MANet: Fine-Tuning Segment Anything Model for Multimodal Remote Sensing Semantic Segmentation

Xianping Ma, Xiaokang Zhang, Man-On Pun, and Bo Huang

***Abstract*—Multimodal remote sensing data, collected from a variety of sensors, provide a comprehensive and integrated perspective of the Earth's surface. By employing multimodal fusion techniques, semantic segmentation offers more detailed insights into geographic scenes compared to single-modality approaches. Building upon recent advancements in vision foundation models, particularly the Segment Anything Model (SAM), this study introduces a novel Multimodal Adapter-based Network (MANet) for multimodal remote sensing semantic segmentation. At the core of this approach is the development of a Multimodal Adapter (MMAdapter), which fine-tunes SAM's image encoder to effectively leverage the model's general knowledge for multimodal data. In addition, a pyramid-based Deep Fusion Module (DFM) is incorporated to further integrate high-level geographic features across multiple scales before decoding. This work not only introduces a novel network for multimodal fusion, but also demonstrates, for the first time, SAM's powerful generalization capabilities with Digital Surface Model (DSM) data. Experimental results on two well-established fine-resolution multimodal remote sensing datasets, ISPRS Vaihingen and ISPRS Potsdam, confirm that the proposed MANet significantly surpasses current models in the task of multimodal semantic segmentation. The source code for this work will be accessible at https://github.com/sstary/SSRS.**

***Index Terms*—Multimodal fusion, Remote Sensing, Semantic Segmentation, Segment Anything Model, Adapter**

I. Introduction

Multimodal remote sensing semantic segmentation involves the process of classifying each pixel in remote sensing images using data from multiple sources or modalities, such as optical images, multispectral images, hyperspectral images, and LiDAR. By integrating diverse types of information, multimodal approaches enhance the accuracy and robustness of segmentation, particularly in complex environments [1, 2]. This technique leverages the complementary strengths of different data types to improve the identification of land objects, making it crucial for applications like land use and land cover [3, 4], environmental monitoring [5], and disaster management [6, 7]. In recent years, deep learning technologies have greatly prompted the development of multimodal fusion methods in remote sensing.

Initially, convolution neural networks (CNNs) were the dominant architecture, known for their ability to extract local spatial features from different modalities with the encoder-decoder framework [8]. These early CNN-based methods to stack multimodal data and fuse features at various stages for improved segmentation performance [9, 10]. With the introduction of the self-attention-based Transformer [11], which excel at modeling global context and long-range dependencies, many hybrid methods integrating Transformers into CNN-based methods have emerged. In particular, Vision Transformer (ViT) [12] and Swin Transformer [13] further introduced Transformer into the field of computer vision, greatly improving the ability of image feature extraction. Combining CNNs for detailed feature extraction with Transformers for capturing global relationships between different data sources marked a new stage in segmentation models [14, 15, 16]. This hybrid approach enhances the ability to fuse information across modalities and scales, leading to more accurate, robust segmentation results in complex scenes. Despite their good performance, the aforementioned models were trained solely on task-specific data, meaning they lack the broad, general knowledge that models like foundation models possess, which were trained on vast amounts of diverse data beyond the scope of a single task.

Interestingly, the recent emergence of large foundation models offers a solution to this challenge. In particular, the Segment Anything Model (SAM) [17] is a cutting-edge image segmentation model designed to tackle a wide range of segmentation tasks across diverse datasets. As presented in Fig. 1, SAM is comprised of three components, including a ViT-based image encoder, a prompt encoder and a mask decoder. Developed by Meta AI, SAM benefits from large-scale training on a vast visual corpus of natural images, enabling it to generalize effectively to unseen objects. Its versatility and robust performance position it as a valuable tool for various applications [18, 19]. However, compared to natural images, remote sensing images exhibit significant differences due to variations in sensors, resolution, spectral ranges, etc [20, 21, 22]. For multimodal tasks, non-optical data such as DSM further increases the discrepancy. This raises a crucial research problem: *How can the capabilities of SAM, acquired from massive natural visual corpora, be leveraged to enhance multimodal remote sensing tasks?*

Existing fine-tuning techniques, such as Adapter [23, 24, 25] and LoRA [26], address part of this challenge. Compared to training and full fine-tuning, these methods keep most

This work was supported in part by the National Natural Science Foundation of China under Grant 42371374 and 41801323, China Postdoctoral Science Foundation under grant 2020M682038 and the Shenzhen Science and Technology Innovation Committee under Grant No. JCYJ20190813170803617. *(Corresponding authors: Man-On Pun)*

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

Xiaokang Zhang is with the School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan 430081, China.(e-mail: natezhangxk@gmail.com).

Bo Huang is with the Department of Geography, The University of Hong Kong, Hong Kong, SAR 999077, China (e-mail: hbcuhk@gmail.com).

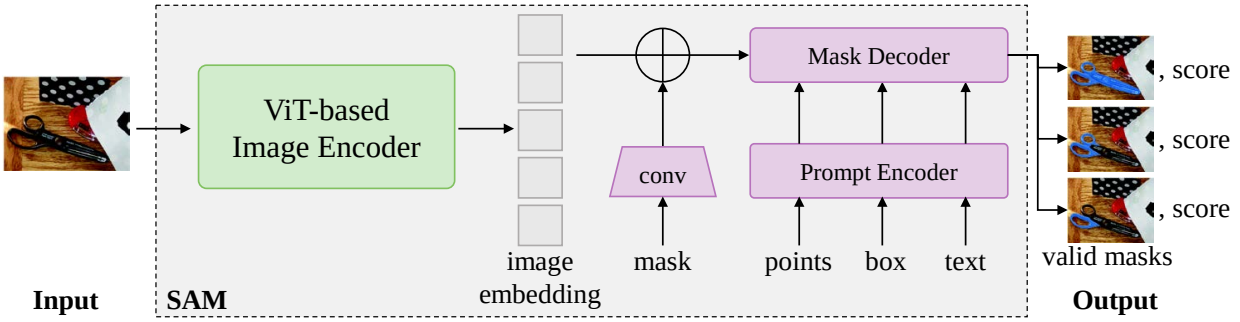

Fig. 1. Schematic of the SAM framework [17]. SAM is made up of three components, with its core image encoder built on ViT blocks.

of the parameters frozen and learns task-specific information with very fewer parameters. This enables parameter-efficient learning and the successful migration of large foundation models to a wider range of specific downstream tasks, even in more constrained hardware environments. The concept of Adapter was first introduced in the natural language processing community [23] as a method to fine-tune large pre-trained models for specific downstream tasks. The core idea behind Adapter is to introduce a parallel, compact, and scalable adaptation module that learns task-specific knowledge during training while the original model branch remains fixed, retaining task-agnostic knowledge. Then the synergy of task-specific and task-agnostic knowledge facilitates efficient fine-tuning of the large model.

In the field of remote sensing, fine-tuning techniques have focused on adapting SAM for single-modality tasks [27, 28]. CWSAM [27] and MeSAM [28] adapted SAM's image encoder and introduced custom mask decoders for remote sensing data. However, as shown in in Table I, an analysis of the SAM parameter distribution across three scales reveals that the majority of SAM's parameters are concentrated in its image encoder, suggesting that most of its general knowledge is encapsulated within this component. Therefore, we consider it unnecessary to modify SAM's prompt encoder and mask decoder for remote sensing-specific tasks, as such modifications increase model adaptation complexity while hindering seamless integration with existing segmentation models.

TABLE I
THE NUMBER OF PARAMETERS IN EACH PART OF THE SAM UNDER DIFFERENT BACKBONES.

Backbone	Image Encoder	Prompt Encoder	Mask Decoder
ViT-B	89.7M	6.2K	4.0M
ViT-L	308.3M	6.2K	4.0M
ViT-H	637.0M	6.2K	4.0M

To address the aforementioned challenges, we introduce a modified multimodal adapter, namely MMAdapter for multimodal fusion remote sensing tasks. Specifically, it consists of two fine-tuned branches that learn modality-specific information and achieve adaptive multimodal fusion with two weighting factors. Leveraging the MMAdapter, we propose a multimodal fusion framework, namely multimodal adapter-based network (MANet). The MANet utilizes a Deep Fusion Module (DFM) to perform multiscale processing and fusion of the deep high-level features from the SAM's image encoder. This is an effective solution for the complex characteristics of remote sensing images. Additionally, we employ a universal semantic segmentation decoder [29] that does not require additional task-specific design efforts, allowing for easy integration with decoders from other remote sensing tasks. The contributions of this work are threefold, as summarized in the following:

- A novel multimodal adapter, namely MMAdapter, is proposed to extract and fuse multimodal remote sensing features by fine-tuning SAM with few parameters, which reveals the powerful generalization capability of SAM on DSM for the first time.
- A novel multimodal fusion network, namely MANet, is proposed by capitalizing on SAM's image encoder and MMAdapter to perform remote sensing semantic segmentation. This is a streamlined and flexible adaptation framework that eliminates most of the redundant modules in SAM.
- Extensive experiments on two well-known fine-resolution multimodal remote sensing datasets, ISPRS Vaihingen and ISPRS Potsdam, confirm that the proposed MANet substantially outperforms existing methods.

The remainder of this paper is structured as follows. Sec. II provides a review of related works on multimodal fusion and SAM. Sec. III presents the structure of the proposed MANet, followed by a detailed description of the extensive experiments in Sec. IV. Finally, the conclusion is given in Sec. V.

II. Related Works

A. Multimodal Remote Sensing Semantic Segmentation

Semantic segmentation is a critical preprocessing step in remote sensing image processing, and leveraging multimodal information often yields better results than relying on a single modality. In recent years, the advent of deep learning has revolutionized the entire field of remote sensing, including semantic segmentation. Based on the classical encoder-decoder framework [8], numerous multimodal fusion approaches based on CNNs and Transformers have driven significant advancements in the field [30, 14, 15, 16]. ResUNet-a [31], an early architecture that based on CNNs, simply stacked multimodal data into four channels. vFuseNet [30] introduced a two-branch encoder to separately extract multimodal features, en-

abling deeper multimodal fusion through element-wise operations at the feature level. The introduction of Transformers [11, 12] has further enriched multimodal networks. For instance, CMFNet [14] employed CNNs for feature extraction and uses a Transformer structure to connect multimodal features across scales, emphasizing the importance of scale in multimodal fusion. Similarly, MFTransNet [16] used CNNs for feature extraction, while enhanced the self-attention module with spatial and channel attention for better feature fusion. FTransUNet [16] presented a multilevel fusion approach to refine the fusion of shallow- and deep-level remote sensing semantic features. Despite the great performance they achieved, we argue that existing models lack sufficient general knowledge, which poses a fundamental limitation to the advancement of multimodal fusion methods.

B. SAM in Remote Sensing

The Segment Anything Model (SAM) [17] holds a unique position as a general-purpose image segmentation model. This vision foundation model was trained on a very large visual corpus. It endows SAM with a remarkable ability to generalize to unseen objects, making it well-suited for applications in diverse scenarios. Nowadays, SAM has already been applied across various fields, such as autonomous driving [32], medical image processing [33] and remote sensing [34, 35, 36, 37]. In remote sensing, SAMRS [34] leveraged SAM to integrate numerous existing remote sensing datasets, using a new prompt named rotated bounding box. Furthermore, some recent works have considered fine-tuning SAM for remote sensing tasks such as semantic segmentation [27, 28, 38] and change detection [39, 40].

For single modality tasks, CWSAM [27] adapted SAM to synthetic aperture radar (SAR) images by introducing a task-specific input module and a class-wise mask decoder. MeSAM [28] incorporated an inception mixer into the SAM's image encoder to retain high-frequency features and introduced a multiscale-connected mask decoder for optical images. For multimodal tasks, RingMo-SAM [38] introduced a prompt encoder tailored to the characteristics of multimodal remote sensing data, along with a category-decoupling mask decoder. It is observed that these methods focus on designing task-specific prompts and mask decoders. However, as discussed in Sec. I, the general knowledge in SAM is primarily centered in the image encoder. While these approaches successfully utilize SAM's general knowledge to remote sensing, their complex architectures greatly hinder their adaptation to existing general semantic segmentation networks. Furthermore, there is currently no SAM-based multimodal approach designed for DSM data.

III. Proposed Method

To begin with, we first present the conventional single-modal Adapter in Sec. III-A before introducing the proposed MMAdapter in Sec. III-B. Building upon the proposed MMAdapter, the proposed MANet is elaborated in Sec. III-B.

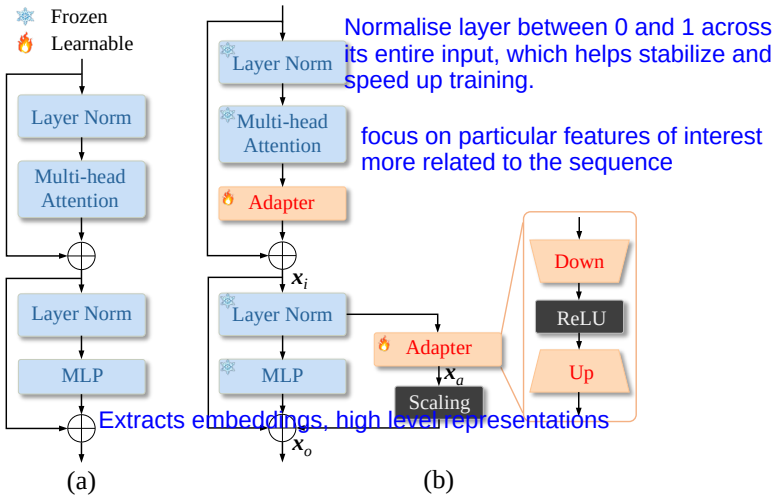

Fig. 2. (a) The ViT block *without* Adapter in SAM's image encoder and (b) the ViT block equipped with standard Adapter [41]. The Adapter facilitates the efficient utilization of general knowledge in specific tasks.

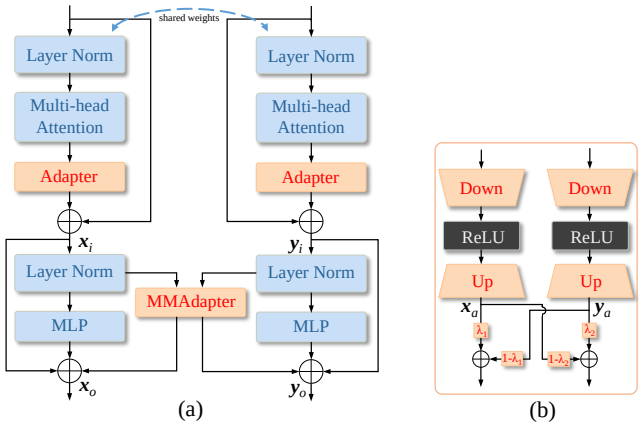

Fig. 3. (a) The ViT block endowed with the proposed MMAdapter and (b) the detailed structure of the MMAdapter. Compared to the standard Adapter, MMAadapter is characterized by two branches of shared weights for multimodal feature extraction. The standard Adapter and the proposed MMAdapter are used to fine-tune and fuse features, respectively.

A. The Standard Adapter

As discussed in Sec.I, SAM general knowledge is confined to its image encoder, specifically the ViT block, whose structure is shown in Fig.2(a). In [41], the Adapter was proposed to enhance the capabilities of ViT blocks for medical tasks through fine-tuning, as illustrated in Fig. 2(b). Instead of adjusting all parameters, the pre-trained SAM parameters remain frozen, while two Adapter modules are introduced to learn task-specific knowledge. Each Adapter consists of a down projection, a ReLU activation, and an up projection. The down projection compresses the input embedding to a lower dimension using a simple MLP layer, and the up projection restores the compressed embedding to its original dimension using another MLP layer. For a specific input feature $\boldsymbol{x}_i \in \mathbb{R}^{h\times w\times c}$, where h, w, and c represent the height, width, and channels of the input feature, the Adapter's process for generating the adapted feature can be expressed as:

$$\boldsymbol{x}_a = \mathrm{ReLU}\left(\mathrm{LN}(\boldsymbol{x}_i)\cdot \boldsymbol{W}_d\right)\cdot \boldsymbol{W}_u, \tag{1}$$

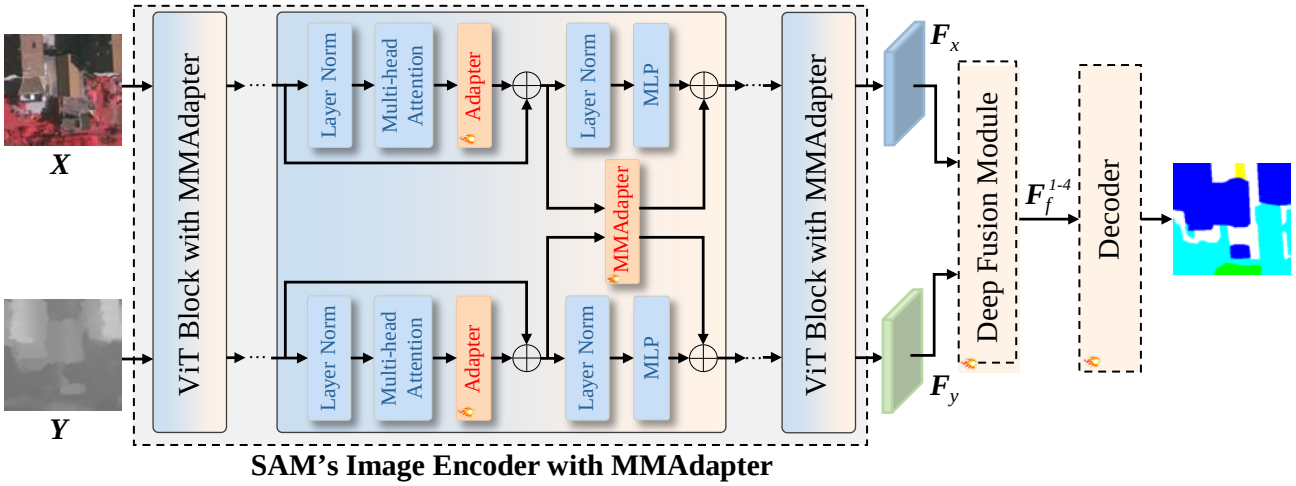

Fig. 4. Overview of the proposed MANet consisting of a SAM's image encoder with MMAdapter, a DFM and a general Decoder.

where $\boldsymbol{W}_d \in \mathbb{R}^{c\times\hat{c}}$ and $\boldsymbol{W}_u \in \mathbb{R}^{\hat{c}\times c}$ are the down projection and up projection, respectively. $\hat{c} \ll c$ is the compressed middle dimension of the Adapter. After that, both the adapted feature $\boldsymbol{x}_a$ and the output of the original MLP branch are fused with $\boldsymbol{x}_i$ by residual connection to generate the output feature $\boldsymbol{x}_o$:

$$\boldsymbol{x}_o = \text{MLP}\left(\text{LN}(\boldsymbol{x}_i)\right) + s \cdot \boldsymbol{x}_a + \boldsymbol{x}_i, \tag{2}$$

where s is a scaling factor to weight the task-specific and task-agnostic knowledge.

Since the Adapter proposed in [41] was designed for single-modal data, it is referred to as the standard Adapter in the sequel.

B. Multimodal Adapter (MMAdapter)

Next, we extend the standard Adapter to multimodal tasks. The proposed MMAdapter is the core multimodal fine-tuning technology of this work. As illustrated in Fig. 3(a), we employ dual branches with shared weights to process multimodal information. The Adapter after the Multi-head Attention is retained to extract the features of each modality independently, while the Adapter during the MLP stage is replaced with the proposed MMAdapter. The detailed structure of MMAdapter is shown in Fig. 3(b). While preserving the core structure of the Adapter, the MMAdapter enables modality interaction through the introduction of two weighting coefficients, λ_1 and λ_2. For two specific multimodal input features, $\boldsymbol{x}_i \in \mathbb{R}^{h\times w\times c}$ and $\boldsymbol{y}_i \in \mathbb{R}^{h\times w\times c}$, the process of generating the adapted features with MMAdapter can be described as:

$$\boldsymbol{x}_a = \text{ReLU}\left(\text{LN}(\boldsymbol{x}_n)\cdot\boldsymbol{W}_{dx}\right)\cdot\boldsymbol{W}_{ux}, \tag{3}$$

$$\boldsymbol{y}_a = \text{ReLU}\left(\text{LN}(\boldsymbol{y}_n)\cdot\boldsymbol{W}_{dy}\right)\cdot\boldsymbol{W}_{uy}, \tag{4}$$

where $\boldsymbol{W}_{dx}, \boldsymbol{W}_{dy} \in \mathbb{R}^{c\times\hat{c}}$ and $\boldsymbol{W}_{ux}, \boldsymbol{W}_{uy} \in \mathbb{R}^{\hat{c}\times c}$ are the down projections and up projections, respectively. After that, the multimodal output features $\boldsymbol{x}_o$ and $\boldsymbol{y}_o$ are generated using λ_1 and λ_2 as follows:

$$\boldsymbol{x}_o = \text{MLP}\left(\text{LN}(\boldsymbol{x}_i)\right) + \lambda_1 \cdot \boldsymbol{x}_a + (1-\lambda_1)\cdot\boldsymbol{y}_a + \boldsymbol{x}_i, \tag{5}$$

$$\boldsymbol{y}_o = \text{MLP}\left(\text{LN}(\boldsymbol{y}_i)\right) + \lambda_2 \cdot \boldsymbol{y}_a + (1-\lambda_2)\cdot\boldsymbol{x}_a + \boldsymbol{y}_i. \tag{6}$$

During the fine-tuning stage, only the newly added parameters are optimized, while the other parameters remain fixed. Detailed annotations are provided in Fig. 2(b), while other subfigures omit these annotations for clarity and conciseness. Finally, it is worth mentioning that the design shown in Fig. 3 can be generalized to more than two modalities in a straightforward manner.

C. The Proposed MANet

Fig. 4 shows an overview of the proposed MANet. The input to MANet is first processed by SAM's image encoder endowed with MMAdapter, responsible for extracting and fusing multimodal remote sensing features using the multimodal fine-tuning mechanism. The output is then fed into DFM, which receives two single-scale multimodal outputs from the encoder and expands them into two sets of multiscale multimodal features using pyramid modules. These high-level abstract features are then fused by four squeeze-and-excitation (SE) Fusion modules to generate a group of multiscale features. Finally, the outputs of DFM are passed to the decoder to produce segmentation prediction maps. In this section, we will elaborate on the key components of the proposed MANet.

- **SAM's image encoder**: We denote the optical images and their corresponding DSM data as $\boldsymbol{X} \in \mathbb{R}^{H\times W\times 3}$ and $\boldsymbol{Y} \in \mathbb{R}^{H\times W\times 1}$, respectively, where H and W represent the height and width of the inputs. The SAM's image encoder, employing a non-hierarchical ViT backbone, first embeds the input into a size of $\mathbb{R}^{h\times w\times c}$, where $h = \frac{H}{16}$, $w = \frac{W}{16}$, and c is the embedding dimension. Then stacked ViT Blocks is used to extract features. This feature size is maintained throughout the encoding process [42]. As illustrated in Fig. 4, both $\boldsymbol{X}$ and $\boldsymbol{Y}$ are input into the SAM's image encoder. It is important to note that the same SAM encoder is used for DSM data, which demonstrates that SAM can be utilized to extract features from non-image data. The SAM's image encoder extracts and fuses multimodal features, generating high-level abstract features $\boldsymbol{F}_x \in \mathbb{R}^{h\times w\times c}$ and $\boldsymbol{F}_y \in \mathbb{R}^{h\times w\times c}$ through the multimodal fine-tuning modules.

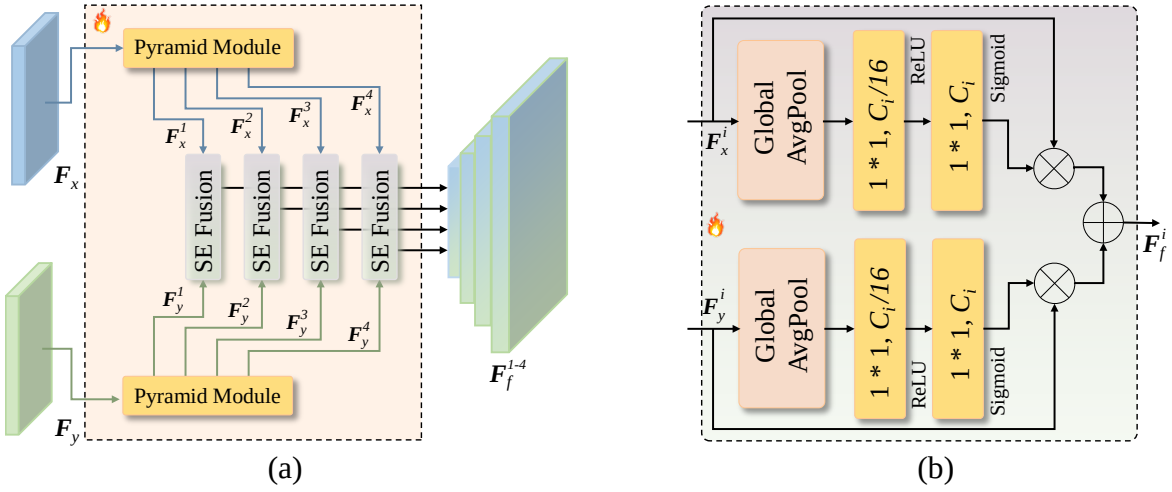

Fig. 5. (a) Schematic of the DFM. There are two Pyramid modules that expand multimodal features for multiscale features before they are fused by four SE Fusion modules. (b) Schematic of the SE Fusion module [16]. Notably, we just employ the existing simple fusion module and do not design this structure specifically, which proves that our primary gains are derived from the multimodal fine-tuning strategy based on the vision foundation model.

- **DFM**: Multiscale features play a critical role in semantic segmentation tasks since dense predicting requires both global and local information. As shown in Fig. 5(a), two pyramid modules, each consisting of a set of parallel convolutions or deconvolutions, are used to generate multiscale feature maps. Starting with the plain ViT feature map at a scale of $\frac{1}{16}$, we produce feature maps at scales of $\{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$ using convolutions with strides of $\{\frac{1}{4}, \frac{1}{2}, 1, 2\}$, where fractional strides indicate deconvolutions [42]. These simple pyramid modules generate two sets of multimodal multiscale features, denoted as $\boldsymbol{F}_x^i$ and $\boldsymbol{F}_y^i$, where $i = \{1, 2, 3, 4\}$ represents the scale index. Subsequently, four SE Fusion modules [16] are employed to further integrate the multimodal features. It is worth mentioning that more advanced fusion modules can yield further improved segmentation performance.
  As illustrated in Fig. 5(b), the SE Fusion module begins by aggregating the global information from the multimodal features. For the i-th fusion module, with an input channel size of C_i, the squeeze and excitation process is performed through two convolutional operations with a kernel size of 1×1, followed by the ReLU and Sigmoid activations. The multimodal outputs are then weighted and combined element-wise, producing the enhanced fused features denoted by $\boldsymbol{F}_f^i$. The outputs from the four SE Fusion modules form the multiscale fusion features, denoted as $\boldsymbol{F}_f^{1\text{-}4}$, which are fed into the decoder for processing. The decoder introduced in UNetformer [29] is employed in this work that reconstructs abstract semantic information into a segmented map by focusing on both global and local information.

IV. Experiments and Discussion

A. Datasets

Vaihingen: It contains 16 fine-resolution Orthophotos, each averaging 2500×2000 pixels. These Orthophotos consist of three channels: Near-Infrared, Red, and Green (NIRRG), and come with a normalized DSM at a 9 cm ground sampling distance. The 16 Orthophotos are split into a training set of 12 patches and a test set of 4 patches. To improve the storage and reading efficiency of large patches, a sliding window of size 256×256 is utilized rather than cropping the patches into smaller images in both training and test stages, which results in about 960 training images and 320 test images.

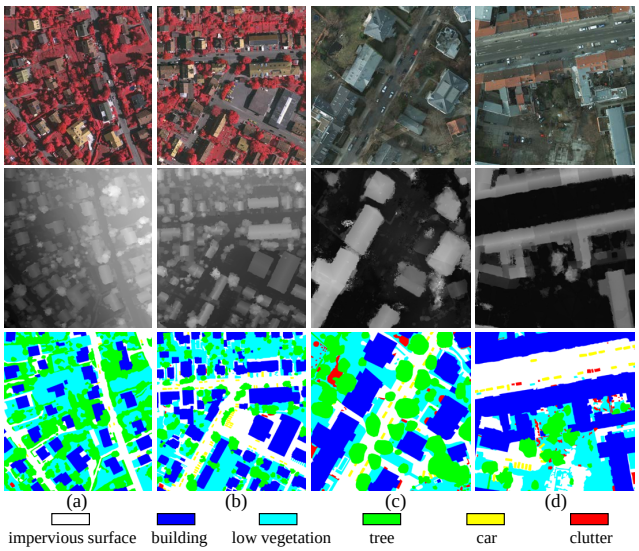

Fig. 6. Here we present four data samples of size 2048×2048 from Vaihingen (first two columns) and Potsdam (last two columns), respectively. The first row displays the orthophotos with NIRRG channels for Vaihingen and RGB channels for Potsdam. The second and third rows present the corresponding pixel-wise depth information and ground truth labels. They show the individual and complementary characteristics of remote sensing data from different sources.

Potsdam: This dataset is much larger than the Vaihingen dataset, which consists of 24 high-resolution Orthophotos, each with the size of 6000×6000 pixels. It includes four multispectral bands: Infrared, Red, Green, and Blue (IRRGB), along with a normalized DSM of 5 cm. The last three bands in this dataset are utilized to diversify our experiments. The

TABLE II
QUANTITATIVE RESULTS ON THE VAIHINGEN DATASET. THE BEST RESULTS ARE IN **BOLD**. THE SECOND BEST RESULTS ARE UNDERLINED. (%).

Method	Backbone	OA						mF1	mIoU
		Bui.	Tre.	Low.	Car	Imp.	Total		
FuseNet [9]	VGG16	96.28	90.28	78.98	81.37	91.66	90.51	87.71	78.71
vFuseNet [30]	VGG16	95.92	91.36	77.64	76.06	91.85	90.49	87.89	78.92
ABCNet [43]	ResNet18	94.10	90.81	78.53	64.12	89.70	89.25	85.34	75.20
MAResU-Net [44]	ResNet18	94.84	89.99	79.09	85.89	92.19	90.17	88.54	79.89
ESANet [45]	ResNet34	95.69	90.50	77.16	85.46	91.39	90.61	88.18	79.42
CMGFNet [46]	ResNet34	97.75	91.60	80.03	87.28	92.35	91.72	90.00	82.26
PSPNet [47]	ResNet101	94.52	90.17	78.84	79.22	92.03	89.94	86.55	76.96
SA-GATE [48]	ResNet101	94.84	92.56	81.29	87.79	91.69	91.10	89.81	81.27
CMFNet [14]	VGG16	97.17	90.82	80.37	85.47	92.36	91.40	89.48	81.44
UNetFormer [29]	ResNet18	96.23	91.85	79.95	86.99	91.85	91.17	89.48	81.97
MFTransNet [15]	ResNet34	96.41	91.48	80.09	86.52	92.11	91.22	89.62	81.61
TransUNet [49]	R50-ViT-B	96.48	92.77	76.14	69.56	91.66	90.96	87.34	78.26
FTransUNet [16]	R50-ViT-B	98.20	91.94	**81.49**	**91.27**	93.01	92.40	91.21	84.23
RS^{3}Mamba [50]	R18-Mamba-T	97.40	92.14	79.56	88.15	92.19	91.64	90.34	82.78
MANet	ViT-B	97.90	91.66	80.98	87.08	93.04	92.12	90.29	82.71
	ViT-L	**98.84**	93.17	81.16	89.23	93.39	92.93	91.51	84.72
	ViT-H	98.38	**93.94**	80.70	90.47	**93.59**	**92.97**	**91.71**	**85.03**

TABLE III
QUANTITATIVE RESULTS ON THE POTSDAM DATASET. THE BEST RESULTS ARE IN **BOLD**. THE SECOND BEST RESULTS ARE UNDERLINED. (%).

Method	Backbone	OA						mF1	mIoU
		Bui.	Tre.	Low.	Car	Imp.	Total		
FuseNet [9]	VGG16	97.48	85.14	87.31	96.10	92.64	90.58	91.60	84.86
vFuseNet [30]	VGG16	97.23	84.29	89.03	95.49	91.62	90.22	91.26	84.26
ABCNet [43]	ResNet18	96.23	78.92	86.40	92.92	88.90	87.52	88.14	79.26
MAResU-Net [44]	ResNet18	96.82	83.97	87.70	95.88	92.19	89.82	90.86	83.61
ESANet [45]	ResNet34	97.10	85.31	87.81	94.08	92.76	89.74	91.22	84.15
CMGFNet [46]	ResNet34	97.41	86.80	86.68	95.68	92.60	90.21	91.40	84.53
PSPNet [47]	ResNet101	97.03	83.13	85.67	88.81	90.91	88.67	88.92	80.36
SA-GATE [48]	ResNet101	96.54	81.18	85.35	**96.63**	90.77	87.91	90.26	82.53
CMFNet [14]	VGG16	97.63	87.40	88.00	95.68	92.84	91.16	92.10	85.63
UNetFormer [29]	ResNet18	97.69	86.47	87.93	95.91	92.27	90.65	91.71	85.05
MFTransNet [15]	ResNet34	97.37	85.71	86.92	96.05	92.45	89.96	91.11	84.04
TransUNet [49]	R50-ViT-B	96.63	82.65	89.98	93.17	91.93	90.01	90.97	83.74
FTransUNet [16]	R50-ViT-B	97.78	88.27	88.48	96.31	93.17	91.34	92.41	86.20
RS^{3}Mamba [50]	R18-Mamba-T	97.70	86.11	89.53	96.23	91.36	90.49	91.69	85.01
MANet	ViT-B	97.93	87.13	87.72	95.68	92.68	90.89	91.79	85.14
	ViT-L	98.31	**88.78**	87.27	96.29	**93.69**	91.62	92.51	86.37
	ViT-H	**98.44**	87.37	**90.36**	96.24	93.17	**91.71**	**92.70**	**86.69**

24 orthophotos are split into 18 patches for training and 6 for testing. Using the same sliding window approach, this dataset contains 10368 training samples and 3456 test samples.

The Vaihingen and Potsdam datasets classify five main foreground categories: Building (Bui.), Tree (Tre.), Low Vegetation (Low.), Car, and Impervious Surface (Imp.). Additionally, a background class labeled as Clutter contains indistinguishable debris and water surfaces. Visual examples from both datasets are presented in Fig. 6. Notably, the sliding window moves with a smaller step size, and the overlapping areas are averaged to the reduced boundary effect during the test stage.

B. Implementation details

All experiments were conducted using PyTorch on a single NVIDIA GeForce RTX 3090 GPU with 24GB of RAM. The stochastic gradient descent algorithm was used to train all the models under consideration with a learning rate of 0.01, momentum of 0.9, weight decay of 0.0005, and a batch size of 10. The batch size is reduced to 6 when ViT-H is applied to meet the memory limit. Basic data augmentation techniques, including random rotation and flipping, are applied after sample collection with the sliding window.

To assess the semantic segmentation performance on multi-modal remote sensing data, we use overall accuracy (OA), mean F1 score (mF1), and mean intersection over union (mIoU). These standard metrics enable a fair comparison between the proposed MANet and other state-of-the-art methods. Specifically, OA evaluates both foreground classes and the background class, while mF1 and mIoU are calculated for the five foreground classes.

C. Performance Comparison

We benchmarked the proposed MANet against fourteen state-of-the-art methods, including ABCNet [43], PSPNet [47], MAResU-Net [44], vFuseNet [30], FuseNet [9], ESANet [45], SA-GATE [48], CMGFNet [46], TransUNet [49], CMFNet [14], UNetFormer [29], MFTransNet [15], FTransUNet [16], and RS^{3}Mamba [50], most of which were specifically designed for remote sensing tasks. In our experiments, ABCNet, PSPNet, MAResU-Net, UNetFormer, and RS^{3}Mamba utilized

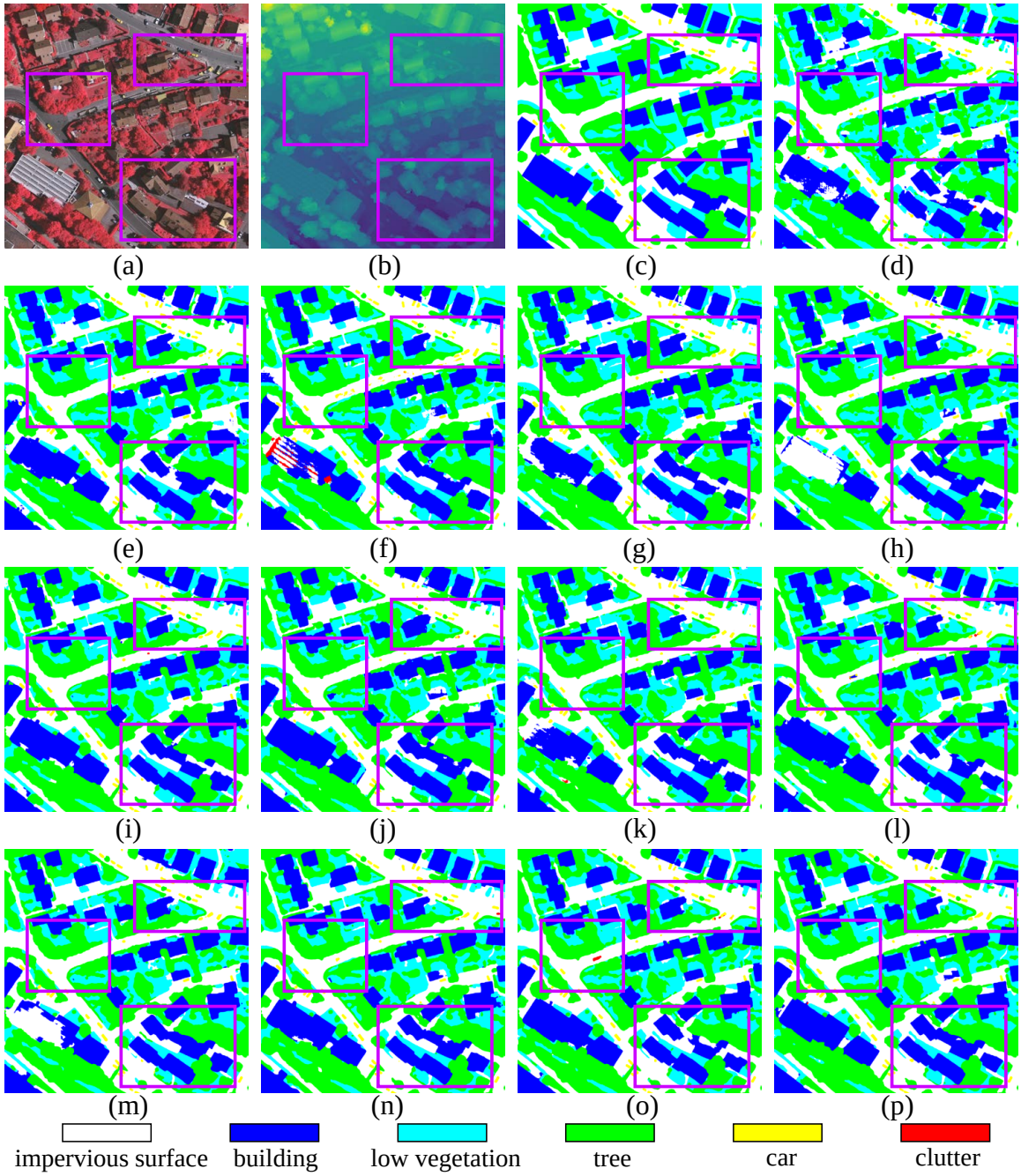

Fig. 7. Visualized comparisons on the Vaihingen test set with the size of 1800 × 1800. (a) NIRRG images, (b) DSM, (c) Ground Truth, (d) MAResU-Net, (e) vFuseNet, (f) FuseNet, (g) ESANet, (h) SA-GATE, (i) CMGFNet, (j) TransUNet, (k) CMFNet, (l) UNetFormer, (m) MFTransNet, (n) FTransUNet, (o) $\mathrm{RS}^3\mathrm{Mamba}$, (p) The proposed MANet. Purple boxes are added to highlight the differences.

only the optical images, to highlight the impact of DSM data and demonstrate the advantages of multimodal methods over single-modal ones. The other methods are state-of-the-art multimodal models based on different network architectures, including CNN and Transformer. The comparative quantitative results are presented in Table II and Table III.

1) Performance Comparison on the Vaihingen dataset: As presented in Table II, the proposed MANet demonstrated substantial improvements in terms of OA, mF1 and mIoU metrics compared to existing segmentation methods. These results confirmed that our MANet can effectively leverage the extensive general knowledge embedded in SAM. In particular, MANet outperformed state-of-the-art models across three specific classes, namely *Building*, *Tree* and *Impervious surface*. In terms of the overall performance, the proposed MANet achieved an OA of 92.97%, an mF1 score of 91.71% and a mIoU of 85.03%, reflecting gains of 0.57%, 0.50% and 0.80% over the second best method FTransUNet, respectively. Moreover, each of the three MANet backbone variants offers unique advantages. Even the smallest variant, ViT-B, is comparable to most methods, further validating that our multimodal fine-tuning approach can efficiently utilize

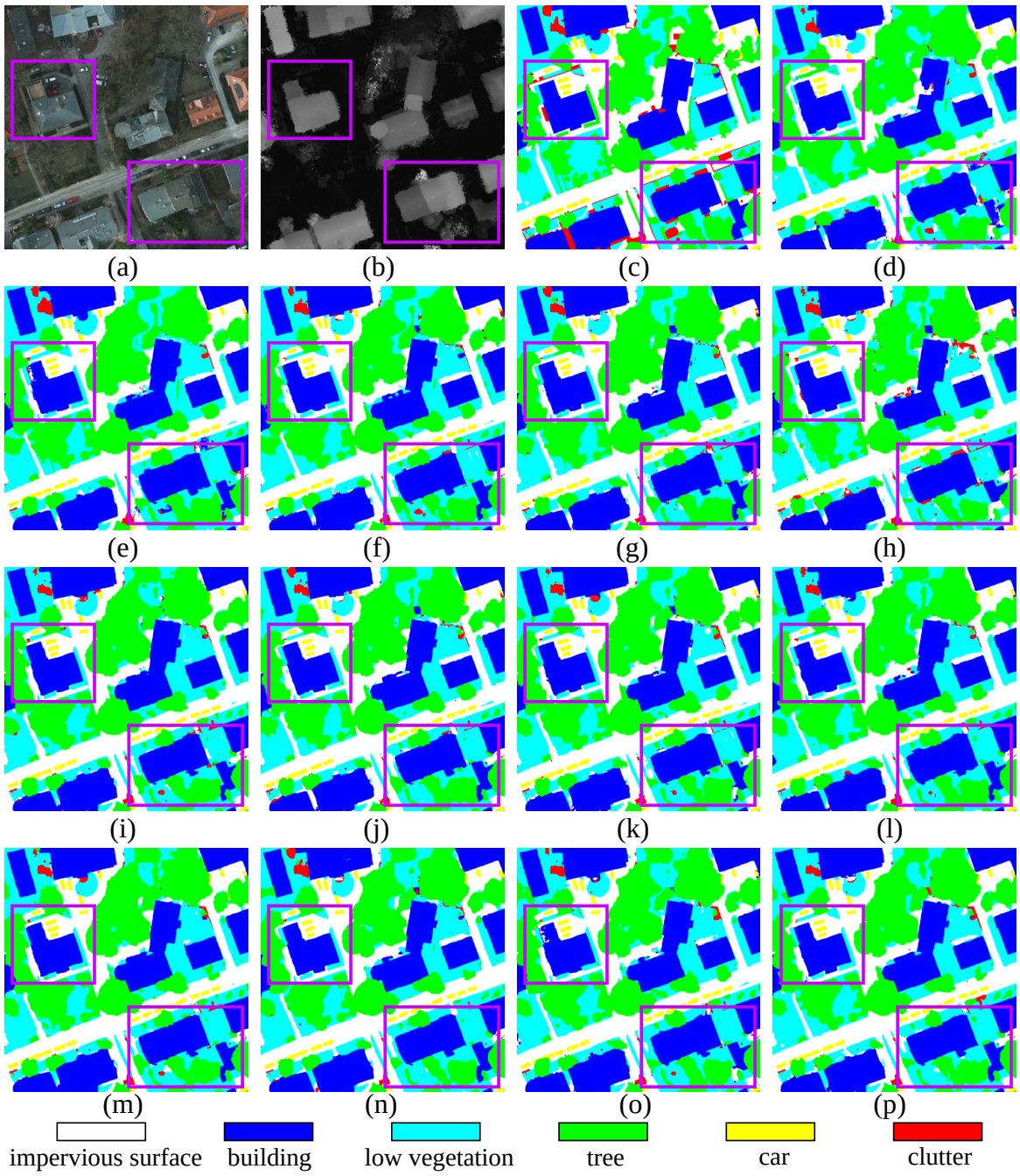

Fig. 8. Visualized comparisons on the Potsdam test set with the size of 2000 × 2000. (a) IRRG images, (b) DSM, (c) Ground Truth, (d) MAResU-Net, (e) vFuseNet, (f) FuseNet, (g) ESANet, (h) SA-GATE, (i) CMGFNet, (j) TransUNet, (k) CMFNet, (l) UNetFormer, (m) MFTransNet, (n) FTransUNet, (o) RS^{3}Mamba, (p) The proposed MANet. Purple boxes are added to highlight the differences.

general knowledge from SAM to assist with the semantic segmentation of multimodal remote sensing data. This result demonstrated the practical value of the proposed MANet and the MMAdapter in guiding the introduction of foundation models, like SAM, into multimodal remote sensing tasks.

Fig. 7 presents a visual comparison of the results produced by various methods and the proposed MANet. MANet demonstrates superior performance in generating sharper and more precise boundaries for ground objects, such as trees, cars, and buildings, resulting in clearer separations. This also helps preserve the integrity of the ground objects. Overall, the visualizations produced by MANet exhibit a cleaner and more organized appearance. We attribute these improvements primarily to SAM's powerful feature extraction capabilities. By applying the multimodal fine-tuning mechanism, SAM's ability to segment every natural element is effectively extended to ground objects.

2) Performance Comparison on the Potsdam dataset: Experiments on the Potsdam dataset yielded results consistent with those from the Vaihingen dataset. As shown in Table III, the corresponding OA, mF1 and mIoU values were 91.71%, 92.70%, 86.69% respectively, which corresponds to

TABLE IV
QUANTITATIVE RESULTS ON THE VAIHINGEN DATASET WITH DIFFERENT MODALITIES AND FINE-TUNING MECHANISM. THE BEST RESULTS ARE IN **BOLD**. (%).

Modality	Fine-tuning	OA						mF1	mIoU
		Bui.	Tre.	Low.	Car	Imp.	Total		
NIIRG	The standard Adapter	96.29	93.09	80.15	89.08	92.59	92.02	90.94	83.69
NIIRG + DSM	The standard Adapter	98.74	91.98	**82.17**	88.59	93.25	92.73	91.23	84.25
NIIRG + DSM	MANet with MMAdapter	**98.84**	**93.17**	81.16	**89.23**	**93.39**	**92.93**	**91.51**	**84.72**

increases of 0.37%, 0.29% and 0.49%, respectively, over FTransUNet. Notably, significant gains were observed for *Building*, *Tree* and *Low Vegetation* compared to other state-of-the-art methods. The MANet with smaller backbones also shows superior performance. This flexibility allows MANet to balance hardware requirements and performance needs across various application scenarios.

Fig. 8 shows a visualization example from Potsdam, where we observed more defined boundaries and intact object representations. These visual improvements are consistent with the mF1 and mIoU metrics shown in TableIII. Undoubtedly, it further validates the practical applicability of the proposed MANet and MMAdapter.

In additional, it is also observed that MANet is hard to achieve the best results simultaneously in identifying *Tree* and *Low Vegetation*. This challenge arises because both categories are characterized by irregular boundaries. Furthermore, their similarity, as well as the staggered or overlapping distribution, make them difficult to be distinguished from each other. Combining SAM with more specialized design to tackle the identification of these challenging categories presents an interesting future direction.

D. Modality Analysis

To illustrate the necessity of multimodal fine-tuning, we conducted a modality analysis, with the results presented in Table IV. In the first experiment, we only used single-modality data and applied the standard Adapter mechanism to fine-tune the SAM's image encoder. This experiment highlights the importance and need for multimodal information. In the second experiment, the SAM's image encoder retained the standard Adapters but excluded the proposed MMAdapter. Therefore, this experiment could still extract remote sensing multimodal features independently, but it lacked crucial information fusion during the encoding process.

Inspection of Table IV reveals the significant effectiveness of incorporating multimodal information. The improvement is particularly pronounced in the categories of *Building* and *Impervious surface*, which tend to have significant and stable surface elevation information. Additionally, this enhancement strengthens the model's ability to distinguish other categories. Overall, the integration of multimodal information provides comprehensive benefits across the board. The introduction of the MMAdapter enables more effective utilization of DSM information, significantly enhancing the model's ability to extract and fuse multimodal information. As a result, the semantic segmentation performance is further improved.

E. Ablation Study

The proposed MANet consists of two core components: the SAM's image encoder with MMAdapter and the DFM. To validate their effectiveness, ablation experiments were conducted by systematically removing specific components. As shown in Table V, two ablation experiments were designed. In the first experiment, the DFM was removed from MANet, leading to a lack of deep analysis and integration of high-level abstract remote sensing semantic features. The second experiment follows the setup of the second experiments in Table IV.

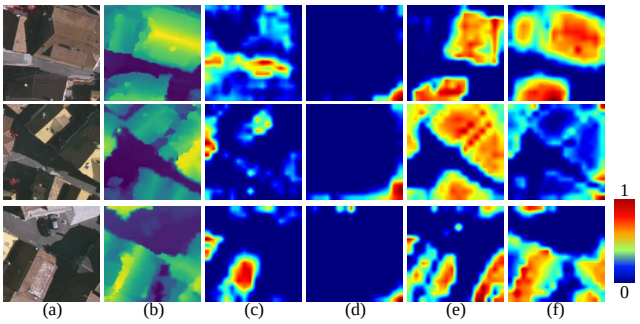

Fig. 9. Three groups of heatmaps. (a) NIIRG images, (b) DSM, (c) heatmaps from NIIRG images and (d) heatmaps from DSM both generated by the original SAM, (e) heatmaps from NIIRG images and (f) heatmaps from DSM both generated by the proposed MMAdapter. The high-value areas in the heatmaps indicate objects identified as buildings by the methods. The effectiveness of our MMAdapter can be clearly observable.

It is important to note that removing all Adapters from the SAM's image encoder severely degrades the model's performance, as confirmed in Fig. 9, which also illustrates the effectiveness of the multimodal fine-tuning mechanism. Inspection of Fig. 9(c) and (e) reveal that SAM, without fine-tuning, cannot extract meaningful features from remote sensing data, rendering it unsuitable for semantic segmentation tasks. However, inspection of Fig. 9(d) and (f) shows that, after fine-tuning with the MMAdapter, the heatmaps change dramatically. Furthermore, Fig. 9(e) clearly demonstrates that SAM, despite being trained on RGB optical images, is also effective when applied to non-optical DSM data. It is observed that DSM can effectively provide supplementary information. Therefore, the fine-tuned SAM's image encoder is capable of recognizing and segmenting remote sensing objects effectively in multimodal tasks.

Inspection of Table V indicates that both the MMAdapter and DFM are essential for enhancing the performance of the proposed MANet. Specifically, the MMAdapter facilitates continuous information fusion, allowing for extraction and fusion

TABLE V
ABLATION STUDY OF THE PROPOSED MANET. THE BEST RESULTS ARE IN **BOLD**.

MMAdapter	DFM	OA(%)	mF1(%)	mIoU(%)
✓		92.73	91.23	84.25
	✓	92.80	91.30	84.35
✓	✓	**92.93**	**91.51**	**84.72**

of multimodal information as the encoding depth increases. The DFM verifies the importance of high-level features in the semantic segmentation of remote sensing data. In this work, we primarily introduce a new framework for leveraging SAM, rather than emphasize the high-level feature fusion techniques. Replacing DFM with a more advanced fusion model is expected to result in further performance improvement.

F. Model Scale Analysis

The improved performance of MANet is largely attributable to the general knowledge provided by the vision foundation model, SAM, which is also a large model. However, the large model does not have advantages in terms of computational complexity or inference speed compared to existing general methods. Consequently, we focus on reporting the model's *trainable* parameter number and memory footprint to measure its hardware requirements.

Table VI presents the results of the model scale for all methods compared in this work. As indicated in Table VI, the proposed *parameter-efficient* fine-tuning technique allows the large foundation models to be used on a single GPU while maintaining a manageable number of trainable parameters and memory costs. In our experiments, we successfully fine-tuned the ViT-L backbone on the same hardware with the same hyperparameters, achieving results that surpassed all existing methods. For the ViT-H backbone, we adjusted the batch size from 10 to 6 due to the GPU memory limitation. This reduction in batch size did not degrade performance, but further improved performance. These results prove the powerful feature extraction and fusion capabilities of the large vision foundation model. This work also offers valuable insights for exploring multimodal tasks with large models under constrained hardware conditions.

V. CONCLUSION

In this study, a novel multimodal fusion framework called MANet has been proposed for semantic segmentation of multimodal remote sensing data by leveraging the general knowledge embedded in the vision foundation model, SAM. Specifically, the SAM's image encoder is equipped with a novel MMAdapter that is utilized to extract and fuse multimodal remote sensing features. The fused deep features are then further exploited by a pyramid-based DFM before being reconstructed into segmentation maps. Comprehensive experiments on two multimodal datasets, ISPRS Vaihingen and ISPRS Potsdam, have demonstrated that MANet outperforms current state-of-the-art segmentation methods. This research is the first to confirm the reliability of SAM on DSM data and offers a promising solution for applying vision foundation models to multimodal data. Furthermore, it has the potential for extension to other remote sensing tasks, such as semi-supervised or unsupervised learning.

TABLE VI
MODEL SCALE ANALYSIS MEASURED BY A 256×256 IMAGE ON A SINGLE NVIDIA GEFORCE RTX 3090 GPU. MIOU VALUES ARE THE RESULTS ON THE VAIHINGEN DATASET. THE BEST RESULTS ARE IN **BOLD**.

Method	Parameter (M)	Memory (MB)	MIoU (%)
ABCNet [43]	**13.39**	1598	75.20
PSPNet [47]	46.72	3124	76.96
MAResU-Net [44]	26.27	1908	79.89
UNetFormer [29]	24.20	1980	81.97
RS^{3}Mamba [50]	43.32	1548	82.78
TransUNet [49]	93.23	3028	78.26
FuseNet [9]	42.08	2284	78.71
vFuseNet [30]	44.17	2618	78.92
ESANet [45]	34.03	1914	79.42
SA-GATE [48]	110.85	3174	81.27
CMFNet [14]	123.63	4058	81.44
MFTransUNet [15]	43.77	**1549**	81.61
CMGFNet [46]	64.20	2463	82.26
FTransUNet [16]	160.88	3463	84.23
MANet (ViT-B)	20.43	2274	82.71
MANet (ViT-L)	56.67	5514	84.36
MANet (ViT-H)	111.28	8934	**85.03**

REFERENCES

[1] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1560–1584, 2015.

[2] J. Li, D. Hong, L. Gao, J. Yao, K. Zheng, B. Zhang, and J. Chanussot, "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102926, 2022.

[3] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[4] D. Hong, J. Hu, J. Yao, J. Chanussot, and X. X. Zhu, "Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 178, pp. 68–80, 2021.

[5] P. Karmakar, S. W. Teng, M. Murshed, S. Pang, Y. Li, and H. Lin, "Crop monitoring by multimodal remote sensing: A review," *Remote Sensing Applications: Society and Environment*, p. 101093, 2023.

[6] N. Algiriyage, R. Prasanna, K. Stock, E. E. Doyle, and D. Johnston, "Multi-source multimodal data and deep learning for disaster response: a systematic review," *SN Computer Science*, vol. 3, pp. 1–29, 2022.

[7] X. Zhang, W. Yu, M.-O. Pun, and W. Shi, "Cross-domain landslide mapping from large-scale remote sensing images using prototype-guided domain-aware progressive representation learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 197, pp. 1–17, 2023.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.

[9] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Asian conference on computer vision*, 2016, pp. 213–228.

[10] X. Zhang, W. Yu, and M.-O. Pun, "Multilevel deformable attention-aggregated networks for change detection in bitemporal remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, pp. 1–11, 2017.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," *International Conference on Learning Representations*, pp. 1–22, 2021.

[13] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[14] X. Ma, X. Zhang, and M.-O. Pun, "A crossmodal multiscale fusion network for semantic segmentation of remote sensing data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 3463–3474, 2022.

[15] 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, p. 722, 2023.

[16] X. Ma, X. Zhang, M.-O. Pun, and M. Liu, "A multilevel multimodal fusion transformer for remote sensing semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[17] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[18] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[19] H. Wang, P. K. A. Vasu, F. Faghri, R. Vemulapalli, M. Farajtabar, S. Mehta, M. Rastegari, O. Tuzel, and H. Pouransari, "SAM-clip: Merging vision foundation models towards semantic and spatial understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3635–3647.

[20] Y. Li, H. Zhang, X. Xue, Y. Jiang, and Q. Shen, "Deep learning for remote sensing image classification: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 6, p. e1264, 2018.

[21] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4096–4105.

[22] X. Ma, X. Zhang, Z. Wang, and M.-O. Pun, "Unsupervised domain adaptation augmented by mutually boosted attention for semantic segmentation of vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[23] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in *International conference on machine learning*. PMLR, 2019, pp. 2790–2799.

[24] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, "Adaptformer: Adapting vision transformers for scalable visual recognition," vol. 35, 2022, pp. 16 664–16 678.

[25] X. He, C. Li, P. Zhang, J. Yang, and X. E. Wang, "Parameter-efficient model adaptation for vision transformers," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 817–825.

[26] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[27] X. Pu, H. Jia, L. Zheng, F. Wang, and F. Xu, "Classwise-SAM-adapter: Parameter efficient fine-tuning adapts segment anything to sar domain for semantic segmentation," *arXiv preprint arXiv:2401.02326*, 2024.

[28] X. Zhou, F. Liang, L. Chen, H. Liu, Q. Song, G. Vivone, and J. Chanussot, "MeSAM: Multiscale enhanced segment anything model for optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, p. 5623515, 2024.

[29] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 196–214, 2022.

[30] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 140, pp. 20–32, 2018.

[31] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020.

[32] D. Li, W. Xie, M. Cao, Y. Wang, J. Zhang, Y. Li, L. Fang, and C. Xu, "FusionSAM: Latent space driven segment anything model for multimodal fusion and segmentation," *arXiv preprint arXiv:2408.13980*, 2024.

[33] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment anything model for medical image analysis: an experimental study," *Medical Image Analysis*, vol. 89, p. 102918, 2023.

[34] D. Wang, J. Zhang, B. Du, M. Xu, L. Liu, D. Tao, and L. Zhang, "SAMRS: Scaling-up remote sensing segmentation dataset with segment anything model," in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[35] Z. Qi, C. Liu, Z. Liu, H. Chen, Y. Wu, Z. Zou, and Z. Shi, "Multi-View remote sensing image segmentation with SAM priors," in *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, 2024, pp. 8446–8449.

[36] H. Chen, J. Song, and N. Yokoya, "Change detection between optical remote sensing imagery and map data via segment anything model (SAM)," *arXiv preprint arXiv:2401.09019*, 2024.

[37] X. Ma, Q. Wu, X. Zhao, X. Zhang, M.-O. Pun, and B. Huang, "SAM-assisted remote sensing imagery semantic segmentation with object and boundary constraints," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024.

[38] Z. Yan, J. Li, X. Li, R. Zhou, W. Zhang, Y. Feng, W. Diao, K. Fu, and X. Sun, "RingMo-SAM: A foundation model for segment anything in multimodal remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[39] L. Ding, K. Zhu, D. Peng, H. Tang, K. Yang, and L. Bruzzone, "Adapting segment anything model for change detection in VHR remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–11, 2024.

[40] L. Mei, Z. Ye, C. Xu, H. Wang, Y. Wang, C. Lei, W. Yang, and Y. Li, "SCD-SAM: Adapting segment anything model for semantic change detection in remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[41] J. Wu, W. Ji, Y. Liu, H. Fu, M. Xu, Y. Xu, and Y. Jin, "Medical SAM adapter: Adapting segment anything model for medical image segmentation," *arXiv preprint arXiv:2304.12620*, 2023.

[42] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring plain vision transformer backbones for object detection," in *European conference on computer vision*. Springer, 2022, pp. 280–296.

[43] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M.

Atkinson, "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery," *ISPRS journal of photogrammetry and remote sensing*, vol. 181, pp. 84–98, 2021.

[44] R. Li, S. Zheng, C. Duan, J. Su, and C. Zhang, "Multistage attention resu-net for semantic segmentation of fine-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[45] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, and H.-M. Gross, "Efficient rgb-d semantic segmentation for indoor scene analysis," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13 525–13 531.

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

[47] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[48] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, and G. Zeng, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation," in *European Conference on Computer Vision*, 2020, pp. 561–577.

[49] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[50] X. Ma, X. Zhang, and M.-O. Pun, "RS3Mamba: Visual state space model for remote sensing image semantic segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.