SiMultiF: A Remote Sensing Multimodal Semantic Segmentation Network With Adaptive Allocation of Modal Weights for Siamese Structures in Multiscene

Shichao Cui, Wei Chen, Wenwu Xiong, Xin Xu, Xinyu Shi, and Canhai Li

***Abstract*— Semantic segmentation of remote sensing images is crucial for resource exploration, precision agriculture, and environmental monitoring. However, conducting semantic segmentation on single-modality data for remote sensing images that contain various scenes, especially unique scenes, is highly challenging. To address this challenge, we propose SiMultiF, a Siamese architecture-based multimodal feature adaptive fusion semantic segmentation network. SiMultiF employs a dual-branch Siamese structure feature extractor. The adaptive feature weight adjustment module (AFWAM) and the multimodal fusion module (MFM) facilitate in-depth understanding and extraction of multimodal data. Specifically, the Siamese structure can extract features from multimodal data concurrently without adding to the number of parameters. The AFWAM module can adaptively identify the importance of different modal data and dynamically adjust the modal weight to enhance the network's comprehension of complex scene data. Additionally, the cross-attention (CA)-based MFM module bridges modality gaps and achieves comprehensive multimodal feature fusion. Numerous experiments have demonstrated that the proposed SiMultiF outperforms other state-of-the-art semantic segmentation models (both multimodal and single modal) on the high-resolution ISPRS Potsdam dataset, ISPRS Vaihingen dataset, and special scene dataset (vegetation polarization dataset with extreme natural lighting contrast). Moreover, the robustness and generalizability of the network in multiscene and multimodal datasets are verified.**

***Index Terms*— Multimodal fusion, multiscene, semantic segmentation, siamese architecture.**

I. Introduction

SEMANTIC segmentation involves classification prediction at the image pixel level and the assignment of a semantic label for every pixel [1], [2]. Semantic segmentation pertains not only to categories but also to the boundary of each type of object [3]. Interpreting remote sensing images relies heavily on the crucial problem of semantic segmentation. With the rapid advancements in semantic segmentation technology, new approaches have emerged. These methods enable more efficient and precise segmentation of terrestrial objects in remote sensing images. This technique is widely used and holds significant importance in various fields, including land use change, smart cities, crop health monitoring, and environment detection [4], [5], [6].

Compared with natural images, remote sensing image data present complex scenes. Remote sensing data are characterized by rich surface object information, blurred outlines, and unbalanced foreground and background information [7]. In particular, in remote sensing image scenes, extreme brightness differences cause a decrease in feature differences between different classes of targets, blurring of class boundaries, disturbance of the continuity of features of the same class, and concealment of detailed information [8]. This leads to uncertainty in the remote sensing semantic segmentation process, thus reducing the segmentation accuracy [9]. On the other hand, remote sensing semantic segmentation is typically conducted via single-modality data, such as optical images [10], infrared images [11], high-resolution images [12], [13], and point cloud data [14], [15]. However, for remote sensing images with multiscene, semantic segmentation methods using single-modality data have limitations. Therefore, semantic segmentation algorithms based on multimodal data have become a research hotspot. Although multimodal data provide a large amount of supplementary information to increase the interpretability of remote sensing images, they also pose major challenges.

The various remote sensing devices' imaging mechanisms capture intricate object features, and semantic gaps exist between these modalities, complicating the process of mapping multimodal data into a unified feature space and performing stable and effective feature fusion. Consequently, increasing the number of modes without a strategy is not necessarily beneficial to segmentation accuracy. Only the multimodal fusion semantic segmentation algorithm under the optimization strategy can enhance the segmentation accuracy [16], [17], [18], [19], [20]. The current fusion techniques may be classified into three main types according to the stage of fusion: early interaction, late interaction, and throughout interaction [16]. Early interaction involves the fusion of data

Received 3 August 2024; revised 10 December 2024, 17 January 2025, and 25 February 2025; accepted 15 March 2025. Date of publication 21 March 2025; date of current version 4 April 2025. This work was supported in part by the Pre-Research on Civil Aerospace Technology During the 14th Five Year Plan Period "Surface Coverage Classification Technology for Multisystem Load Data Fusion" under Grant D010206, in part by the Fundamental Research Funds for the Central Universities (Ph.D. Top Innovative Talents Fund of CUMTB BBJ2024039), in part by the National Natural Science Foundation of China under Grant 42371351, in part by the Key Research and Development Program of Hebei Province under Grant 23310101D, and in part by China Scholarship Council under Grant 202406430025. *(Corresponding author: Wei Chen.)*

Shichao Cui, Wei Chen, and Wenwu Xiong are with the College of Geoscience and Surveying Engineering, China University of Mining and Technology, Beijing 100083, China (e-mail: cc_csc@163.com; chenw@cumtb.edu.cn; x1308567250@gmail.com).

Xin Xu is with China Siwei Surveying and Mapping Technology Company Ltd., Beijing 100083, China (e-mail: xuxin_5277@163.com).

Xinyu Shi is with the State Key Laboratory of Information Engineering in Surveying, Wuhan University, Wuhan, Hubei 430079, China (e-mail: xinyushi@whu.edu.cn).

Canhai Li is with the Land Satellite Remote Sensing Application Center, Ministry of Natural Resources, Beijing 100048, China (e-mail: lich@lasac.cn).

Digital Object Identifier 10.1109/TGRS.2025.3553713

before feature extraction, which is usually used for isomorphic data [21]. Late interaction is the direct fusion of the segmentation results obtained from the two modes and is usually suitable for heterogeneous data [20]. Throughout interaction, also known as feature fusion, fusion is performed in the feature extraction process [22]. Early or late fusion is easier, but it does not fully utilize the information between different modalities, potentially leading to semantic information loss. Conversely, throughout the interaction, the modal features of different scales are continuously fused, fully considering the semantic gap. To increase the network's sensitivity to feature information from various modes, this study uses the through-interaction fusion method. The challenge lies in effectively integrating multimodal data into a unified feature space, ensuring both stability and efficiency in feature fusion, without compromising the individual information of each modality. In addition, although the data between different modalities are closely related, multimodal fusion within the same framework has seldom been explored before.

However, the existing multimodal methods are usually aimed at specific scenarios and specific modes but lack multimodal networks for multiple scenarios (especially special scenes) [12], [23], [24], [25]. Although some multimodal models can be extended to multiple situations depending on their paradigm, their multimodal structure can lead to problems such as the doubling of network parameters, increased cost, and low efficiency. Therefore, in various complex remote sensing scenarios and multimodal data, effectively extracting important features in different modalities, reducing the influence of the semantic divide between these modes, and effectively integrating complementary information are highly important. Moreover, maintaining network accuracy at a low cost is one of the challenges facing multimode networks.

To address these challenges, this study proposes a multiscene-oriented multimodal fusion segmentation network with a Siamese structure (SiMultiF). Specifically, the network utilizes a Siamese structure, employing a dual-branch architecture with shared weights to simultaneously extract features from multimodal images. This structure ensures that multimodal data can be hierarchically fused via parallel inputs while controlling the network parameters to prevent exponential growth. In addition, we design an adaptive feature weight adjustment module (AFWAM) and a multimodal fusion module (MFM) to capture the most valuable information of different modal data accurately, and high-quality fusion of complementary information between modes is ensured. The robustness and generalizability of our network are validated through experiments and comprehensive evaluation, as is its ability to couple multimodal information to enhance semantic segmentation accuracy in multiscene remote sensing images. The primary advancements and contributions of this study are as follows.

1) This article introduces SiMultiF, a Siamese-structured transformer-based multimodal fusion network, designed to handle multiscene remote sensing data. It simultaneously extracts features from multimodal data, reducing network parameters and model complexity. This enhances the network's generalization capability and adaptability to multiscene images, enabling effective multiscale fusion to improve segmentation accuracy.
2) An AFWAM is proposed to capture valuable information across modalities. Positioned before multiscale feature fusion, AFWAM identifies key features and dynamically adjusts their weights. This adjustment improves feature extraction and ensures more reliable modal representations, enhancing the model's interpretability for complex data.
3) The MFM, built on a cross-attention (CA) mechanism, integrates global semantic information with local details from different modalities. It reduces modality discrepancies and fully fuses complementary information, significantly improving the network's fusion quality and accuracy.

This article is organized as follows. Section II offers a summary of the pertinent research. In Section III, we comprehensively explain the proposed method. Section IV describes the dataset and experimental demonstrations, including comparison experiments, ablation studies, visualization results, and corresponding analyses. Finally, Section V summarizes our work and offers insights into prospects.

II. Related Works

This section will examine the semantic segmentation methods for remote sensing images across different scenarios, as well as the methods for multimodal semantic segmentation.

A. Semantic Segmentation of Remote Sensing Images Based on Attention Mechanisms

An increasing number of researchers are exploring the complementary nature of multimodal data to increase precision and robustness in semantic segmentation, given the constraints of single sensors in extracting full features from remote sensing images. In multimodal semantic segmentation fusion strategies, the simpler strategy is to sum the pixel values of the features of different modes or the segmentation results or achieve them through channel superposition [26], [27]. While these methods are straightforward and computationally efficient, they may not be able to fully mine the correlations and differences between multimodal data. More complex fusion strategies, such as methods based on attention mechanisms, have received more attention from scholars. He et al. [28] studied the modality gap between LiDAR and red–green–blue (RGB) cameras and designed an attention fusion module within a dual-branch segmentation network, achieving 3-D semantic segmentation. Hu et al. [29] proposed an architecture with three parallel branches, utilizing an attention complement module for RGB-depth (RGB-D) image semantic segmentation by integrating feature information. Su et al. [30] integrated multimodal information via channelwise spatial attention blocks and a dual-encoder structure. Similarly, Liu et al. [31] developed a multimodal semantic segmentation network named FFCANet, with a global context module and a dual-branch encoder architecture. Zhang [32] proposed a bidirectional feature attention fusion network (DFFNet) for remote sensing image segmentation, which incorporates boundary information. Luo et al. [33]

applied the cross-modal feature fusion (CMFF) method based on self-attention for building extraction. CA has garnered significant attention from researchers because of its ability to establish connections between different modalities, facilitating information exchange and integration. Ma et al. [34] designed a cross-modal multiscale fusion network, CMTrans, which can effectively model long-range dependencies across multiscale feature maps derived from multimodal data. MSFNet [35] employs CA to fuse information from different scales and modalities in a multistage scheme. Li et al. [36] designed MCANet for semantic segmentation of optical and synthetic aperture radar (SAR) images, utilizing pseudo-Siamese feature extraction and a multimodal CA mechanism. CMX [37] leverages CA to reconstruct feature fusion in token dimensions, capturing modal correlations in the token space. Additionally, through the integration of CA and self-attention, multilevel multimodal fusion is achieved [38], combining convolutional neural network (CNN) and a vision transformer (ViT) into a unified fusion framework.

B. Transformer-Based Multimodal Semantic Segmentation for Remote Sensing Images

Owing to the excellent performance of the transformer algorithm in semantic segmentation tasks, transformer architectures have increasingly been adopted in the multimodal fusion domain. Xie et al. [39] proposed SegFormer, which combines multiscale feature fusion with transformer modules to achieve superior performance without the need for complex postprocessing. ViT [40] demonstrated the potential of replacing traditional convolutional features with pure transformer representations, thereby increasing the segmentation accuracy. Fan et al. [23] utilized transformer blocks to obtain multimodal data for identifying urban informal settlements (UISs) via high spatial resolution remote sensing images and polygonal data. ST-UNet [41] embeds the Swin transformer into UNet to effectively combine the local modeling capability of the CNN with the global modeling capacity of the Swin transformer. CMX [37] and TFIV [42] leveraged transformer fusion models to increase the accuracy of the RGB modality and supplementary modality (RGB-X). Wang et al. [43] used a transformer to dynamically detect noninformative tokens and replace them with cross-modal features, achieving efficient feature fusion. Additionally, Wang et al. [44] proposed a context exchange network (CEN) for multimodal and multitask dense image prediction, which adaptively exchanges channels between different modality subnetworks at its core. Wang et al. [45] proposed a transformer-based model with multiple lightweight modality fusion and adaptive multimodal matching modules to effectively model intramodal and intermodal relationships. Ma et al. [46] improved the transformer network by innovating interaction mechanisms across adjacent scale features, effectively capturing contextual cues while maintaining low computational complexity.

C. Semantic Segmentation of Remote Sensing Images in Various Scenes

Semantic segmentation plays a crucial role in understanding scenes. Recently, many researchers have developed advanced semantic segmentation networks for various remote sensing scenarios. Li et al. [47] introduced thermal infrared images to address backlighting conditions in natural images and proposed a feature fusion-based RGB-thermal (RGB-T) semantic segmentation network named LASNet, which follows the localization, activation, and sharpening steps. Peng et al. [48] addressed the image semantic segmentation problem of small-scale objects in aerial scenes with CF-Net. It uses channel attention mechanisms to select information and increases the receptive field of low-level feature maps by using cross-fusion blocks. A unique context-aware network that can process vast quantities of point cloud data directly and efficiently and is appropriate for urban settings was created by Liu et al. [49]. Qiang et al. [50] designed a hierarchical point cloud transformer framework to perform semantic segmentation of multisource vegetation point cloud data in vegetation scenes. High-resolution imagery semantic segmentation is an indispensable part of the remote sensing image segmentation field [13], [51], [52], [53]. Sun et al. [54] introduced a novel semantic segmentation algorithm, RSProtoSeg, which is based on unlearnable prototypes. This method optimizes the spatial relationships between foreground–background prototypes and intraclass prototypes, effectively mitigating foreground–background distribution imbalances. Guo et al. [55] proposed PIF-Net, a multimodal fusion network for ultrahigh-resolution imagery and aerial point cloud images, demonstrating the significant potential of multimodal networks in remote sensing applications.

III. Methodology

In Section III-A, we provide the general architecture of SiMultiF. Section III-B elaborates on the Siamese structure. Sections III-C and III-D offer detailed descriptions of AFWAM and MFM, respectively. Section III-E provides a detailed introduction to the loss function.

A. SiMultiF Framework

As depicted in Fig. 1, our proposed multimodal network, SiMultiF, uses an encoder–decoder architecture. This framework comprises a single dual-branch feature extractor composed of an AFWAM module, an MFM module, and a UperNet-based decoder.

Specifically, we simultaneously input images of different modes into the feature extractor of the Siamese structure. The feature maps of the two modes are obtained after stage 1 of the feature extractor. By automatically identifying the importance of modal information, two feature maps are input to the AFWAM module, which assigns different weights to different modal features. Then, the two weight-adjusted feature maps are input to the MFM module. This module can efficiently incorporate the enhancing information from the two modalities. After that, the two feature maps that the AFWAM module modified are input to the next stage of the feature extractor, and the above steps are repeated. Ultimately, the UperNet-based decoder receives fully fused feature maps from the four stages. The adjusted and fused feature map from Stage 4 is input into

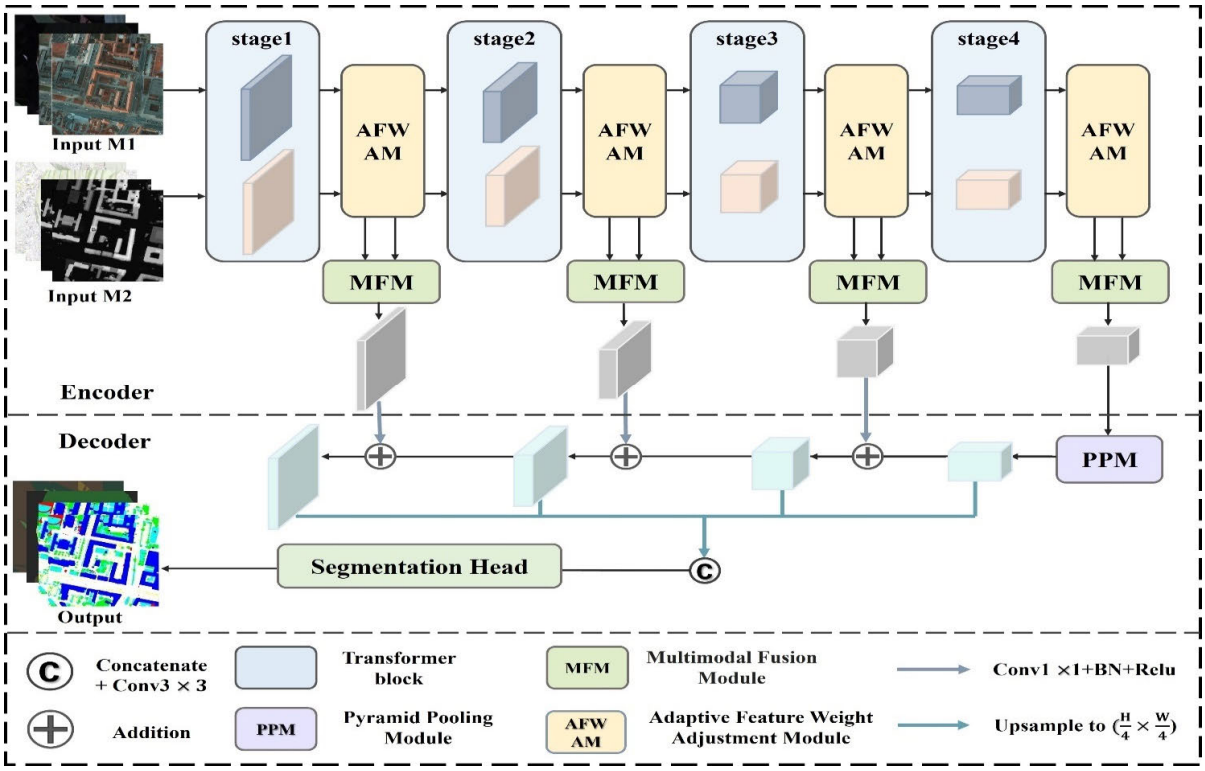

Fig. 1. Diagram of the overall structure of SiMultiF.

the pyramid pooling module (PPM) module, which produces a feature map that includes global spatial relationships and local detail information. The feature pyramid network (FPN) aggregates the multiscale feature data fused through the MFM, encompassing shallow spatial information and deep high-level semantics from complementary modalities, thereby restoring detailed information. Finally, the segmentation head refines the output features to produce the ultimate outcome of the segmentation process.

B. Siamese Structure

For extracting multimodal data, most methods use the pseudo-Siamese structure to simultaneously extract multimodal features via the same two networks but do not share weights; one can also use two different networks, as shown in Fig. 2(a). This structure can be applied to data with certain differences. However, because the weights are not shared, the general network has more parameters. In contrast, as illustrated in Fig. 2(b), weight sharing and the Siamese structure drastically reduce the overall number of parameters and complexities of the network yet preserve extraction accuracy.

In the Siamese network, a similarity metric can be learned from the data, and the learned metric is used to compare and match new samples of an unknown category. The Siamese network is implemented by sharing weights. The basic idea is to construct two subnetworks with the same structure by sharing parameters, and each subnetwork accepts an input sample for feature extraction. Compared with multimodal data such as pictures, text, and speech, which have large gaps, multimodal remote sensing data have strong similarities. Further experiments show that the Siamese network structure is suitable for multimodal semantic segmentation tasks. In addition to most existing mode fusion networks, this structure fails to increase the total number of network parameters, reduces network complexity, improves network generalization, or makes the network more friendly for multiscene remote sensing images. Therefore, as a feature extractor, we employ a Siamese network [56] structure involving two weight-sharing Swin transformers [57] as a backbone. Each branch is composed of four stages. Window multihead self-attention (W-MSA) and shifted W-MSA (SW-MSA) block feature information across different modes, and different scales are extracted layer by layer.

C. Adaptive Feature Weight Adjustment Module

Although multimodal data provide a large amount of complementary information, they can also contain some redundant information. To capture the most valuable information in different modes accurately, increase the sensitivity of the network to the complementary information between modes, and achieve information interaction between modes, we design the AFWAM module. It can adaptively identify the importance of feature information of different modes at different scales. It achieves cross-modal information interaction and adaptively improves the sensitivity to intermodal information by dynamically adjusting the feature weights. In addition, to increase the sensitivity of the network to modal complementary information, we use the AFWAM-adjusted information as the input of the next stage instead of the outcome features of the preceding

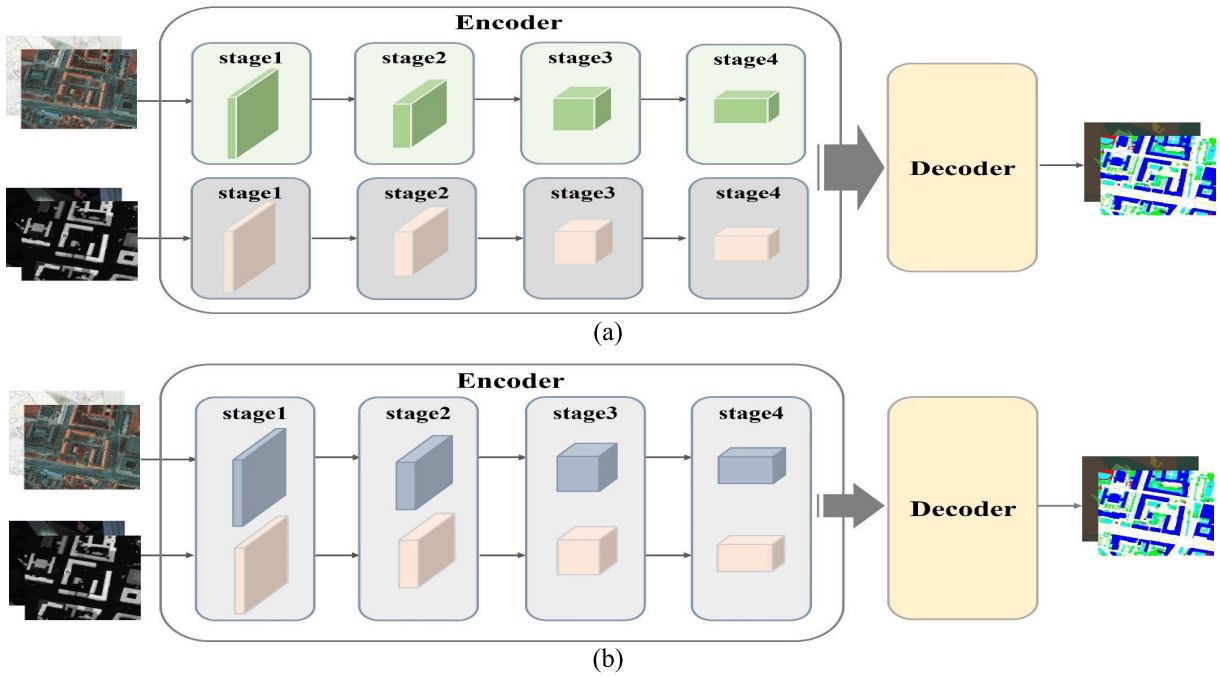

Fig. 2. Pseudo-Siamese structure and Siamese structure network. (a) Pseudo-Siamese structure. (b) Siamese structure.

stage as direct inputs. Fig. 3 shows the AFWAM module's full comprehensive diagram.

Previous attention mechanisms typically employ channel attention and spatial attention sequentially, which enhances feature information. Channel attention extracts channel-wise importance weights without considering contextual information to recalibrate features. Spatial attention can be used to calculate a spatial importance map to adaptively indicate the significance of various regions. However, spatial attention focuses only on the distribution of feature information at the spatial scale and ignores the information distribution at the feature level, whereas channel attention is strictly restricted to the channel dimension's feature arrangement. Our AFWAM overcomes these limitations by capturing the distribution of feature information across the spatial dimension for each feature channel. The mechanism simultaneously considers feature information from both spatial and channel dimensions, obtaining pixel-level feature weights, thereby enhancing the ability to adaptively adjust feature weights in complex scenes. Specifically, this module takes the feature maps F and F' obtained by two modalities feature extractors as inputs. First, channel attention and spatial attention operations are performed on the two features in parallel according to (1)–(4) to obtain $W_S(F)$, $W_C(F)$, $W'_S(F')$, and $W'_C(F')$, respectively. Next, we fully integrate the channel attention weights and spatial attention weights to obtain the coarse-grained feature information weights for each modality, yielding the modality-specific coarse-grained weight matrices $W_{SC}(F)$ and $W'_{SC}(F')$, as shown in (5) and (6)

$$W_S(F) = \mathrm{Conv}_{1\times1}\big(\mathrm{cat}\big(\big(\mathrm{AvgPool}_S(F), \mathrm{MaxPool}_S(F)\big)\big)\big) \tag{1}$$

$$W'_S(F') = \mathrm{Conv}_{1\times1}\big(\mathrm{cat}\big(\big(\mathrm{AvgPool}_S(F'), \mathrm{MaxPool}_S(F')\big)\big)\big) \tag{2}$$

$$W_C(F) = \mathrm{Conv}_{7\times7}\big(\max\big(0, \mathrm{Conv}_{1\times1}\big(\mathrm{AvgPool}_C(F)\big)\big)\big) \tag{3}$$

$$W'_C(F') = \mathrm{Conv}_{7\times7}\big(\max\big(0, \mathrm{Conv}_{1\times1}\big(\mathrm{AvgPool}_C(F')\big)\big)\big) \tag{4}$$

$$W_{SC}(F) = W_S(F) \oplus W_C(F) \tag{5}$$

$$W'_{SC}(F') = W'_S(F') \oplus W'_C(F'). \tag{6}$$

The globally average pooling operation across each spatial dimension is denoted by $\mathrm{Avgpool}_C$, $\mathrm{MaxPool}_S$ represents the global max pooling operation across the channel, and the global average pooling procedure across every channel dimension is represented via $\mathrm{Avgpool}_S$. Cat denotes the channelwise concatenation operation. Conv involves a convolution operation. $\mathrm{Max}(0, x)$ denotes the ReLU activation function. The elementwise addition function is denoted by $\oplus$.

After obtaining the coarse-grained feature information weights for the modalities, the next step is to refine the feature weights ω and ω', and we utilize the input features to guide the generation of the final feature weights. In particular, to fully mix the feature information of the channel dimension, after the concatenation of the channels of $W_{SC}(F)$ and F as well as $W'_{SC}(F')$ and F', each channel is rearranged in an alternating manner through the channel shuffling operation [58]. This enables the establishment of relationships between different channels and increases information interaction between the channels. The refined feature weights are then obtained through convolution and sigmoid activation. The formulas for this process are as follows:

$$\omega = \sigma\big(\mathrm{Conv}_{7\times7}\big(S^C\mathrm{cat}((W_{SC}(F), F))\big)\big) \tag{7}$$

$$\omega' = \sigma\big(\mathrm{Conv}_{7\times7}\big(S^C\mathrm{cat}\big(\big(W'_{SC}(F'), F'\big)\big)\big)\big) \tag{8}$$

where σ denotes the function known as the sigmoid function and S^C represents the channel shuffling function.

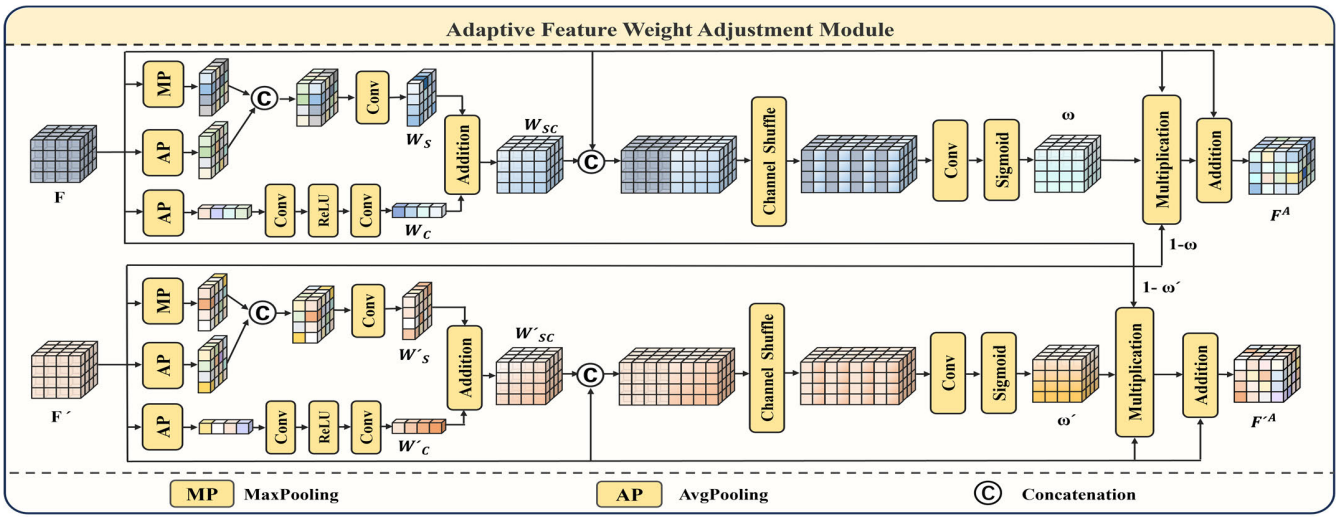

Fig. 3. Detailed structure of the AFWAM.

Finally, we combine information from different modalities using a weighted summation method, which adjusts the sensitivity of the network to crucial information within each modality based on their respective feature weights. The input features are added through skip connections to relieve the problem of diminishing gradients and simplify the acquisition procedure. This equation constitutes the following formula:

$$F^A = F + \omega F + F' \times (1 - \omega) \tag{9}$$

$$F'^A = F' + \omega' F' + F \times \left(1 - \omega'\right). \tag{10}$$

D. Multimodal Fusion Module

To completely use the complimentary data shared by different modalities, we design an attention mechanism-based MFM, depicted in Fig. 4, that effectively and stably fuses features from different modalities. In this module, we design an efficient self-attention-based CA method. The CA method offers complementary interactions between sequences rather than correction-based interactions from a feature map perspective. This mechanism better facilitates the exchange and fusion of modal information, and features are further fused through mixed channel embedding. This module enhances the fusion quality of the network, thus improving the model's capacity for generalization.

The input of the MFM module includes the AFWAM-adjusted F_{A} and F'_{A} features of each stage. First, we flatten the features of the two branches and then use linear embedding to generate F_Y and F_I as well as F'_Y and F'_I, which retain the initial features, which are used in the subsequent residual structure for information exchange. F_I and F'_I are encoded as query (Q), key (K), and value (V), or query′ (Q'), key′ (K'), and value′ (V'). The transpose of K is multiplied by V to calculate the global context vector, with the weight determined via Softmax. The global attention map is obtained according to (11) and (12). According to (13) and (14), the interaction vector is multiplied by the context vector from another modal path to convert a CA component by converting the attention function. F_Y and F'_Y that are used for the interaction and retain the initial features are merged to obtain F_{CA} and F'_{CA}

$$F_{\mathrm{GC}} = \mathrm{SoftMax}\left(K^T \times V\right) \tag{11}$$

$$F'_{\mathrm{GC}} = \mathrm{SoftMax}\left(K'^T \times V'\right) \tag{12}$$

$$F_{\mathrm{CA}} = \mathrm{cat}\left(F_Y, Q \times F'_{\mathrm{GC}}\right) \tag{13}$$

$$F'_{\mathrm{CA}} = \mathrm{cat}\left(F'_Y, Q' \times F_{\mathrm{GC}}\right) \tag{14}$$

where F_{GC} and F'_{GC} represent the global attention weight maps of the two modalities and F_{CA} and F'_{CA} represent the crossover results.

Next, we fuse the obtained crossover results with the original features via the residual connection method and obtain the F_{CP} and F'_{CP} through normalization, as expressed in (15) and (16). By concatenating them in series and then processing them with a multilayer perceptron (MLP) through an MLP, we can obtain the stage-specific feature output, denoted F_M, via (17)

$$F_{\mathrm{CP}} = \mathrm{Normal}\left(\mathrm{channelEmbedding}\left(F^A\right) + F_{\mathrm{CA}}\right) \tag{15}$$

$$F'_{\mathrm{CP}} = \mathrm{Normal}\left(\mathrm{channelEmbedding}\left(F'^A\right) + F'_{\mathrm{CA}}\right) \tag{16}$$

$$F_M = \mathrm{MLP}\left(\mathrm{cat}\left(F_{\mathrm{CP}} + F'_{\mathrm{CP}}\right)\right). \tag{17}$$

In the latter part of MFM, we employ convolutional layers for deep feature fusion. In addition, in this channel-by-channel fusion process, the information of the surrounding region should be used to assist in segmentation. Therefore, we add a 3×3 depth convolutional layer in the middle to implement the skip connection structure to promote in-depth fusion, and the formula is as follows:

$$\begin{aligned} F_F = \mathrm{Normal}(&\mathrm{Conv1} \times 1(F_M) \\ &+ \mathrm{Conv1} \times 1(\mathrm{ReLU}(\mathrm{DWConv3} \times 3 \\ &\times (\mathrm{Conv1} \times 1(F_M))))). \end{aligned} \tag{18}$$

E. Loss Function

In this study, the cross-entropy loss function is adopted as follows:

$$L = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k} \log \hat{y}_{i,k} \tag{19}$$

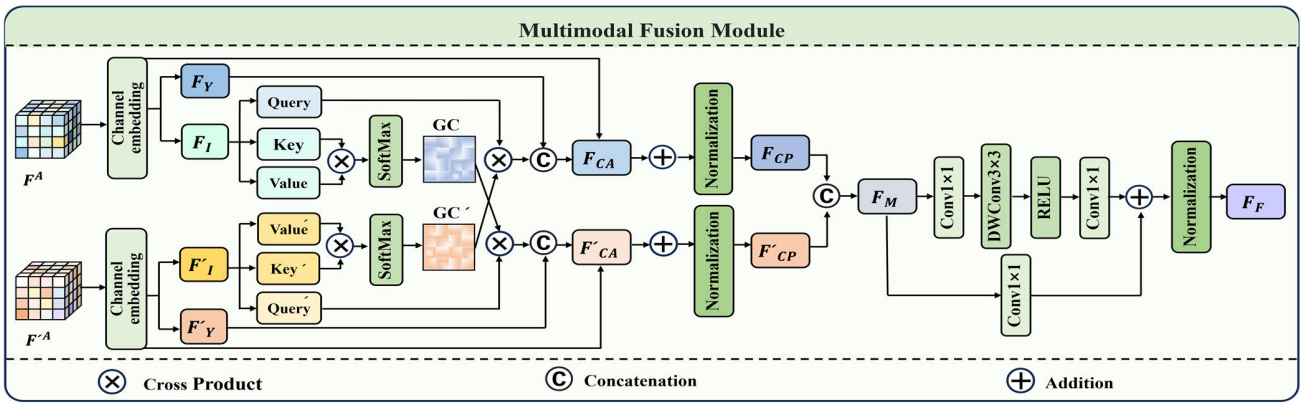

Fig. 4. Detailed structure of MFM.

where L denotes the cross-entropy loss, k is the number of categories, N is the total number of samples, y_i is the true category distribution, and $\hat{y}_i$ is the probability of the ith category being predicted by the model.

IV. Experimental Settings

The datasets used in the network experiment are introduced in this section, along with details on the experimental parameter settings, and the experimental results are presented and analyzed. We test the suggested approach via a range of datasets, including complex and specialized scenarios. First, we introduce the remote sensing multimodal dataset that was employed in this work. Then, we describe the training details and evaluation metrics for the semantic segmentation task. Additionally, we further analyze the model's performance as measured by normalization methods, network costs, and other aspects. Finally, we conduct comparative and ablation experiments on remote sensing multimodal datasets across different scenarios. In contrast with recent state-of-the-art (SOTA) networks and evaluating module performance, we use intricate flexibility and applicability with the proposed method in complex remote sensing imagery.

A. Datasets

1) ISPRS Potsdam Datasets: There are 38 extremely high-resolution real-world orthophoto images in the ISPRS Potsdam dataset, each with a matching digital surface model (DSM). Each image has pixels measuring 6000 × 6000 pixels with a spatial resolution of 5 cm. A pair of patches are offered by the dataset: four multispectral channels, including infrared, red, green, and blue (IRRGB) channels, and DSM images. The dataset is refined into four modes of data: RGB, DSM, IRGB, and RGBIR, as depicted in Fig. 5. The dataset consists of six classifications: buildings (Bui.), trees (Tre.), impervious surfaces (Imp.), cars (Car.), low vegetation (Low.), and Clutter (Clu.). In addition, we selected 24 DSM images that were normalized and divided into testing and training sets at random in a proportion of three to one. The training set contains orthophotos indexed by 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, and 4_11, whereas tests 2_11, 3_10, 4_10, 5_11, 6_11,

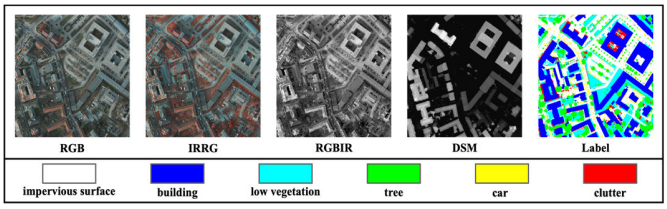

Fig. 5. Visualization of the ISPRS Potsdam dataset used.

and 7_12. To read large images without cropping in advance, learning batches are continuously gathered via a window sliding method. The 256 × 256 pixels with a stride of 256 are the dimensions of the sliding window that are specified during training. For the testing phase, the stride size was set to 32. At that stage, a lower stride size can assist in alleviating the border effect by averaging the expected results in the sections that overlap.

2) ISPRS Vaihingen Dataset: The ISPRS Vaihingen dataset comprises 33 high-resolution real orthophoto images of varying sizes, with an average size of 2500 × 2000 pixels. Each orthophoto features three channels: near-infrared, red, and green (NIRRG), along with a standardized DSM with a ground sampling distance (GSD) of 9 cmas, depicted in Fig. 6. This dataset encompasses five foreground classes: buildings (Bui.), trees (Tre.), low vegetation (Low.), cars (car), and impervious surfaces (Imp.), as well as one background class (Clutter). Additionally, we chose 16 images from the dataset as the training and testing sets, adopting the same dataset partitioning method as described in the literature [38]. Specifically, the training set includes orthophotos indexed by 1, 3, 23, 26, 7, 11, 13, 28, 17, 32, 34, and 37, whereas the test set comprises orthophotos indexed by 5, 21, 15, and 30.

3) Vegetation Polarization Dataset With Extreme Natural Light Contrast: The vegetation polarization dataset with extreme natural lighting contrast contains four categories, namely, illuminated vegetation, illuminated soil, shaded vegetation, and shaded soil. In the scene of natural lighting contrast, the four components of the geometric optics of vegetation had certain differences in terms of light intensity and polarization; among these components, the light intensity images of the sunlit vegetation and sunlit soil were significantly different, but the

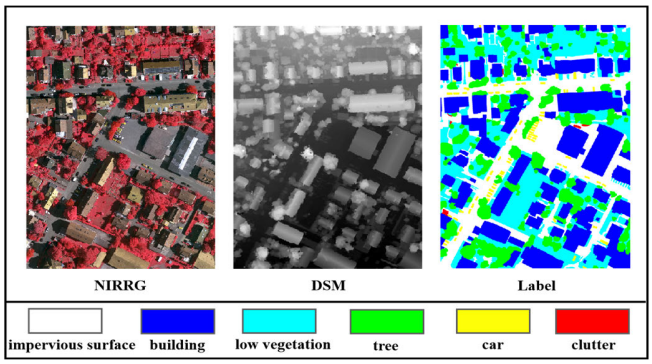

Fig. 6. Visualization of the ISPRS Vaihingen dataset used.

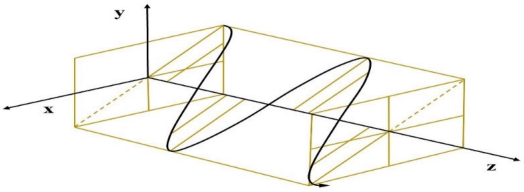

Fig. 7. Schematic diagram of linearly polarized light.

hue and light intensity images of the shaded soil and shaded vegetation were not different. Obviously, there are certain difficulties in quantification and automated segmentation. The unique directional vibration information of polarization can provide the scattering features associated with the vegetation in the shaded region and provide unique information for the differentiation of soil and shaded vegetation [59]. The trajectory of the electric vector of light on a plane parallel to the propagation axis can be divided into polarized light with different polarization states. Fig. 7 shows a simplified diagram of linearly polarized light.

When a light wave is transmitted along the z-axis, the electric field components in each direction are expressed as

$$\begin{cases} E_x = E_{0x}(\omega t - kz + \delta_1) \\ E_y = E_{0y}(\omega t - kz + \delta_2) \\ E_z = 0 \end{cases} \tag{20}$$

where ω represents the angular frequency, t stands for time, k denotes the wave vector (representing the spatial periodicity of the wave), and φ represents the phase. E_{0x} and E_{0y} are the amplitudes, and δ is the phase. When the phase in the x- and y-directions satisfies (21), the trajectory of the electric field vector becomes a straight line, which is referred to as linear polarization. The specific formula is as follows:

$$\delta = \delta_2 - \delta_1 = \pi m (m = 0, \pm 1, \pm 2, \ldots) \tag{21}$$

$$\frac{E_y}{E_x} = (-1)^m \frac{E_{0y}}{E_{0x}}. \tag{22}$$

By employing a Stokes vector representation, the polarization situation of the light mirrored by a surface is determined [60], which employs four vectors $S = (\mathrm{IQUV})^T$, to depict the degree of polarization in a light beam. These parameters I, Q, U, and V are independent and can be expressed in terms of the electric field components E_x and E_y

$$S = \begin{bmatrix} I \\ Q \\ U \\ V \end{bmatrix} = \begin{bmatrix} E_x^2 + E_y^2 \\ E_x^2 - E_y^2 \\ 2E_x E_y \cos\delta \\ 2E_x E_y \sin\delta \end{bmatrix}. \tag{23}$$

The variables of interest are as follows: I is the overall incident light intensity; Q and U denote the direction and intensity of linear polarization, respectively; and V indicates the circularly polarized component.

Each ground object has distinct natural properties reflected in its degree of linear polarization (DoLP), and the ratio of the light intensity of the linearly polarized light within a beam to the total light intensity can be determined. The calculation formula of the DoLP represented by the Stokes vector is shown in the following equation:

$$\mathrm{DoLP} = \frac{\sqrt{(Q^2 + U^2)}}{I}. \tag{24}$$

Triton TRI050S-Q triple-channel polarization imaging equipment was used together with Sony's IMX250MYR CMOS polarization lens sensor for data acquisition. This dataset includes 495 vegetation images of different light levels under natural outdoor light, as shown in Fig. 8. The pixel size of each photograph is 1224 × 1024. In this study, images of vegetation were collected via two modalities, namely, the DoLP and the light intensity. During the experiment, training, validation, and test sets were selected at random from the dataset at a 7:2:1 ratio. To optimize network performance, data augmentation was applied to the training and validation sets. The images were rotated and nonoverlapping and cropped to 512 × 512 pixels. The test set images were cropped without overlap.

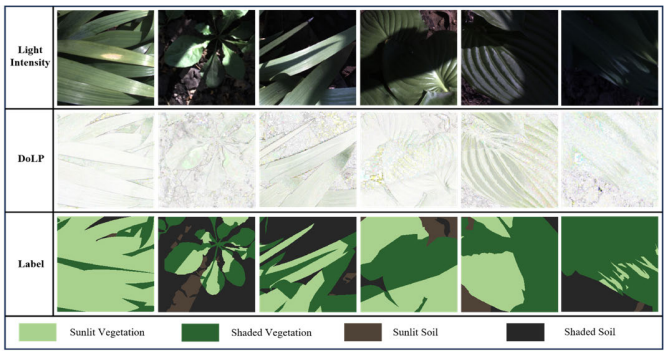

Fig. 8. Samples of the vegetation polarization dataset with extreme natural lighting contrast. The intensity image is displayed in the first line. The images from the DoLP modality are displayed in the second line. The labels corresponding to each entry are displayed in the third line.

B. Evaluation Metrics

To conduct a comprehensive evaluation of how well our suggested multimodal fusion segmentation network SiMultiF performs, the overall accuracy (OA), the mean intersection over union (mIoU), and the $F1$ score ($F1$) were used as the

evaluation indicators, and the formula was as follows:

$$\mathrm{OA} = \frac{\sum_{k=1}^{P} \mathrm{TP}_K + \mathrm{TN}_K}{\sum_{k=1}^{P} \mathrm{TP}_K + \mathrm{FP}_K + \mathrm{TN}_K + \mathrm{FN}_K} \tag{25}$$

$$\mathrm{IoU} = \frac{1}{P} \sum_{k=1}^{P} \frac{\mathrm{TP}_K}{\mathrm{TP}_K + \mathrm{FP}_K + \mathrm{FN}_K} \tag{26}$$

$$F1 = 2\frac{Q \times R}{Q + R} \tag{27}$$

where TP_k represents the true positive, FP_k represents the false positive, TN_k represents the true negative, and FN_k represents the false negative. K stands for an item that is a member of the kth class. In addition, both Q and R are defined as

$$Q = \frac{1}{P} \sum_{k=1}^{P} \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k} \tag{28}$$

$$R = \frac{1}{P} \sum_{k=1}^{P} \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k}. \tag{29}$$

C. Implementation Details

In this research endeavor, all experiments were conducted leveraging the PyTorch deep learning framework, exclusively utilizing a solitary NVIDIA GeForce RTX 4090 GPU endowed with 24 GB of memory. Model training adhered to the AdamW optimization algorithm, employing a momentum factor of 0.9 and a decay rate of 0.01. For our custom network architecture, the learning rate was calibrated to 0.00006, whereas the learning rates for the remaining networks were prescribed in accordance with their respective original specifications. With respect to data handling, the vegetation polarization dataset with extreme natural lighting contrast was uniformly processed with a batch size of 8, and the ISPRS Potsdam dataset and ISPRS Vaihingen dataset were configured with a batch size of 16. Furthermore, after sample collection via the sliding window approach in the ISPRS Potsdam dataset and ISPRS Vaihingen dataset, simple data augmentation methods, including flips and random rotations, were used to improve model generalizability and supplement the dataset. For the ISPRS Potsdam dataset and ISPRS Vaihingen dataset, all the models were trained for 50 epochs, and for the vegetation polarization dataset with extreme natural lighting contrast, all the models were trained for 100 epochs.

D. Performance Comparison

Comparative performance analysis of our freshly developed SiMultiF network was performed against 11 cutting-edge networks within the current research landscape, namely, the Swin transformer [57], Res2Net [61], GhostNet V2 [62], ConvNeXt [63], RMT [64], MaNet [65], UNetFormer [66], BEDSN [67], NLFNet [68], CMX [37], and FTransUNet [38]. In our experiments, the Swin transformer, Res2Net, Ghost V2, ConvNeXt, RMT, MANet, BEDSN, and UNetFormer methods considered only dominant modal information, whereas the other methods considered cross-modal data information. To demonstrate the applicability of the SiMultiF model in complex scene remote sensing images, we perform comparative experiments via the ISPRS Potsdam dataset, ISPRS Vaihingen dataset, and vegetation polarization dataset with extreme natural lighting contrast. We use the $F1$ score, OA, and mIoU of each segmentation class to evaluate the accuracy of each model. Experiments validate the practicability of the proposed model with respect to the multimodal semantic segmentation responsibility of multiscale remote sensing scenes, as well as the generalizability and robustness of SiMultiF.

1) Performance Comparison on the ISPRS Potsdam Datasets: In the ISPRS potsdam dataset experiment medium, the Swin transformer, Res2Net, Ghost V2, ConvNeXt, RMT, MANet, BEDSN, and UNetFormer consider only visible image feature image (represented by the GRB in Table I) information. For the dual-input multimode model, we choose the DSM data as the input for the second mode. As shown in Table I, the classification accuracies of impervious surfaces, buildings, low vegetation, trees, cars, and clutter were 93.89%, 97.40%, 85.77%, 84.42%, 95.01%, and 67.14%, respectively; the OA reached 91.02% and the mIoU was 78.76%. The experimental results prove that our SiMultiF network has achieved substantial improvements over other SOTA methods on impervious surfaces, buildings, low vegetation, and clutter. In addition, the SiMultiF multimodal input method has clear advantages over the unimodal input network. Compared with the best results for the dual-input multimodal network, the OA and mIoU of FTransUNet increased by 0.81% and 0.77%, respectively. In the ISPRS Potsdam dataset, the features of low vegetation and tree classes exhibit considerable overlap, which could hinder the network's ability to effectively distinguish between them in complex scenes. Although the proposed SiMultiF network benefits from multimodal feature fusion, it is possible that during the fusion process, the network adaptively adjusts the modal feature weights to avoid redundant information, potentially overlooking subtle differences between similar features. This may have led to suboptimal $F1$ scores for the tree and car classes.

Fig. 9 shows a visualization example for all methods considered on the ISPRS Potsdam dataset. Our SiMultiF algorithm can classify buildings and low vegetation as well as clutter more accurately. This is because our network better utilizes complementary information in RGB and DSM. Two red dashed boxes have been added to Fig. 9(a)–(o). The upper left box shows the mixed segmentation of buildings and low vegetation. In this complex scene, SiMultiF has more accurate segmentation accuracy and clearer boundaries. The lower right box shows the area of clutter. For impervious surfaces, most other methods will confuse clutter with impervious surfaces, whereas SiMultiF will segment clutter more completely.

2) Performance Comparison on the ISPRS Vaihingen Datasets: In the experiments conducted on the ISPRS Vaihingen dataset, the single-input and dual-input network configurations are consistent with those used in the ISPRS Potsdam dataset. The dual-input network takes the NIRRG and DSM data as inputs. Since the clutter class occupies a small proportion of the ISPRS Vaihingen dataset [67] and is not included in the test set, we consider only the five foreground classes in Table II. According to Table II, the $F1$ score for the

TABLE I
PERFORMANCE COMPARISON ON THE ISPRS POTSDAM DATASETS. WE INTRODUCE THREE INDICATORS, $F1$, OA, AND MIOU FOR THE FOUR CATEGORIES

Method	Data	F1(%)						OA(%)	mIoU (%)
		Imp.	Bui.	Low.	Tree.	Car.	Clu.		
Swin Transformer	RGB	90.98	95.36	83.37	82.13	94.14	63.35	88.29	74.20
Res2Net	RGB	90.88	95.26	82.55	77.96	85.64	61.60	87.75	73.96
GhostNet V2	RGB	91.00	95.47	83.32	82.26	94.19	62.42	88.32	75.08
ConvNeXt	RGB	91.76	95.80	84.17	85.29	**95.83**	64.61	89.34	75.56
RMT	RGB	88.26	92.28	84.46	84.65	95.24	57.69	87.40	73.68
MANet	RGB	90.82	96.05	79.50	77.05	87.51	55.89	86.64	70.22
UNetFormer	RGB	91.93	95.91	84.04	82.66	94.30	59.60	88.89	75.30
BEDSN	RGB	91.78	97.00	84.55	84.44	94.66	60.69	89.57	76.63
NLFNet	RGB, DSM	92.10	95.29	84.47	84.07	94.83	64.29	89.19	76.51
CMX	RGB, DSM	92.02	95.82	84.13	82.42	94.23	59.72	88.94	75.34
FTransUNet	RGB, DSM	92.37	97.15	85.62	**85.39**	95.30	64.27	90.21	77.99
Ours	RGB, DSM	**93.89**	**97.40**	**85.77**	84.42	95.01	**67.14**	**91.02**	**78.76**

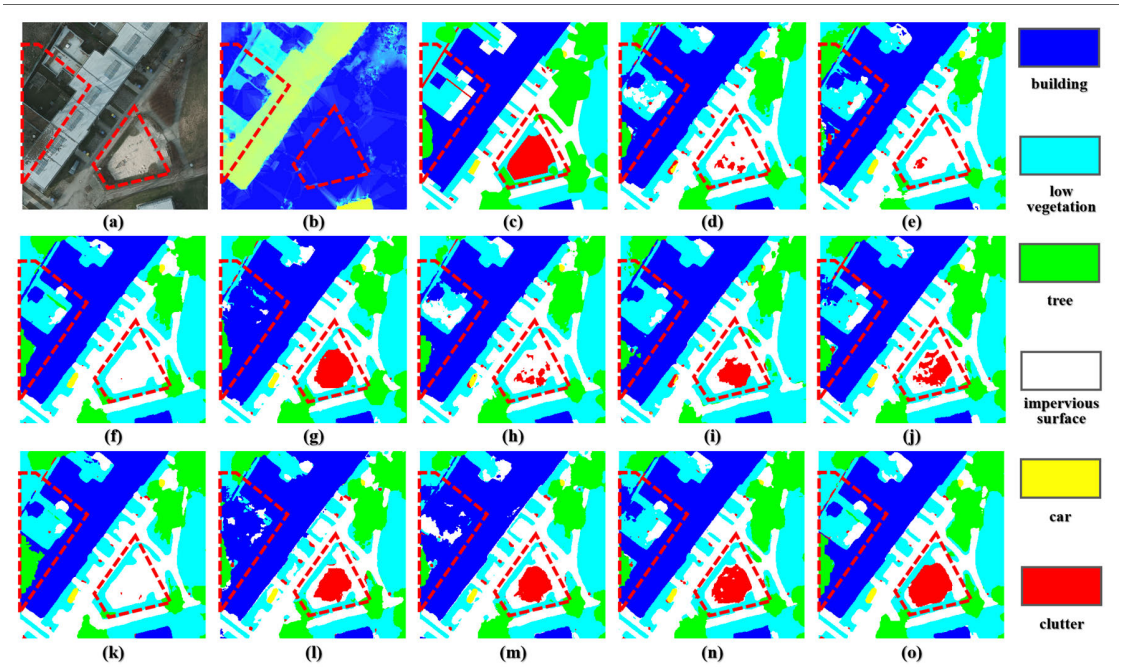

Fig. 9. Qualitative performance comparison was performed on the ISPRS Potsdam dataset test. (a) RGB images. (b) DSM. (c) Ground truth. (d) Swin transformer. (e) Res2Net. (f) GhostNet V2. (g) ConvNeXt. (h) RMT. (i) MANet. (j) UNetFormer. (k) BEDSN. (l) NLFNet. (m) CMX. (n) FTransUNet. (o) SiMulitiF. Two red boxes have been added to all the subgraphs to highlight differences.

five classes is slightly lower than that of the ISPRS Potsdam dataset because of the relatively small data size and fewer training epochs. The class car is particularly underrepresented in the test set, leading to a noticeable decline in accuracy. Comparing the robustness of SiMultiF and FTransUNet, both outperform other networks. FTransUNet achieves the highest accuracy in the building and car classes, with values of 96.44% and 85.33%, respectively. SiMultiF achieves the best performance in the impervious surface, low vegetation, and tree classes, with accuracies of 92.68%, 82.91%, and 92.45%, respectively. The OA and mIoU of SiMultiF are 91.86% and 81.16%, respectively, which are higher than

TABLE II
PERFORMANCE COMPARISON ON THE VARIOUS DATASETS. WE INTRODUCE THREE INDICATORS, $F1$, OA, AND MIOU FOR THE FOUR CATEGORIES

Method	Data	F1(%)					OA (%)	mIoU(%)
		Imp.	Bui.	Low.	Tree.	Car.		
Swin Transformer	RGB	88.40	93.35	76.41	89.50	59.84	87.99	70.45
Res2Net	RGB	89.99	93.63	75.49	88.95	68.53	88.37	72.54
GhostNet V2	RGB	88.88	94.46	72.43	87.36	46.20	87.25	68.29
ConvNeXt	RGB	90.27	95.33	76.09	88.00	68.46	88.89	73.03
RMT	RGB	90.76	95.40	75.62	88.70	70.30	89.04	73.59
MANet	RGB	89.52	93.77	76.88	89.70	68.53	88.78	71.26
UNetFormer	RGB	89.78	95.03	77.93	89.80	71.77	89.15	74.66
BEDSN	RGB	90.77	95.38	78.95	90.85	78.51	90.16	77.34
NLFNet	RGB, DSM	91.41	95.99	78.39	90.20	77.55	90.32	77.28
CMX	RGB, DSM	90.72	95.45	79.07	89.84	73.72	89.78	75.93
FTransUNet	RGB, DSM	92.26	**96.44**	80.74	91.27	**85.33**	91.37	80.96
Ours	RGB, DSM	**92.68**	96.02	**82.91**	**92.45**	82.28	**91.86**	**81.16**

those of FTransUNet by 0.49% and 0.20%, respectively. Overall, SiMultiF demonstrates the highest accuracy compared with existing SOTA models. Specifically, compared with the existing single-modal BEDSN method with better precision, SiMultiF increases the mIoU by 3.82%, and the $F1$ score for the low vegetation class improves by 3.96%. Moreover, compared with the existing multimodal method NLFNet, the mIoU improves by 3.88%. These results indicate that the proposed SiMultiF network exhibits excellent robustness and generalization capabilities.

As shown in Fig. 10, our SiMultiF network can more accurately classify buildings, low vegetation, and trees. This is because our network can adaptively adjust and integrate DSM data. In Fig. 10(a)–(o), we added a red box to highlight the segmentation results. In densely built-up areas, SiMultiF provides more accurate segmentation, with smoother edge delineation and less interference from shadows, indicating that our network effectively leverages both DSM and NIRRG data.

3) Performance Comparison of the Vegetation Polarization Dataset With Extreme Natural Lighting Contrast: In the experiments conducted on the vegetation polarization dataset with extreme natural lighting contrast, the single-input network considered only the light intensity modality information (denoted as GRB in Table III). The NLFNet, CMX, FTransUNet, and SiMultiF methods incorporate DoLP modality information. According to Table III, in contrast to the unimodal Swin transformer, the proposed SiMultiF significantly improved the $F1$, OA, and mIoU. Specifically, in contrast to the single-input RGB modality data, the mIoU increased by 2.03%, and it further improved by 3.29% compared with the single-input DoLP modality data. This confirms that the new SiMultiF method can extract important complementary information from the DoLP mode and successfully combines the characteristics of the two modalities, enhancing the accuracy of vegetation extraction, particularly in scenes with extreme lighting contrasts. Compared with the existing SOTA models, SiMultiF outperforms the other methods in four categories. Specifically, compared with the existing single-mode RMT method, which has better accuracy, SiMultiF improves the mIoU by 1.55%. Compared with the BEDSN method, SiMultiF increased the mIoU by 1.23%. The classification accuracy for shaded vegetation increased by 1.13%. Additionally, compared with the existing multimodal method FTransunet, the mIoU method achieves an improvement of 0.54%. Compared with the existing NLFNet methods, the mIoU increased by 1.1%, and compared with CMX, this indicator increased by 1.44%. These results can be explained by SiMultiF selecting the most valuable modal information by dynamically adjusting the modal weights, which is conducive to better extraction and fusion of multilevel multimodal features. In terms of overall effectiveness, the proposed SiMultiF algorithm achieved an OA of 93.67%, the $F1$ scores of different categories were 95.43%, 94.31%, 93.30%, and 91.67%, and the mIoU was 88.14%. These outcomes demonstrated that the proposed SiMultiF network achieves very good robustness.

As illustrated in Fig. 11, our network maintains outstanding performance on specialized datasets, particularly under extreme lighting conditions. By adaptively selecting complementary information from both light intensity and DoLP data and effectively integrating cross-modal information, we can enhance the segmentation of objects under extreme lighting conditions. As depicted in the red boxes in Fig. 11(a)–(p), SiMultiF exhibits enhanced differentiation between vegetation and soil in shadows. Additionally, our

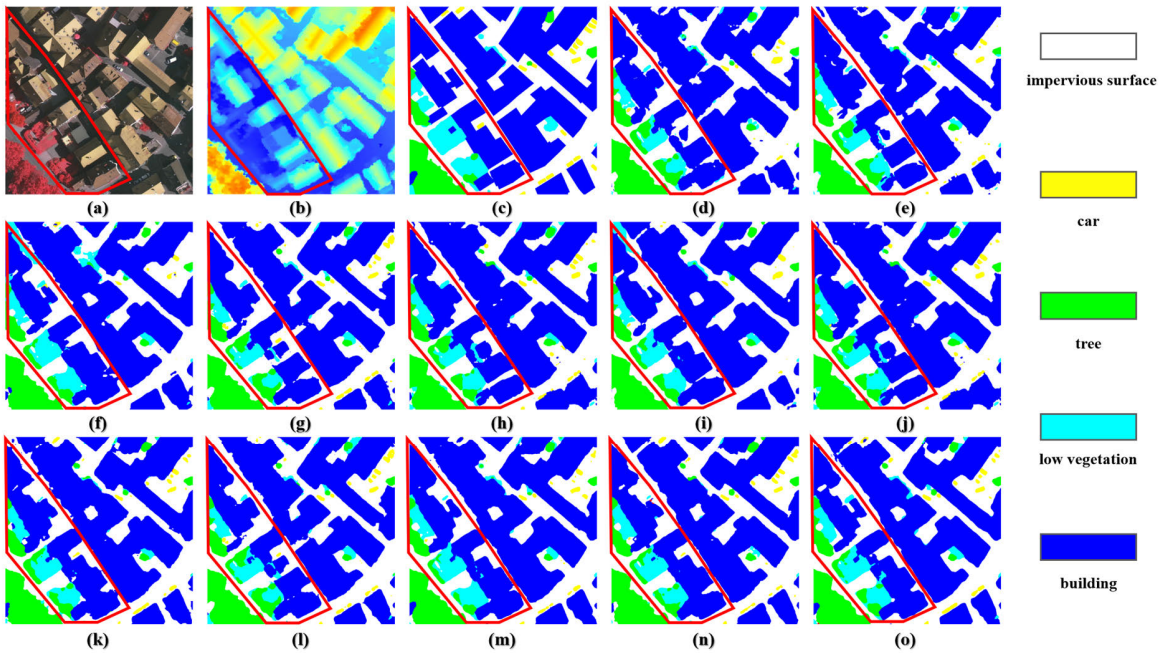

Fig. 10. Performance of various networks compared on the ISPRS Vaihingen dataset test set. (a) NIRRG images. (b) DSM. (c) Ground truth. (d) Swin transformer. (e) Res2Net. (f) GhostNet V2. (g) ConvNeXt. (h) RMT. (i) MANet. (j) UNetFormer. (k) BEDSN. (l) NLFNet. (m) CMX. (n) FTransUNet. (o) SiMulitiF. An orange box was added to all the subgraphs to highlight differences.

TABLE III

ANALYSIS OF THE RESULTS OF THE VEGETATION POLARIZATION AND LIGHT INTENSITY DATASETS FOR SHADOW SCENE. WE INTRODUCE THREE INDICATORS, $F1$, OA, AND THE MIOU FOR THE FOUR CATEGORIES

Method	Data	F1(%)				OA (%)	mIoU (%)
		Sunlit Soil	Shaded Soil	Sunlit Vegetation	Shaded Vegetation		
Swin Transformer	RGB	93.89	93.04	92.58	90.58	92.53	86.11
	DoLP	93.80	91.77	92.07	89.49	91.77	84.85
Res2net	RGB	94.85	93.54	91.80	89.42	92.42	85.95
GhostNet V2	RGB	94.37	93.51	91.84	89.94	92.42	85.94
ConvNeXt	RGB	94.22	93.17	92.05	89.94	92.34	85.82
RMT	RGB	94.95	93.56	92.08	90.54	92.76	86.59
MANet	RGB	94.28	93.36	92.07	89.78	92.39	85.87
UNetFormer	RGB	94.34	93.13	91.97	89.74	92.31	85.74
BEDSN	RGB	95.31	94.07	91.95	90.55	92.96	86.91
CMX	RGB, DoLP	94.94	93.43	92.42	90.58	92.83	86.70
NLFNet	RGB, DoLP	95.08	93.99	92.42	90.69	93.04	87.04
FTransunet	RGB, DoLP	95.01	94.03	92.98	91.50	93.37	87.60
Ours	RGB, DoLP	**95.43**	**94.31**	**93.30**	**91.67**	**93.67**	**88.14**

network boasts smoother classification segmentation lines. Furthermore, SiMultiF demonstrates strong generalizability, making it well suited for remote sensing image semantic segmentation tasks in specialized settings.

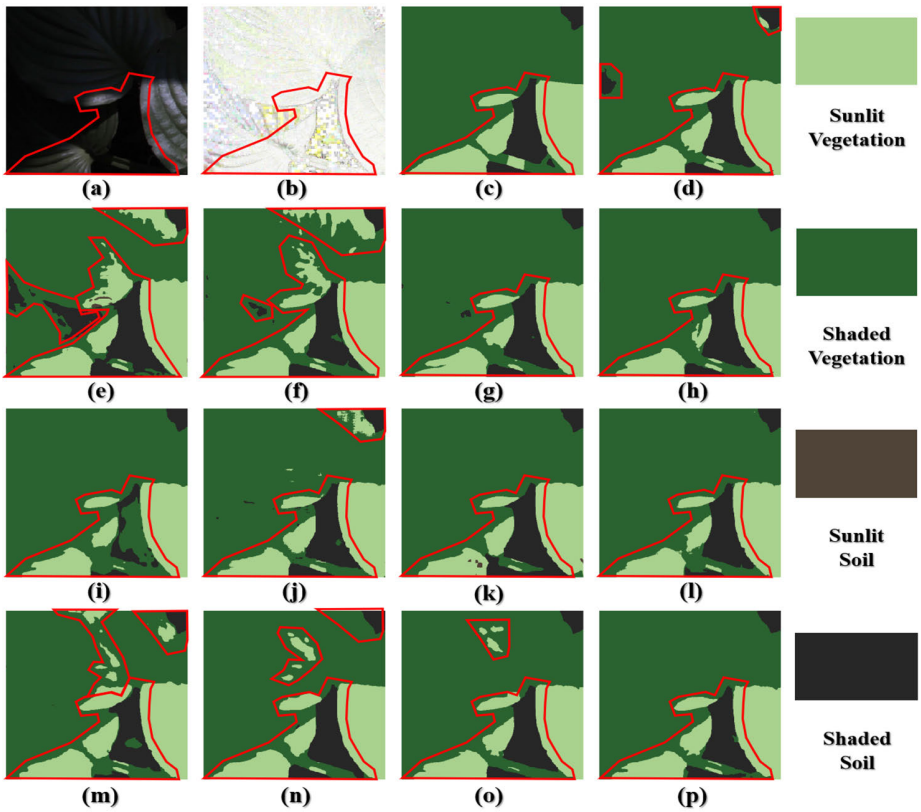

Fig. 11. Performance of the various networks on the vegetation polarization and light intensity datasets for shadow scene images. (a) Light intensity image. (b) DoLP image. (c) Ground truth. (d) Swin transformer (light intensity image). (e) Swin transformer (DoLP image). (f) Res2Net. (g) GhostNet V2. (h) ConvNeXt. (i) RMT. (j) MANet. (k) UNetFormer. (l) BEDSN. (m) CMX. (n) NLFNet. (o) FTransUNet. (p) SiMulitiF. An orange box was added to all the subgraphs to highlight differences.

E. Analysis of Multimodal Data of the Siamese Structure via the Normalization Method

One of the subjects of this study is the use of the weight-sharing method in the Siamese structure in the semantic segmentation task of multimodal data. We compared four different normalization methods, batch normalization [69], layer normalization [70], group normalization [71], and instance normalization [72], on the SiMultiF weight-sharing twin-branch encoder part of the vegetation polarization dataset with extreme natural lighting contrast. Through experimental analysis, when the Siamese structure is used, the selection of the network normalization method for multimodal data has a great effect on network accuracy. Batch normalization normalizes the data of each mini-batch, layer normalization normalizes all the channels of each sample, and group normalization is a technique that normalizes the values of each group channel of all the samples. The instance normalization normalizes every channel of every sample. When the batch normalization layer is used and the backpropagation technique is employed for modifying weights, the weights are updated in the two branches, and the direction of the weight update has a strong relationship with the mean and variance of the data in the batch in each branch. However, in the Siamese structure, when the data distributions of the two branches are inconsistent, the mean and variance are largely different, and convergence of the model is challenging. Moreover, the distinctions in the training mode and the validation mode of the batch normalization layer led to poor generalization model effectiveness. During the training process, in our Siamese structure, the weights are combined into two branches to take their respective mean values, and the variance is updated twice. In the verification mode, the two branches are calculated based on the statistical mean and variance, which are bound to be quite different from the real data distribution (when the two data distributions are quite different), leading to poor results, as shown in Fig. 12. As shown in Table IV, batch normalization is suitable for transformer networks, but it is not suitable for multimodal data in Siamese structure networks. This problem can be solved by changing the training strategy and method, but the process is more complicated. Replacing with other normalization methods is a relatively simple strategy, and through comparison, layer normalization is the method with the best accuracy in SiMultiF.

F. Ablation Study

At this point, we perform an extensive assessment of the enhancements introduced by the method we propose via experimental results. This process involves analyzing the effects of the network architecture, the adaptive modal weight adjustment module, the MFM, and the selection of multimodal data on the segmentation accuracy. This experiment was evaluated using the same experimental settings as the comparative experiment.

TABLE IV
ANALYSIS OF MULTIMODAL DATA OF THE SIAMESE STRUCTURE VIA THE NORMALIZATION METHOD

Method	Normalization	Data	F1(%)				OA (%)	mIoU (%)
			Sunlit Soil	Shaded Soil	Sunlit Vegetation	Shaded Vegetation		
Swin Transformer	BatchNorm	RGB	94.65	93.65	91.62	81.82	92.43	85.98
SiMultiF	BatchNorm	RGB,DoLP	69.71	54.84	55.54	61.12	60.14	43.43
	LayerNorm	RGB,DoLP	**95.43**	**94.31**	**93.30**	**91.67**	**93.67**	**88.14**
	GroupNorm	RGB,DoLP	88.51	86.31	86.38	83.59	86.17	75.79
	InstanceNorm	RGB,DoLP	95.15	94.17	92.68	91.20	93.30	87.48

TABLE V
ABLATION ANALYSIS OF EVERY PART OF SIMULTIF (%). THE VEGETATION POLARIZATION DATASET WITH EXTREME NATURAL LIGHTING CONTRAST IS ABBREVIATED AS POLARIZATION IN THE TABLE

AFWAM	Fusion method			mIoU (%)		Params (M)
	Add	Concat	MFM	Postdam	Polarization	
			√	77.47	87.49	42.30
√	√			77.16	87.29	30.09
√		√		77.22	87.32	33.06
√			√	78.76	88.14	42.62

TABLE VI
NETWORK STRUCTURE STUDIED BY ABLATION. THE VEGETATION POLARIZATION DATASET WITH EXTREME NATURAL LIGHTING CONTRAST IS ABBREVIATED AS POLARIZATION IN THE TABLE

Siamese structure	AFWAM	Fusion method		mIoU (%)		Params (M)
		Concat	MFM	Postdam	Polarization	
		√		76.69	86.53	68.89
	√		√	78.35	87.84	70.95
√	√		√	78.76	88.14	42.62

1) Conducting Ablation Experiments on Each Module: Validation experiments were performed by removing certain modules while keeping the Siamese architecture intact with the goal of validating the efficacy of each unique module within SiMultiF. As demonstrated in Table V, four studies involving ablation were designed based on our adaptive modal weight adjustment module and MFM. In the first experiment, we used only the MFM fusion scheme to segment the important features retrieved from both branches simultaneously. The experimental results highlighted the importance of our adaptive modal weight adjustment module. In the second experiment, we retain the AFWAM and choose the simple addition method as the fusion strategy. In the third experiment, we maintain the AFWAM method and choose the concatenation method as the fusion strategy. The fourth experiment involved all our innovative methods. The outcomes of the experiment demonstrate that the removal of each component leads to different degrees of accuracy loss, indicating the necessity of each component.

2) Network Structure Ablation Experiment: As shown in Table VI, to verify the applicability of the weight-sharing Siamese structure in SiMultiF with the semantic segmentation challenge of multimodal remote sensing images, we used a Siamese structure without weight sharing and kept other modules unchanged for ablation experiments. Moreover, compared with the Siamese structure with no weight sharing, we do

TABLE VII
DATA SELECTION ABLATION EXPERIMENTS OF THE SIMULTIF NETWORK ON THE ISPRS POTSDAM DATASET

Method	Data	F1(%)						OA (%)	mIoU (%)
		Imp.	Bui.	Low.	Tree.	Car.	Clu.		
Swin Transformer	RGB	90.98	95.36	83.37	82.13	94.14	63.35	88.29	74.20
SiMultiF	RGB, DSM	93.66	97.18	85.14	83.48	94.42	66.13	90.34	77.86
SiMultiF	IRGB, DSM	**93.89**	**97.40**	**85.77**	**84.42**	**95.01**	**67.14**	**91.02**	**78.76**

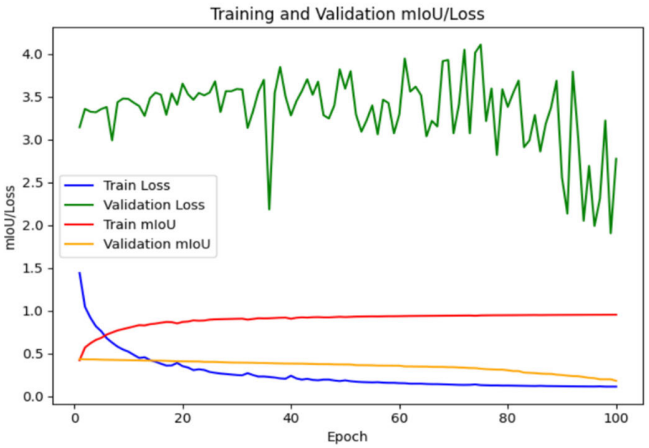

Fig. 12. Miou and loss of the training set and validation set of the SiMultiF network when batch normalization was used.

not add any modules and use only the concatenation method to fuse multimodal features in the decoder section to show how well our proposed structure is. The experimental results demonstrate that the weight-sharing Siamese structure in the Siamese structure is suitable for multiscale and multimodal remote sensing images. This architecture substantially lowers the quantity of network parameters and simplifies network complexity. When combined with our proposed modules, it achieves a low parameter count while maintaining the highest level of accuracy, making it the most efficient approach.

3) Selective Ablation Experiment With Multimodal Data: Data selection analysis was performed on the ISPRS Potsdam dataset. This analysis aimed to contrast the differences between IRRG and RGB in unimodal and multimodal segmentation tasks, as displayed in Table VII. The outcomes show that the NIRRG provides more information than the RGB does in semantic segmentation. This might be the case as the NIRRG can better characterize plants (trees and low vegetation). In addition, the use of DSM data significantly improved the overall segmentation performance. Especially for buildings and impervious surfaces, since these two types of ground objects usually have different elevation characteristics, the accuracy is significantly improved after the DSM data are added. This finding indicates that compared with single-modal data, multimodal data fusion technology is an important way to achieve performance improvement.

V. CONCLUSION

This article introduces a new approach based on Siamese architecture, which is tailored to allow semantic segmentation of multiscale and multimodal remote sensing images. Furthermore, we propose SiMultiF, which aims to solve the problem of the increase in the parameter doubling cost resulting from the increase in the number of modes through the weight-sharing dual-branch structure. The proposed AFWAM module can adaptively analyze the importance of multimodal features, dynamically adjust the modal weights, and preliminary perform modal feature complementarity. Moreover, the MFM module fully fuses the global semantic information and local detail information of different modalities based on CA to effectively represent the complementary information of different modalities. Furthermore, by using multiscale connections, we improve the use of multiscale features by bridging a link across the image encoder with a mask decoder and enabling precise object recognition across scales in multimodal remote sensing imagery. The effectiveness of all the components of SiMultiF was verified via ablation experiments. Experimental evaluations conducted on various scene-specific datasets reveal that the performance of the proposed SiMultiF model surpasses that of current SOTA segmentation methods, demonstrating its remarkable efficacy.

REFERENCES

[1] M. Zhang, Y. Zhou, J. Zhao, Y. Man, B. Liu, and R. Yao, "A survey of semi- and weakly supervised semantic segmentation of images," *Artif. Intell. Rev.*, vol. 53, no. 6, pp. 4259–4288, Aug. 2020.

[2] K. Zhu, N. N. Xiong, and M. Lu, "A survey of weakly-supervised semantic segmentation," in *Proc. IEEE IEEE 9th Intl. Conf. Big Data Secur. Cloud (BigDataSecurity) Intl. Conf. High Perform. Smart Comput., (HPSC) IEEE Intl. Conf. Intell. Data Secur. (IDS)*, May 2023, pp. 10–15.

[3] B. Li, Y. Shi, Z. Qi, and Z. Chen, "A survey on semantic segmentation," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2018, pp. 1233–1240.

[4] D. Feng, J. Wang, M. Fu, G. Liu, M. Zhang, and R. Tang, "Spatiotemporal variation and influencing factors of vegetation cover in the ecologically fragile areas of China from 2000 to 2015: A case study in Shaanxi province," *Environ. Sci. Pollut. Res.*, vol. 26, no. 28, pp. 28977–28992, Oct. 2019.

[5] P. Kumar et al., "PPSF: A privacy-preserving and secure framework using blockchain-based machine-learning for IoT-driven smart cities," *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 3, pp. 2326–2341, Jul. 2021.

[6] R. Li, S. Zheng, C. Duan, L. Wang, and C. Zhang, "Land cover classification from remote sensing images based on multi-scale fully convolutional network," *Geo-Spatial Inf. Sci.*, vol. 25, no. 2, pp. 278–294, Jan. 2022.

[7] 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.

[8] W. Guo, U. K. Rage, and S. Ninomiya, "Illumination invariant segmentation of vegetation for time series wheat images based on decision tree model," *Comput. Electron. Agricult.*, vol. 96, pp. 58–66, Aug. 2013.

[9] W. Xu et al., "Shadow detection and removal in apple image segmentation under natural light conditions using an ultrametric contour map," *Biosyst. Eng.*, vol. 184, pp. 142–154, Aug. 2019.

[10] X. Zhou et al., "MeSAM: Multiscale enhanced segment anything model for optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5623515.

[11] Y. Chen, L. Li, X. Liu, and X. Su, "A multi-task framework for infrared small target detection and segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5003109.

[12] Y. Cai, L. Fan, and Y. Fang, "SBSS: Stacking-based semantic segmentation framework for very high-resolution remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5600514.

[13] J. Chen, J. Zhu, Y. Guo, G. Sun, Y. Zhang, and M. Deng, "Unsupervised domain adaptation for semantic segmentation of high-resolution remote sensing imagery driven by category-certainty attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5616915.

[14] H. Shuai and Q. Liu, "Geometry-injected image-based point cloud semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5701510.

[15] J. Xu, R. Zhang, J. Dou, Y. Zhu, J. Sun, and S. Pu, "RPVNet: A deep and efficient range-point-voxel fusion network for LiDAR point cloud segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16004–16013.

[16] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12113–12132, Oct. 2023.

[17] S. Jabeen, X. Li, M. S. Amin, O. Bourahla, S. Y. Li, and A. Jabbar, "A review on methods and applications in multimodal deep learning," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 2, pp. 1–41, Jun. 2023.

[18] S. Sun, S. Zhi, Q. Liao, J. Heikkilä, and L. Liu, "Unbiased scene graph generation via two-stage causal modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12562–12580, Oct. 2023.

[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] N. Audebert, B. L. Saux, and S. Lefèvre, "Semantic segmentation of Earth observation data using multimodal and multi-scale deep networks," in *Proc. 13th ACCV*, Nov. 2017, pp. 180–196.

[21] Z. Li and H. Leung, "Fusion of multispectral and panchromatic images using a restoration-based method," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 5, pp. 1482–1491, May 2009.

[22] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, Jun. 2018.

[23] R. Fan, F. Li, W. Han, J. Yan, J. Li, and L. Wang, "Fine-scale urban informal settlements mapping by fusing remote sensing images and building data via a transformer-based multimodal fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5630316.

[24] L. Liu et al., "HierU-net: A hierarchical semantic segmentation method for land cover mapping," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4404614.

[25] S. Zhou et al., "DSM-assisted unsupervised domain adaptive network for semantic segmentation of remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5608216.

[26] L. H. Hughes, M. Schmitt, L. Mou, Y. Wang, and X. X. Zhu, "Identifying corresponding patches in SAR and optical images with a pseudo-Siamese CNN," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 784–788, May 2018.

[27] X. L. Huang et al., "UBC," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun.*, Aug. 2022, pp. 1412–1420.

[28] D. He, F. Abid, and J.-H. Kim, "Multimodal fusion and data augmentation for 3D semantic segmentation," in *Proc. 22nd Int. Conf. Control, Autom. Syst. (ICCAS)*, Nov. 2022, pp. 1143–1148.

[29] X. Hu, K. Yang, L. Fei, and K. Wang, "ACNET: Attention based network to exploit complementary features for RGBD semantic segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1440–1444.

[30] R. Su, J. Liu, D. Zhang, C. Cheng, and M. Ye, "Multimodal glioma image segmentation using dual encoder structure and channel spatial attention block," *Frontiers Neurosci.*, vol. 14, Oct. 2020, Art. no. 586197.

[31] Y. Liu, O. Yoshie, and H. Watanabe, "Application of multi-modal fusion attention mechanism in semantic segmentation," in *Proc. 16th ACCV*, Dec. 2023, pp. 378–397.

[32] C. Zhang, "Based on multi-feature information attention fusion for multi-modal remote sensing image semantic segmentation," in *Proc. 2021 IEEE ICMA*, Aug. 2021, pp. 71–76.

[33] H. Luo, X. Feng, B. Du, and Y. Zhang, "A multimodal feature fusion network for building extraction with very high-resolution remote sensing image and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5621819.

[34] 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.

[35] X. Ma, X. Zhang, M.-O. Pun, and M. Liu, "MSFNET: Multi-stage fusion network for semantic segmentation of fine-resolution remote sensing data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2022, pp. 2833–2836.

[36] 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, Feb. 2022, Art. no. 102638.

[37] 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.

[38] 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.

[39] E. Z. Xie, W. H. Wang, Z. D. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. 35th NeurIPS*, Dec. 2021, pp. 12077–12090.

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

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

[42] J. Li, B. Yang, L. Bai, H. Dou, C. Li, and L. Ma, "TFIV: Multigrained token fusion for infrared and visible image via transformer," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 2526414.

[43] 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. (CVPR)*, Jun. 2022, pp. 12176–12185.

[44] 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.

[45] Q. Wang, W. Chen, Z. Huang, H. Tang, and L. Yang, "MultiSenseSeg: A cost-effective unified multimodal semantic segmentation model for remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4703724.

[46] X. Ma, X. Xu, X. Zhang, and M.-O. Pun, "Adjacent-scale multimodal fusion networks for semantic segmentation of remote sensing data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 20116–20128, 2024.

[47] G. Li, Y. Wang, Z. Liu, X. Zhang, and D. Zeng, "RGB-T semantic segmentation with location, activation, and sharpening," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1223–1235, Mar. 2023.

[48] 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.

[49] C. Liu et al., "Context-aware network for semantic segmentation toward large-scale point clouds in urban environments," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5703915.

[50] X. Qiang, W. He, S. Chen, Q. Lv, and F. Huang, "Hierarchical point cloud transformer: A unified vegetation semantic segmentation model for multisource point clouds based on deep learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4411816.

[51] S. Saha, L. Mou, C. Qiu, X. X. Zhu, F. Bovolo, and L. Bruzzone, "Unsupervised deep joint segmentation of multitemporal high-resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8780–8792, Dec. 2020.
[52] L. Zhang, Z. Tan, G. Zhang, W. Zhang, and Z. Li, "Learn more and learn usefully: Truncation compensation network for semantic segmentation of high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4403814.
[53] X. Zhang, Z. Weng, P. Zhu, X. Han, J. Zhu, and L. Jiao, "ESDINet: Efficient shallow-deep interaction network for semantic segmentation of high-resolution aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5607615.
[54] W. J. Sun, J. Zhang, Y. J. Lei, and D. F. Hong, "RSProtoSeg: High spatial resolution remote sensing images segmentation based on non-learnable prototypes," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5626610.
[55] Z. Guo, R. Xu, C.-C. Feng, and Z. Zeng, "PIF-net: A deep point-image fusion network for multimodality semantic segmentation of very high-resolution imagery and aerial point cloud," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5700615.
[56] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR05)*, vol. 1, pp. 539–546.
[57] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[58] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.
[59] S. Y. Li, J. N. Jiao, and C. Wang, "Research on polarized multi-spectral system and fusion algorithm for remote sensing of vegetation status at night," *Remote Sens.*, vol. 13, no. 17, p. 24, Sep. 2021.
[60] J. R. Schott, *Fundamentals of Polarimetric Remote Sensing*. Bellingham, WA, USA: SPIE, 2009, pp. 1–244.
[61] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.
[62] Y. Tang, K. Han, J. Guo, C. Xu, C. Xu, and Y. Wang, "GhostNetV2: Enhance cheap operation with long-range attention," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2022, pp. 9969–9982.
[63] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11966–11976.
[64] Q. Fan, H. Huang, M. Chen, H. Liu, and R. He, "RMT: Retentive networks meet vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 5641–5651.
[65] R. Li et al., "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607713.
[66] L. Wang et al., "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 190, pp. 196–214, Aug. 2022.
[67] X. Li, L. Xie, C. Wang, J. Miao, H. Shen, and L. Zhang, "Boundary-enhanced dual-stream network for semantic segmentation of high-resolution remote sensing images," *GIScience Remote Sens.*, vol. 61, no. 1, Dec. 2024, Art. no. 2356355.
[68] R. Yan, K. Yang, and K. Wang, "NLFNet: Non-local fusion towards generalized multimodal semantic segmentation across RGB-depth, polarization, and thermal images," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Dec. 2021, pp. 1129–1135.
[69] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2015, pp. 448–456.
[70] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
[71] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jan. 2018, pp. 3–19.
[72] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*.

Shichao Cui received the B.S. degree in geographical information science from Liaocheng University, Liaocheng, China, in 2018. She is currently pursuing the Ph.D. degree in surveying science and technology with China University of Mining and Technology, Beijing, China.

Her research interests include remote sensing image semantic segmentation and multimodal fusion.

Wei Chen received the B.S. degree in geographical information system from Beijing Normal University, Beijing, China, in 2008, and the Ph.D. degree in photogrammetry and remote sensing from Peking University, Beijing, in 2013.

He is currently an Associate Professor with the College of Geoscience and Surveying Engineering, China University of Mining and Technology, Beijing. His research interests include radiometric calibration and remote sensing applications.

Wenwu Xiong received the B.S. degree from Southwest Petroleum University, Chengdu, China, in 2021. He is currently pursuing the Ph.D. degree in remote sensing science and technology with China University of Mining and Technology, Beijing, China.

His research interests include object detection in UAV imagery, and model lightweighting and deployment.

Xin Xu received the B.S. degree in resource exploration and the M.S. degree in geoinformation science from China University of Mining and Technology, Beijing, China, in 2020 and 2023, respectively.

He served as an AI Algorithm Engineer with China Siwei Surveying and Mapping Technology Company Ltd., Beijing, from 2023 to 2024. He is currently an Algorithm Engineer with Baidu, Beijing, focusing on multimodal technologies and AIGC-related research and development.

Xinyu Shi received the B.S. degree in geographical information science from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2020, and the M.S. degree in surveying and mapping engineering from China University of Mining and Technology, Beijing, China. She is currently pursuing the Ph.D. degree in photogrammetry and remote sensing with Wuhan University, Wuhan, China.

Her research interests include nighttime light radiative transfer and radiometric calibration.

Canhai Li received the master's degree in mapping and geographic information engineering from Wuhan University, Wuhan, China, in 2008.

He is currently a Senior Engineer with the Land Satellite Remote Sensing Application Center, MNR, Beijing, China. His research interests include multisource remote sensing image fusion and classification.