72	94.25	-	89.89	81.93
MFNet [73]	93.40	96.85	87.18	88.78	96.11	91.02	92.46	86.22
CEGFNet [57]	90.61	94.61	85.12	85.20	94.88	-	90.08	82.23
TokenFusion [49]	**93.56**	97.15	87.37	88.67	95.87	91.11	92.52	86.33
CEN [74]	92.03	96.71	86.70	88.35	94.52	90.12	91.66	84.82
MFTransNet [79]	90.87	95.79	82.55	83.90	90.16	87.93	88.65	79.96
CMX [50]	92.86	96.36	88.16	89.17	95.88	90.94	92.48	86.20
PACSCNet [17]	92.89	96.26	87.13	87.63	95.48	90.43	91.88	85.21
FTransUNet [20]	92.91	96.16	87.64	88.93	95.69	90.65	92.27	85.83
LMF-Net	93.36	96.99	**88.33**	89.35	**96.16**	**91.40**	**92.84**	**86.83**

The first column lists the name of different multi-modal semantic segmentation approaches. The next five columns show the $F1$ value for each category in potsdam. And three kinds of overall performance indicators are recorded in the remaining columns. Bold font is the maximum value in its column.

in this row, the strip-shaped *impervious surfaces* extracted by our method is more complete and does not occur interruption phenomenon. Likewise, the *low vegetation* with similar shape is effectively classified in row2 and row3 as well. In particular, as displayed in row4, for object with special structure, such as concave and cavity, LMF-Net still clearly identifies the outline of this *building*.

3) Complexity Analysis: Apart from the classification capability, we continue to have a comprehensive assessment about the computation complexity of different methods. The comparison results between LMF-Net and other well-performing models are listed in Table III. As a note, FLOPs indicator is measured by a 512×512 image slice and the mF_1 is obtained on the Vaihingen dataset. It can be found that the proposed LMF-Net

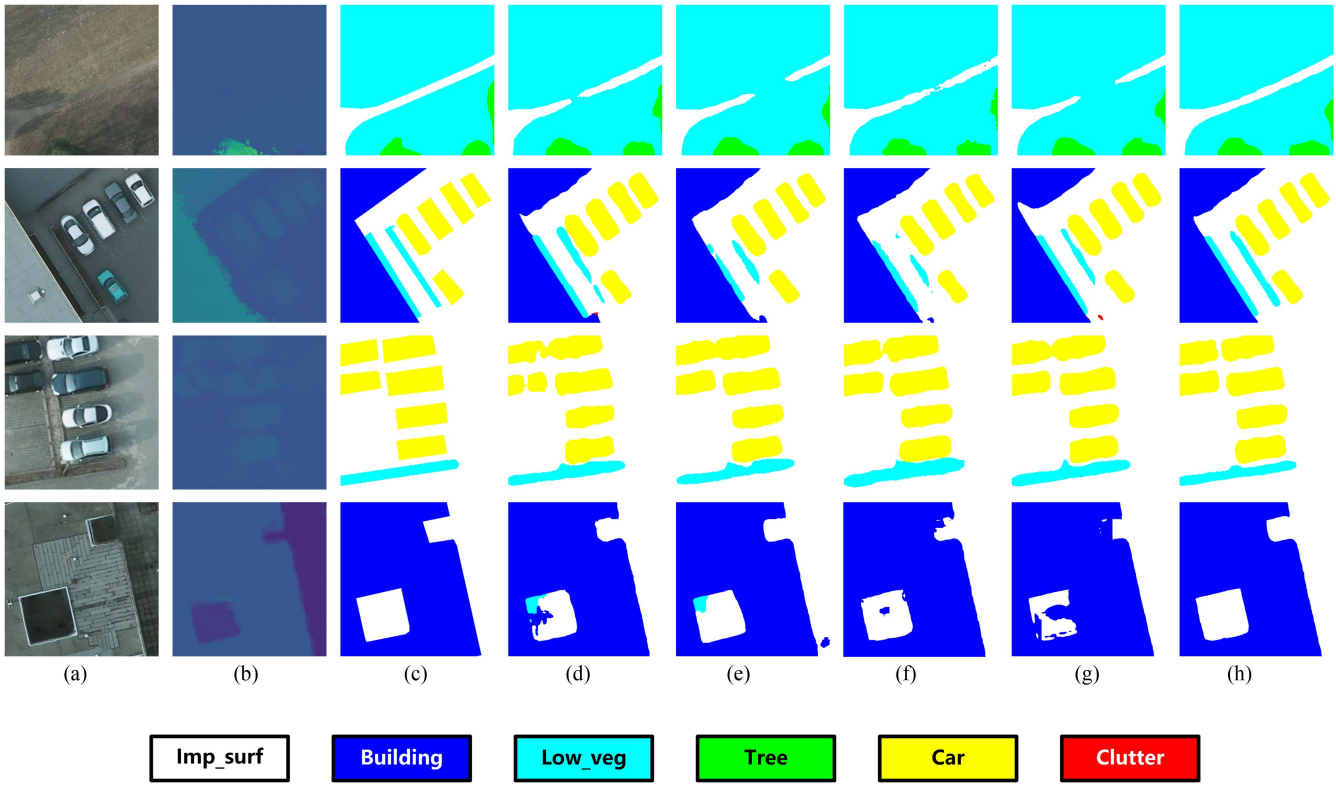

Fig. 10. Visualization of results achieved for the Potsdam dataset. The label legend of each category can be referenced to the bottom of this figure. (a) Optical images. (b) DSM images. (c) Ground truth. (d) Results of SA-GATE [68]. (e) Results of TokenFusion [49]. (f) Results of PACSCNet [17]. (g) Results of FTransUNet [20]. (h) Results of the proposed LMF-Net. Best viewed in color.

TABLE III
EVALUATION OF THE LMF-NET AND SEVERAL ADVANCED APPROACHES IN TERMS OF COMPUTATION COMPLEXITY

Method	Params(M)	FLOPs(G)	mF_1(%)
DSMFNet [66]	50.92	518.00	84.98
SA-GATE [68]	110.85	164.88	90.14
CMFNet [58]	106.29	312.70	88.54
CIMFNet [72]	71.41	118.80	89.52
MFNet [73]	94.67	85.72	90.13
TokenFusion [49]	45.91	80.04	89.93
CEN [74]	118.12	526.49	88.94
PACSCNet [17]	94.05	150.26	90.37
CMX [50]	66.56	**57.07**	90.27
FTransUNet [20]	203.99	264.61	89.52
LMF-Net	**42.90**	71.42	**90.67**

FLOPs is measured by a 512 × 512 image slice. Bold font is the best value in its column.

has the minimum number of network parameters and the second least FLOPs among all comparison approaches. Furthermore, LMF-Net is able to achieve the optimal mF_1 score even in case where the parameters and FLOPs are less than the second best method by exceeding 50M and 70G, respectively. The above-mentioned analysis suggests that our LMF-Net has relatively strong competitiveness in both identification capability and model complexity.

TABLE IV
ABLATION STUDIES REGARDING THE SEGMENTATION RESULTS AND COMPUTATION OVERHEAD OF DIFFERENT MODULES IN THE PRESENTED LMF-NET ON THE VAIHINGEN AND POTSDAM DATASET

Module	mIoU$_V$(%)	mIoU$_P$(%)	Params(M)	FLOPs(G)
baseline	80.54	84.30	44.68	72.77
+MSPF	82.22	85.86	40.90	69.41
+LEF	82.54	86.08	46.61	73.08
+LFF	82.53	85.95	44.75	74.47
+LEF&LFF	83.20	87.09	46.68	74.78
LMF-Net	83.15	86.83	42.90	71.42

V. DISCUSSION

In this section, we have a detailed discussion about our LMF-Net and analyze its efficacy. Section V-A provides the ablation studies to verify the performance of each module. Section V-B compares three different fusion schemes. Then, the design choice of pooling operator types is elucidated in Section V-C. Finally, we show some heatmap samples in Section V-D.

A. Ablation Studies About Different Modules

Table IV reports the ablation studies on three critical modules, i.e., MSPF, LEF, and LFF, on the Vaihingen and Potsdam dataset. mIoU$_V$ and mIoU$_P$ separately represent the mIoU performance on the Vaihingen and Potsdam dataset. The baseline here consists

TABLE V
EFFECT OF THREE FUSION SCHEMES IN DIFFERENT MODULES ON THE VAIHINGEN DATASET

Module	Scheme	mIoU(%)	mF_1(%)
baseline	-	80.54	89.00
MSPF	I	80.76	89.13
MSPF	F	82.06	89.99
MSPF	A	82.22	90.09
LEF	I	79.16	88.04
LEF	F	80.97	89.28
LEF	A	82.54	90.28
LFF	I	80.70	89.07
LFF	F	82.41	90.21
LFF	A	82.53	90.27

of a two-stream SegFormer [34] models, but there is no information interaction between the two streams. Then, we successively plug three modules into the baseline approach to validate their capability.

As is shown in Table IV, the proposed MSPF, LEF, and LFF all contribute to significant benefits on both two datasets, achieving an increment of over 1.5% in mIoU. The combination of the two learnable fusion schemes is still able to further improve the performance, realizing the mIoU indicator of up to 83.20% and 87.09% on these two datasets, respectively. Nonetheless, it should not be neglected that the computation complexity of the LEF and the LFF is relatively high. Compared to the baseline model, the combination of them leads to an increase of about 2 M Parameters and 2 G FLOPs as well. As a result, we introduce the MSPF to appropriately reduce the computation cost. It can be seen that Parameters and FLOPs decrease to 42.90 M and 71.42 G, respectively. They are even lower than the baseline model. This indicates that the MSPF makes the whole pipeline more compact. Although the introduction of pooling causes some information loss, there is only a negligible decrease in performance. The proposed model still shows competent segmentation behavior on both the Vaihingen and Potsdam dataset. Thus, the above-mentioned ablation studies prove that the union of three carefully designed modules ensures the superiority of the LMF-Net in terms of segmentation performance and computation efficiency.

B. Comparison of Three Fusion Schemes

Then, we assess the effectiveness of three fusion schemes in MSPF, LEF, and LFF. The results achieved for the Vaihingen dataset are fully recorded in Table V. Here, I means that the optical branch and the DSM branch are independent, and there is no information exchange in each proposed module. For F, it represents that we only fuse optical branch and DSM branch, yet the fused embeddings/features are not assigned to each branch any more. The meaning of A is that we not only combine the embeddings/features of optical and DSM modality, but also assign the fusion results to these two branches.

From Table V, we can find that in contrast to the baseline, scheme I does not bringing about noticeable improvement. For the LEF, the value of mIoU and mF_1 even decreases in a certain extent when adopting scheme I. This implies that unreasonable utilization of multimodal data probably gives rise to performance degradation. Nevertheless, modality fusion schemes, namely, F and A, have more superior effects for the MSPF and the LFF. There is an obvious increment in mIoU and mF_1 indicator. Even for the LEF, the results of scheme F also surpass those of the baseline and scheme I, which suggests that modality fusion is more conducive to fully develop the complementary advantage of optical and DSM embeddings/features. In addition, compared to scheme F, the performance is further enhanced by applying scheme A, especially for the LEF. It demonstrates that taking into account both complementary information and specificity information of different modal data is beneficial for the neural network to learn more discriminative representations. For this reason, based on the results of this experiment, we select scheme A in designing the proposed LMF-Net architecture.

TABLE VI
COMPARISON OF LEVERAGING TWO POOLING OPERATORS ON THE VAIHINGEN DATASET

Pooling Type	mIoU(%)	mF_1(%)
$MaxPool$	82.04	89.99
$AvgPool$	82.22	90.09

C. Design Choice of Pooling Operator

Generally, there are two widely utilized pooling operators, i.e., MaxPool and AvgPool. We individually employ these two pooling operators in the MSPF module and observe their effects. The results achieved for the Vaihingen dataset are documented in Table VI. It can be noticed that both of these pooling operators perform well in the MSPF while AvgPool is slightly ahead of MaxPool, reaching 82.22% and 90.09% in the matter of mIoU and mF_1 metric, respectively. One possible reason could be that the extracted features are comparatively smooth through taking pixel average. Therefore, when compared with other excellent models, the pooling operators are all set to AvgPool in the LMF-Net.

D. Visualization of Heatmaps

A heatmap typically highlights the region of interest. It is able to reveal the activation status of intermediate feature maps in a model. In order to have a better explanation for the effectiveness of the proposed method, we provide several heatmaps of the baseline model and LMF-Net in Fig. 11. The colorbar legend about these heatmaps is jet. Note that regions with high temperatures indicate stronger feature responses and vice versa.

As shown in Fig. 11, saliency regions present how the baseline model and LMF-Net determine which category a pixel belongs to. Compared with the baseline model, the contour of feature responses of the LMF-Net is closer to the ground truth and the corresponding activated values are higher. Besides, our method can extract more accurate local details about some small objects, e.g., cars. These phenomena mean that our method has certain advantages in facilitating the acquisition of discriminative semantic features. It also contributes to a better prediction effect.

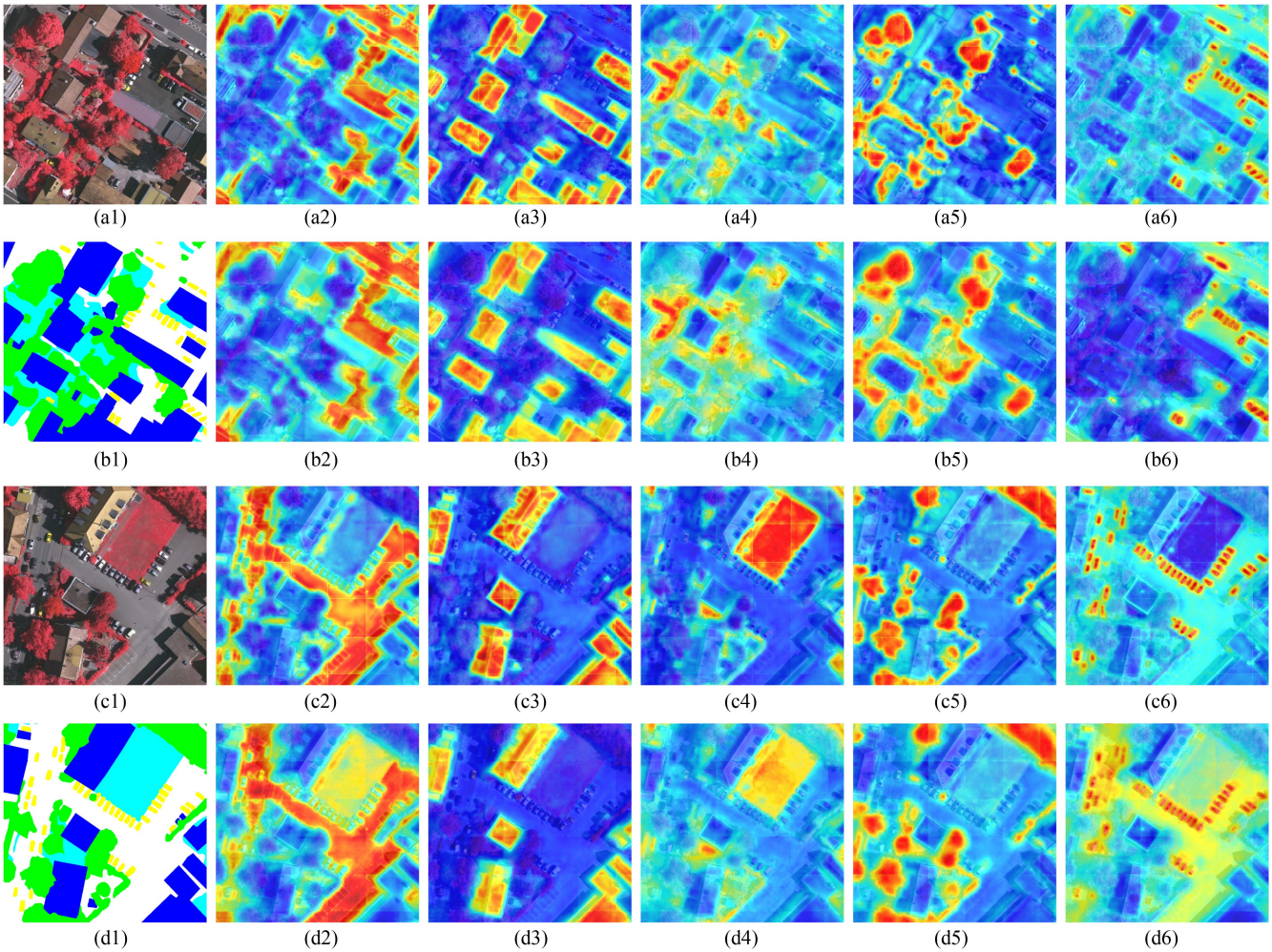

Fig. 11. Visualization of heatmaps. (a1) is the optical image. (b1) represents the ground truth of the same area as (a1). (a2)∼(a6) indicate the heatmaps generated by the proposed LMF-Net on the category of *impervious surfaces*, *building*, *low vegetation*, *tree*, and *car* while (b2)∼(b6) show the heatmaps obtained by the baseline model for these five categories. (c) and (d) are presented in the same form. Best viewed in color.

VI. Conclusion

In this article, we propose an LMF-Net for remote sensing semantic segmentation. To be specific, we introduce an MSPF module by leveraging pooling operator. MSPF can generate key-value pairs with multimodal complementary information in a parameter-free mean. Then, to further harness the cross-modal collaboration embeddings/features in heterogeneous data fusion, we elaborate two learnable fusion modules, namely, LEF and LFF. To objectively and comprehensively validate the performance of our LMF-Net, we perform sufficient experiments on two challenging public datasets, ISPRS Vaihingen and ISPRS Potsdam, and compare the LMF-Net with plenty of remarkable models. The achieved results fully demonstrate that the proposed LMF-Net has powerful competitiveness in both segmentation performance and computation efficiency. Besides, the effect of MSPF, LEF, and LFF is deeply discussed as well. Concerning future research, we will still devote ourselves to the research of multimodal remote sensing data fusion and will attempt to adopt some semisupervised or unsupervised learning strategies to deal with the challenge of insufficient annotations.

Acknowledgment

The authors deeply appreciate the effort of the anonymous reviewers and would like to thank them for their valuable comments.

References

[1] X. Sun et al., "Revealing influencing factors on global waste distribution via deep-learning based dumpsite detection from satellite imagery," *Nature Commun.*, vol. 14, no. 1, 2023, Art. no. 1444.

[2] D. Xu et al., "Quantization of the coupling mechanism between ECO-environmental quality and urbanization from multisource remote sensing data," *J. Cleaner Prod.*, vol. 321, 2021, Art. no. 128948.

[3] X. Sun et al., "Fair1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 184, pp. 116–130, 2022.

[4] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 296–307, 2020.

[5] Z. Zhu, J. Kang, W. Diao, Y. Feng, J. Li, and J. Ni, "SIRS: Multi-task joint learning for remote sensing foreground-entity image-text retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5625615.

[6] C. Li et al., "Injecting linguistic into visual backbone: Query-aware multimodal fusion network for remote sensing visual grounding," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5637814.

[7] H. Wu, P. Huang, M. Zhang, W. Tang, and X. Yu, "CMTFNet: CNN and multiscale transformer fusion network for remote-sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 2004612.

[8] W. Wu, Z. Shao, X. Huang, J. Teng, S. Guo, and D. Li, "Quantifying the sensitivity of SAR and optical images three-level fusions in land cover classification to registration errors," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, 2022, Art. no. 102868.

[9] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

[10] F. Yuan, K. Li, C. Wang, and Z. Fang, "A lightweight network for smoke semantic segmentation," *Pattern Recognit.*, vol. 137, 2023, Art. no. 109289.

[11] F. Yuan, L. Zhang, X. Xia, Q. Huang, and X. Li, "A gated recurrent network with dual classification assistance for smoke semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 4409–4422, 2021.

[12] H.-F. Zhong, Q. Sun, H.-M. Sun, and R.-S. Jia, "NT-Net: A semantic segmentation network for extracting lake water bodies from optical remote sensing images based on transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5627513.

[13] Y. Li, B. Dang, Y. Zhang, and Z. Du, "Water body classification from high-resolution optical remote sensing imagery: Achievements and perspectives," *ISPRS J. Photogrammetry Remote Sens.*, vol. 187, pp. 306–327, 2022.

[14] Z. Zhu, B. Dong, Q. Bu, and J. Ni, "A parameter-efficient differentiable active contour network for precisely building instance segmentation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2024, pp. 8355–8359.

[15] S. Dong, W. Zhou, C. Xu, and W. Yan, "EGFNet: Edge-aware guidance fusion network for RGB–thermal urban scene parsing," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 1, pp. 657–669, Jan. 2024.

[16] W. Zhou, S. Dong, M. Fang, and L. Yu, "CACFNet: Cross-modal attention cascaded fusion network for RGB-T urban scene parsing," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 1919–1929, Jan. 2024.

[17] X. Fan, W. Zhou, X. Qian, and W. Yan, "Progressive adjacent-layer coordination symmetric cascade network for semantic segmentation of multimodal remote sensing images," *Expert Syst. Appl.*, vol. 238, 2024, Art. no. 121999.

[18] A. Zheng, J. He, M. Wang, C. Li, and B. Luo, "Category-wise fusion and enhancement learning for multimodal remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4416212.

[19] S. Xiao et al., "MOCG: Modality characteristics-guided semantic segmentation in multimodal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5625818.

[20] X. Ma, X. Zhang, M.-O. Pun, and M. Liu, "A multilevel multimodal fusion transformer for remote sensing semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5403215.

[21] J. Niemeyer, F. Rottensteiner, and U. Soergel, "Contextual classification of lidar data and building object detection in urban areas," *ISPRS J. Photogrammetry Remote Sens.*, vol. 87, pp. 152–165, 2014.

[22] 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.

[23] L.-C. 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.

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

[25] 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. 6230–6239.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.* Munich, Germany, Oct. 2015, pp. 234–241.

[27] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet: A nested U-Net architecture for medical image segmentation," in *Proc. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support: 4th Int. Workshop, DLMIA, 8th Int. Workshop, ML-CDS, Held Conjunction MICCAI*, Granada, Spain, Sep. 2018, pp. 3–11.

[28] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Bi-directional ConvLSTM U-Net with densley connected convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 406–415.

[29] Y. Li et al., "Learning dynamic routing for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8550–8559.

[30] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6877–6886.

[31] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 6000–6010.

[32] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Representations*, Virtual Event, Austria, May 2021.

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

[34] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 12077–12090.

[35] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1800–1807.

[36] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 12159–12168.

[37] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[38] J. Wang et al., "RTFormer: Efficient design for real-time semantic segmentation with transformer," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 7423–7436.

[39] B. Dong, P. Wang, and F. Wang, "Head-free lightweight semantic segmentation with linear transformer," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 1, pp. 516–524.

[40] Z. Xu, D. Wu, C. Yu, X. Chu, N. Sang, and C. Gao, "SCTNet: Single-branch CNN with transformer semantic information for real-time segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 6, pp. 6378–6386.

[41] 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. Photogrammetry Remote Sens.*, vol. 162, pp. 94–114, 2020.

[42] 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.

[43] A. Ma, J. Wang, Y. Zhong, and Z. Zheng, "FactSeg: Foreground activation-driven small object semantic segmentation in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5606216.

[44] G. Deng, Z. Wu, C. Wang, M. Xu, and Y. Zhong, "CCANet: Class-constraint coarse-to-fine attentional deep network for subdecimeter aerial image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4401120.

[45] J. Wang, X. Chen, W. Jiang, L. Hua, J. Liu, and H. Sui, "PVNet: A novel semantic segmentation model for extracting high-quality photovoltaic panels in large-scale systems from high-resolution remote sensing imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 119, 2023, Art. no. 103309.

[46] Y. Du, Q. Sheng, W. Zhang, C. Zhu, J. Li, and B. Wang, "From local context-aware to non-local: A road extraction network via guidance of multi-spectral image," *ISPRS J. Photogrammetry Remote Sens.*, vol. 203, pp. 230–245, 2023.

[47] J. Chen, D. Zhang, Y. Wu, Y. Chen, and X. Yan, "A context feature enhancement network for building extraction from high-resolution remote sensing imagery," *Remote Sens.*, vol. 14, no. 9, 2022, Art. no. 2276.

[48] J. Yang, B. Matsushita, and H. Zhang, "Improving building rooftop segmentation accuracy through the optimization of UNet basic elements and image foreground-background balance," *ISPRS J. Photogrammetry Remote Sens.*, vol. 201, pp. 123–137, 2023.

[49] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang, "Multimodal token fusion for vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12176–12185.

[50] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 14679–14694, Dec. 2023.

[51] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5517010.

[52] B. Ren et al., "A dual-stream high resolution network: Deep fusion of GF-2 and GF-3 data for land cover classification," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, 2022, Art. no. 102896.
[53] W. Wu, S. Guo, Z. Shao, and D. Li, "CroFuseNet: A semantic segmentation network for urban impervious surface extraction based on cross fusion of optical and SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2573–2588, 2023.
[54] X. Li et al., "MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 106, 2022, Art. no. 102638.
[55] H. Gao, H. Feng, Y. Zhang, S. Xu, and B. Zhang, "AMSSE-Net: Adaptive multiscale spatial–spectral enhancement network for classification of hyperspectral and lidar data," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5531317.
[56] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot, "Multimodal fusion transformer for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5515620.
[57] W. Zhou, J. Jin, J. Lei, and J.-N. Hwang, "CEGFNet: CEGFNet: Common extraction and gate fusion network for scene parsing of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5405110.
[58] X. Ma, X. Zhang, and M.-O. Pun, "A crossmodal multiscale fusion network for semantic segmentation of remote sensing data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3463–3474, 2022.
[59] J. Ma, W. Zhou, J. Lei, and L. Yu, "Adjacent Bi-hierarchical network for scene parsing of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 3000705.
[60] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 548–558.
[61] H. Fan et al., "MultiScale vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6804–6815.
[62] W. Yu et al., "MetaFormer is actually what you need for vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10809–10819.
[63] Y.-H. Wu, Y. Liu, X. Zhan, and M.-M. Cheng, "P2T: Pyramid pooling transformer for scene understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12760–12771, Nov. 2023.
[64] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
[65] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[66] Z. Cao et al., "End-to-end DSM fusion networks for semantic segmentation in high-resolution aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 11, pp. 1766–1770, Nov. 2019.
[67] K. Yue, L. Yang, R. Li, W. Hu, F. Zhang, and W. Li, "TreeUNet: Adaptive tree convolutional neural networks for subdecimeter aerial image segmentation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 156, pp. 1–13, 2019.
[68] 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.
[69] C. Peng, K. Zhang, Y. Ma, and J. Ma, "Cross fusion net: A fast semantic segmentation network for small-scale semantic information capturing in aerial scenes," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5601313.
[70] L. Ding, H. Tang, and L. Bruzzone, "LANet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 426–435, Jan. 2021.
[71] Z. Cao, W. Diao, X. Sun, X. Lyu, M. Yan, and K. Fu, "C3Net: Cross-modal feature recalibrated, cross-scale semantic aggregated and compact network for semantic segmentation of multi-modal high-resolution aerial images," *Remote Sens.*, vol. 13, no. 3, 2021, Art. no. 528.
[72] W. Zhou, J. Jin, J. Lei, and L. Yu, "CIMFNet: Cross-layer interaction and multiscale fusion network for semantic segmentation of high-resolution remote sensing images," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 4, pp. 666–676, Jun. 2022.
[73] 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, 2022, Art. no. 5404418.
[74] Y. Wang, F. Sun, W. Huang, F. He, and D. Tao, "Channel exchanging networks for multimodal and multitask dense image prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5481–5496, May 2023.
[75] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[76] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," in *Proc. Int. Conf. Learn. Representation*, 2018.
[77] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking wider to see better," 2015, *arXiv:1506.04579*.
[78] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 140, pp. 20–32, 2018.
[79] S. He, H. Yang, X. Zhang, and X. Li, "MFTransNet: A multi-modal fusion with CNN-transformer network for semantic segmentation of HSR remote sensing images," *Mathematics*, vol. 11, no. 3, 2023, Art. no. 722.

Jihao Li (Member, IEEE) received the B.Sc. degree in electronic information engineering from Xidian University, Xi'an, China, in 2017, and the Ph.D. degree in signal and information processing from Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, in 2022.

He is currently an Assistant Researcher with Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision and remote sensing image interpretation.

Wenkai Zhang (Member, IEEE) received the B.Sc. degree in electronic information engineering from the China University of petroleum, Qingdao, China, in 2013, and the M.Sc. and Ph.D. degrees in electronic information engineering from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2018.

He is currently an Assistant Professor with Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include remote sensing image semantic segmentation and multimodal information processing.

Weihang Zhang (Student Member, IEEE) received the B.Sc. degree in electronic information engineering from Xidian University, Xi'an, China, in 2021. He is currently working toward the Ph.D. degree in signal and information processing with Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

His research interests include remote sensing image retrieval and multimodal remote sensing image interpretation.

Ruixue Zhou (Member, IEEE) received the B.Sc. degree in electronic information engineering from Shandong University, Jinan, China, in 2018, and the Ph.D. degree in communication and information system from Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, in 2023.

She is currently an Assistant Researcher with Aerospace Information Research Institute, Chinese Academy of Sciences. Her research interests include remote sensing image processing, particularly weakly supervised semantic segmentation, and scene understanding.

Chongyang Li (Member, IEEE) received the B.Sc. degree in electronic information engineering from the Dalian University of Technology, Dalian, China, in 2022. He is currently working toward the Ph.D. degree in signal and information processing with Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

His research interests include multimodal remote sensing image interpretation and multimodal signal processing.

Boyuan Tong (Student Member, IEEE) received the B.Sc. degree in electronic information engineering from Central South University, Changsha, China, in 2022. He is currently working toward the Ph.D. degree in communication and information system with the University of Chinese Academy of Sciences, Beijing, China, and the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing.

His research interests include computer vision, pattern recognition, especially on remote sensing image interpretation and multimodal signal processing.

Xian Sun (Senior Member, IEEE) received the B.Sc. degree from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2004, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2009, all in electronic information engineering.

He is currently a Professor with Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision, geospatial data mining, and remote sensing image understanding.

Kun Fu (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electronic information engineering from the National University of Defense Technology, Changsha, China, in 1995, 1999, and 2002, respectively.

He is currently a Professor with Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, remote sensing image understanding, geospatial data mining, and visualization.