A Multilevel Multimodal Fusion Transformer for Remote Sensing Semantic Segmentation

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

***Abstract*—Accurate semantic segmentation of remote sensing data plays a crucial role in the success of geoscience research and applications. Recently, multimodal fusion-based segmentation models have attracted much attention due to their outstanding performance as compared to conventional single-modal techniques. However, most of these models perform their fusion operation using convolutional neural networks (CNN) or the vision transformer (Vit), resulting in insufficient local-global contextual modeling and representative capabilities. In this work, a multilevel multimodal fusion scheme called FTransUNet is proposed to provide a robust and effective multimodal fusion backbone for semantic segmentation by integrating both CNN and Vit into one unified fusion framework. Firstly, the shallow-level features are first extracted and fused through convolutional layers and shallow-level feature fusion (SFF) modules. After that, deep-level features characterizing semantic information and spatial relationships are extracted and fused by a well-designed Fusion Vit (FVit). It applies Adaptively Mutually Boosted Attention (Ada-MBA) layers and Self-Attention (SA) layers alternately in a three-stage scheme to learn cross-modality representations of high inter-class separability and low intra-class variations. Specifically, the proposed Ada-MBA computes SA and Cross-Attention (CA) in parallel to enhance intra- and cross-modality contextual information simultaneously while steering attention distribution towards semantic-aware regions. As a result, FTransUNet can fuse shallow-level and deep-level features in a multilevel manner, taking full advantage of CNN and transformer to accurately characterize local details and global semantics, respectively. Extensive experiments confirm the superior performance of the proposed FTransUNet compared with other multimodal fusion approaches on two fine-resolution remote sensing datasets, namely ISPRS Vaihingen and Potsdam. The source code in this work is available at https://github.com/sstary/SSRS.**

***Index Terms*—Multilevel Multimodal Fusion, Vision Transformer, Remote Sensing, Semantic Segmentation**

I. Introduction

Recent technological advances in earth observation have led to the increasing accessibility to fine-resolution remote sensing data of various modalities such as optical, multispectral and hyperspectral imagery, synthetic aperture radar (SAR) and light detection and ranging (LiDAR). Effective integration of these multimodal data can provide more comprehensive characterizations of the land surface for many tasks in geoscience research, including change detection [1, 2], land cover mapping [3, 4], object extraction [5, 6], and other tasks [7–9]. In particular, semantic segmentation as a pixel-wise classification task, which aims to classify each pixel into a specific land cover type, has drawn significant attention. In the literature, various semantic segmentation methods have been proposed by using random forest [10], support vector machine [11, 12], and conditional random field [13]. However, these methods are handicapped by their weak abstract and semantic feature extraction capabilities. Recently, deep learning (DL)-based methods have been successfully developed for remote sensing semantic segmentation by exploiting convolutional neural networks (CNN) [14–25]. Despite their excellent performance, these DL-based methods are hindered by the fact that the convolutional operation generally aggregates information from a smaller receptive field. As a result, these methods can only extract local details while overlooking long-range dependencies.

Interestingly, the computer vision (CV) community has encountered challenges similar to those in remote sensing. To cope with this problem, vision transformer (Vit) [26] was developed for CV applications to compute the relationships between each pair of elements in the feature map by capitalizing on the self-attention mechanism [27, 28], which enhances the modeling of global contextual information. Empowered with these robust backbones, i.e., CNN and transformers, semantic segmentation for single modality [29–32] achieved outstanding performance in the CV community. However, only a few works were devoted to the multimodal fusion tasks, emphasizing the fusion of multimodal information according to the characteristics of different modalities [14, 33, 34]. Compared with single-modal data that is limited by sensors, multimodal data is able to show the characteristics of the target from different perspectives. Thereby, better semantic segmentation performance can be achieved by utilizing the complementary properties derived from different modalities. However, the incompatibility of the multimodal data makes the fusion tasks challenging. Compared with natural images, high-resolution remote sensing images have more severe spectral heterogeneity and more complex spatial structures [3, 35]. Moreover, the ground objects in remote sensing data exhibit more significant variations in scales and shapes, which makes it challenging to locate and recognize objects. As a result, CNNs and transformers-based models derived from CV still exhibit limitations in effectively learning discriminative integrated features. To solve the multimodal fusion problem in

This work was supported in part by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence under Grant 2022B1212010001 and National Natural Science Foundation of China under Grant 42371374 and 41801323. *(Corresponding author: Man-On Pun; Xiaokang Zhang)*

Xianping Ma and Man-On Pun are with the School of Science and Engineering, the Future Network of Intelligence Institute (FNii), the Chinese University of Hong Kong, Shenzhen, Shenzhen 518172, China (e-mail: 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).

Ming Liu is with MizarVision, Shanghai, China (liuming@mizarvision.com).

remote sensing, three different fusion strategies have been explored: early fusion, middle fusion and late fusion. Generally, early fusion necessitates proper alignment of multimodal data and may lack robustness for task-independent information, while late fusion is limited in its ability to exploit cross-correlations between multimodal data [17, 36]. In contrast, middle fusion can capture cross-modal dependencies of feature representations, making it more effective in the context of representation learning [37]. However, in this line of research, existing works normally utilize single-level feature fusion based on summation or concatenation and ignore long-range cross-modal dependencies at different feature levels [38].

Motivated by the discussions above, this work proposes a multilevel multimodal fusion strategy, namely Fusion TransUNet (FTransUNet), to overcome the challenges aforementioned in the semantic segmentation of remote sensing data. More specifically, this work first develops a dual-branch model in which the extracted shallow-level fine-grained feature maps from each convolutional layer are fused by shallow-level feature fusion (SFF) module that consists of two squeeze-and-excitation (SE) modules [39]. Note that the SE module can be directly replaced with other upgraded versions to achieve shallow-level feature fusion. This CNN-based shallow-level feature fusion is designed to characterize objects of various scales and shapes. Next, deep-level contextual features are extracted and fused by the proposed Fusion Vision transformer (FVit). In sharp contrast to the TransUNet [31] that targets long-range dependencies for a single modality, the proposed FVit focuses on information enhancement inside the modalities and information exchanges across modalities in a three-stage strategy by applying novel Ada-MBA layers and SA layers alternately. More specifically, the proposed FVit first computes SA for each modality in the first stage before the Ada-MBA layer captures the long-range relationships between different modalities by performing CA and SA simultaneously in the second stage. In particular, two SA modules retain and enhance the intra-modal information while two CA modules guide feature fusion by the mutual-association guidance mechanism in Ada-MBA layer, promoting the learning of cross-modality representations with high inter-class separability and low intra-class variations. After that, the fused deep-level features are enhanced by SA again in the third stage. Empowered with our multilevel strategy, multimodal remote sensing images are encoded effectively, resolving the problem that remote sensing images are more complicated than natural images. Finally, the resulting shallow-level and deep-level features are fed into a cascaded decoder in which features are fused and upsampled with skip connections to recover the input image size. The contributions of this work are threefold, as summarized in the following:

- A Fusion Vision transformer (FVit) is proposed by capitalizing on SA layers and novel Ada-MBA layers. More specifically, FVit extracts and integrates global contextual information using the mutual-association guidance mechanism in a well-designed three-stage structure that alternately applies Ada-MBA layers and SA layers;
- A novel multilevel fusion scheme, namely FTransUNet, is proposed to address the multimodal fusion problem by combining FVit with CNN blocks to learn representative features across different modalities. In particular, detailed shallow-level features and contextual deep-level features are successively fused in a multilevel manner to facilitate the learning towards semantic-aware regions, hereby promoting the performance of semantic segmentation for remote sensing images.
- Extensive experiments on two fine-resolution remote sensing datasets, ISPRS Vaihingen and Potsdam, confirm that the proposed FTransUNet substantially outperforms existing models.

The remainder of this paper is organized as follows. Sec. II first reviews the related works on CNN-based and transformer-based segmentation methods. After that, Sec. III presents the structure of the proposed FTransUNet, whereas Sec. IV provides details on the extensive experiments conducted. Finally, the conclusion is given in Sec. V.

II. Related works

A. Single-Modal Semantic Segmentation

The seminal work [29] proposed the first CNN-based end-to-end model called fully convolutional network (FCN) for semantic segmentation. However, FCN suffers from blurred edges and inaccurate segmentation due to its over-simplified upsampling operation in its decoder design. To circumvent this problem, UNet [30] adopts a classical encoder-decoder network with an expanding path in its decoder. More specifically, the encoder in the UNet extracts multiscale features with gradual downsampling, while the decoder learns more contextual semantic information by restoring the spatial resolution step by step. Despite their good performance, CNN-based methods are ineffective in extracting global semantic information and long-range dependencies of input images due to the limited receptive fields of CNN [40, 41].

To overcome this challenge, the transformer architecture has been recently proposed in [27] to provide an effective alternative for capturing long-range relationships in natural language processing (NLP). More recently, a transformer-based model called Vit has been successfully developed in [26] for CV applications. Empowered with its capability of sequence-to-sequence modeling, the Vit demonstrates significantly improved performance in extracting global contextual information as compared to the CNN-based methods, which has inspired many following works [31, 42–46].

B. Multimodal Semantic Segmentation based on CNN

Driven by technological advances in earth observation, multimodal remote sensing data such as optical images, multispectral images and digital surface model (DSM) data, have become increasingly available. Since DSM represents height information for ground objects, it can provide vital information for identifying objects of consistent heights, such as roads and buildings in small geographical areas. Furthermore, DSM can provide clear boundary information that is helpful for all categories. As a result, it can help improve the overall segmentation performance [19, 47]. To utilize the multimodal data in remote sensing, pioneering multimodal fusion schemes were developed based on the DL framework to utilize the information derived from multiple sensors. For instance, ResUNet-a was developed in [48] by stacking Red-Green-Blue (RGB) images and the DSM into synthesized inputs of four channels. Clearly, such a primitive fusion of multimodal data cannot effectively handle the heterogeneous statistical properties and noise levels across modalities [49, 50]. To cope with this problem, FuseNet [14] devised a simple two-branch network architecture to fuse RGB and DSM data. More specifically, the RGB and DSM data are processed in parallel by each branch during the encoding stage while feature maps generated by the two branches are added element-wise after each convolutional block. Inspired by FuseNet, vFuseNet [15] was established to leverage multiscale fusion strategies. Despite its many advantages, vFuseNet is handicapped by its simple fusion design using element-wise addition of multimodal feature maps. CMGFNet [47] proposed a gated fusion module to combine two modalities for building extraction, which adaptively learns the discriminative features and removes irrelevant information. IIHN [51] constructed a hypergraph network by introducing an interpretable intuition mechanism. CIMFNet [52] and ABHNet [53] both explored feature fusion from the perspective of adjacent levels. The former was based on the cross-layer gate fusion mechanism while the latter was based on attention mechanisms and residual connections. However, these methods tend to be ineffective in extracting the global semantics as they ignore the long-range spatial relationships.

C. Multimodal Semantic Segmentation based on CNN and Transformer

More recently, the transformer architecture has been considered for multimodal fusion in semantic segmentation for its outstanding capability in extracting contextual information [20, 26, 32, 54] and fusing different modalities [19, 38, 55–59]. For instance, TransFuser was developed in [56] to combine feature maps using multiple transformers after each convolutional block. Furthermore, SwinFusion [57] was proposed to first extract shallow-level features using convolutional layers before generating deep-level features using the SA-based Swin Transformer [54], which has shown that the transformer can also serve as the backbone for multimodal fusion tasks. In remote sensing, CMFNet [19] improved the skip connections by exploiting a crossmodal multiscale fusion transformer, which allows the CMFNet to learn robust representations about surface objects of significant variations in scale. EDFT [58] constructed a two-branch network based on Swin Transformer and fused multimodal features by a depth-aware SA module. Similarly, MFTransNet [59] constructed a two-branch network based on CNN while fusing features with an SA module followed by a channel attention module and a spatial attention module. Despite their great performance, these fusion methods did not explore the general network from the perspective of using both CNN and Transformer for feature extraction and fusion, which incurs insufficient modeling of fine-grained local and contextual global information, and subsequently poor feature representations.

TABLE I
MULTIMODAL FUSION STRATEGIES BASED ON CNN AND TRANSFORMER.

	Feature Extractation	**Feature Fusion**
Class 1	CNN	Transformer
Class 2	Transformer	CNN or Transformer
Ours	CNN, Transformer	CNN, Transformer

As listed in Table I, the existing methods can be divided into two categories, namely Class 1 including TransFuser [56], SwinFusion [57], CMFNet [19], MFTransUNet [59] and Class 2 including EDFT [58]. In sharp contrast to these existing methods that utilize CNN or Transformer in feature extraction or feature fusion, the proposed FTransUNet takes full advantage of CNN and a well-designed three-stage Transformer in both feature extraction and feature fusion simultaneously. This proposed strategy is shown critical for effective multimodal multilevel fusion using CNN and Transformer backbones.

III. PROPOSED METHOD

Fig. 1 depicts the overview of the proposed framework. Specifically, in the CNN backbone, the SFF module is employed for shallow-level feature fusion, while an Ada-MBA layer in FVit is designated for deep-level feature fusion. More specifically, the proposed FTransUNet extracts the visible image features (VIS) and DSM features separately using two ResNet branches [60]. The resulting shallow-level features of different scales are enhanced by SFF modules after each CNN block. After that, the fused features are flattened into sequences and subsequently, fed into FVit in which the deep-level features are further fused to learn more contextual representational information by Ada-MBA. Finally, the fused outputs from FVit are added together and reshaped before being fed into a decoder of cascaded upsamplers [31], which helps to recover spatial information with finer precision.

For brevity of presentation, dual-modality data, namely VIS images and the DSM data, are used to elaborate the proposed FTransUNet. Furthermore, the VIS images are regarded as the main modality, while the DSM data serve as the assisting modality since the main modality generally provides more information for the earth surface classification than the assisting modality. It should be emphasized that this fusion scheme in FTransUNet can be extended to other multimodal fusion tasks in a straightforward manner. In the following, we will elaborate on the key components of the proposed FTransUNet.

A. CNN Fusion

We denote VIS images and their corresponding DSM data by $X \in \mathbb{R}^{H \times W \times 3}$ and $Y \in \mathbb{R}^{H \times W \times 1}$, respectively, where H and W are the height and width of the inputs. Note that the VIS images are of three channels, whereas the DSM data one channel. Adopting an architecture similar to TransUNet [31], the proposed FTransUNet utilizes a dual-branch encoder that first extracts multiscale features from each modality using one branch. More specifically, each branch of this encoder consists of four convolutional layers for multiscale feature extraction. The downsampled feature maps of size $C_i \times \frac{H}{2^{i-1}} \times \frac{W}{2^{i-1}}$ are

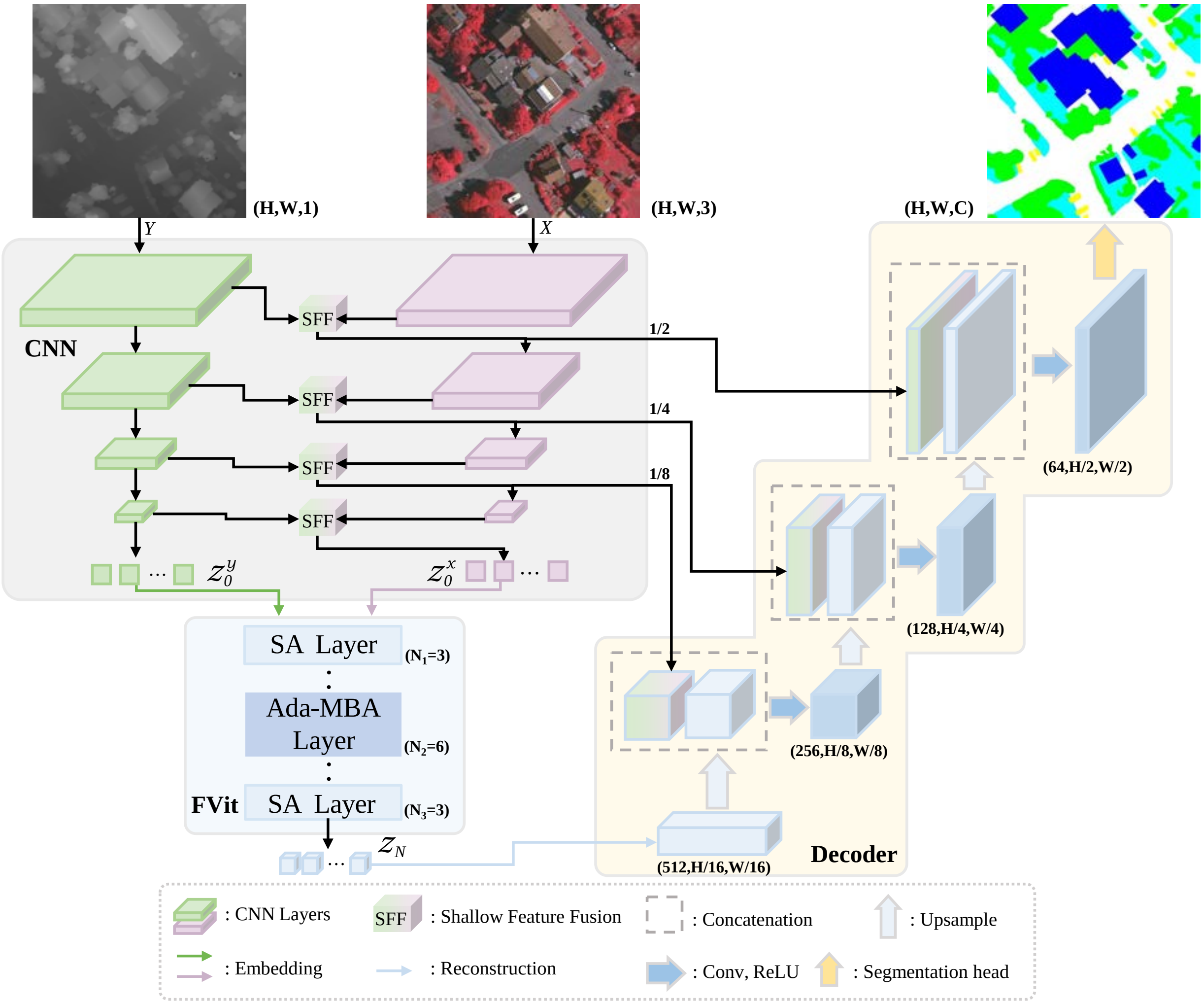

Fig. 1. Overview of the proposed Fusion TransUNet (FTransUNet) with the encoder consisting of CNN and FVit on the left and the decoder shown on the right. CNN blocks contain two branches comprised of two ResNets that extract multimodal shallow-level features and fuse them using four SFF modules. FVit consists of three stages to achieve deep-level feature extraction and fusion. The decoder generates segmentation maps by exploiting fused features using cascaded upsampling techniques [31].

produced by the i-th encoder layer, where i is the layer index of the CNN encoder. These shallow-level features extracted by the convolutional operation are then fused by the SFF modules. In particular, features derived from the assisting modality are fused into those from the main modality, i.e., VIS, before the fused features are input into the next VIS encoder branch. Furthermore, skip connections are utilized by directly feeding the outputs of the SFF modules to the corresponding decoder layers designed to recover detailed local and contextual information.

As shown in Fig. 2, the SFF module first aggregates global information using Global Global Average Pool (AvgPool) in the VIS and DSM branches, respectively. Given the input channel size C_i for the i-th SFF module, the squeeze and excitation process is carried out by the AvgPool and two convolutional operations of kernel size 1×1, followed by the ReLU and the Sigmoid functions. Finally, features from VIS and DSM are weighted and added element-wise, which generates the resulting fused shallow-level features.

B. FVit

Denote by x_I and y_I the VIS and DSM feature maps of dimension $C_I \times \frac{H}{2^{I-1}} \times \frac{W}{2^{I-1}}$, respectively, where I and C_I are the layer index and the output channel size for the last layer in the CNN backbone, respectively. x_I and y_I are first tokenized using two embedding layers and one reshape operation. The embedding layers change the input's channel size from C_I to C_{hid} before the reshape operation flattens the outputs from the embedding layers into two sequences of 2D patches denoted by z_0^x and z_0^y of size $C_{\text{hid}} \times L$, where $L = (H \times W)/(2^{I-1} \times 2^{I-1})$ is the sequence length. Specific position embeddings are added to the vectorized patches z_0^x and z_0^y to retain position information. After that, the tokens z_0^x and z_0^y are fed into FVit.

Input of the FVit encoder sequentially goes through three processes, including SA layers for deep-level feature enhancement, Ada-MBA layers for deep-level feature fusion and another part of SA layers for fused feature enhancement, with the number of layers N_1, N_2 and N_3, respectively. Denote by z_n^x and z_n^y the hidden features for the n-th layer in the

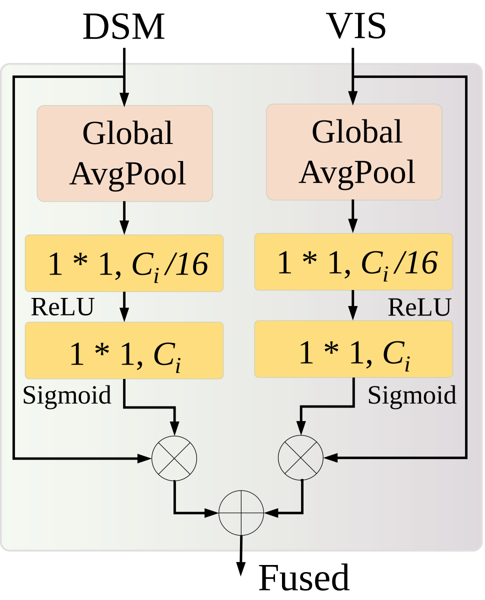

Fig. 2. The proposed SFF module for shallow-level feature fusion in the CNN blocks. There are two branches enhancing the features in channel-wise, and the multimodal information is fused by element-wise addition.

VIS branch and the DSM branch, respectively, where $n \in 1, 2, ..., N_1 + N_2 + N_3$. It is worth noting that the processes preserve the dimension of the feature maps as $C_{\text{hid}} \times L$ throughout FVit. In particular, the SA layer is comprised of two SA modules, two Multi-Layer Perceptron (MLP) modules and layer normalization (LN) operators, as shown in Fig. 3(a). Given multimodal features inputs denoted by z^x_{n-1} and z^y_{n-1}, the SA layer is designed to derive the global relationships for each modality using the Multihead Self-Attention mechanism proposed in [26]. Mathematically, the output of the n-th SA layer for $n = 1, 2, 3$ can be written as follows:

$$\begin{aligned} p^x_n &= SA\left(LN(z^x_{n-1})\right) + z^x_{n-1}, &(1)\\ p^y_n &= SA\left(LN(z^y_{n-1})\right) + z^y_{n-1}, &(2)\\ z^x_n &= MLP\left(LN(p^x_n)\right) + p^x_n, &(3)\\ z^y_n &= MLP\left(LN(p^y_n)\right) + p^y_n. &(4) \end{aligned}$$

After the deep-level feature enhancement performed by the SA layer, FVit further fuses multimodal features in an abstract semantic space with rich contextual information using N_2 Ada-MBA layers, as shown in Fig. 3(b). In this deep-level feature fusion stage, CA and SA are computed simultaneously in the Ada-MBA module to learn the correlation between the main and assisting modalities. The output from the Ada-MBA layer can be written as:

$$(g^x_n, g^y_n) = Ada\text{-}MBA(LN(z^x_{n-1}), LN(z^y_{n-1})). \quad (5)$$

Let $q^x_n = g^x_n + z^x_{n-1}$ and $q^y_n = g^y_n + z^y_{n-1}$. The fused output from the n-th layer for $n = 4, 5, ..., 9$ can be written as:

$$\begin{aligned} z^x_n &= MLP\left(LN(q^x_n)\right) + q^x_n, &(6)\\ z^y_n &= MLP\left(LN(q^y_n)\right) + q^y_n. &(7) \end{aligned}$$

Fig. 4 illustrates the structure of the proposed Ada-MBA module. Empowered by the multihead design [26, 27], the

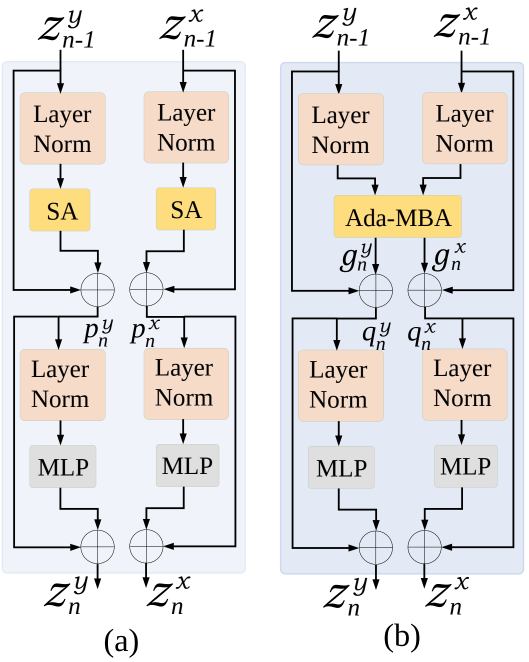

Fig. 3. (a) The proposed SA layer in FVit. (b) The proposed Ada-MBA layer for deep-level feature fusion in FVit. The SA layer and the Ada-MBA layer are similar in structure, but they have different attention modules. The former extracts the multimodal features while the latter fuses these deep-level semantic features.

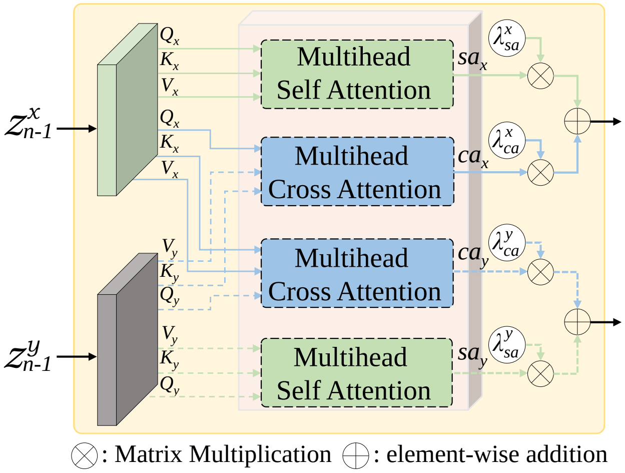

Fig. 4. Schematic of the Ada-MBA module. This module includes two multihead SA modules and two multihead CA modules, which realizes the intra-modal information extraction and the inter-modal information fusion, respectively. It should be noted that the two Q_x's in the figure are identical. For presentational clarity, we add a label for SA and another label for CA while other matrices are the same.

proposed Ada-MBA module divides its multimodal feature inputs z^x_{n-1} and z^y_{n-1} into H equal segments denoted by $z^x_{n-1,h}$ and $z^y_{n-1,h}$ for $h = 1, 2, ..., H$, where H is the number of heads. Since the operations of each head are identical, we will omit the index h in the following discussions on the fusion attention module in the multihead mechanism. Two groups of matrices $\{Q_x, K_x, V_x\}$ and $\{Q_y, K_y, V_y\}$ for multimodal information can be computed by using linear projections U^x_{qkv} and U^y_{qkv}, respectively. After that, SA information (sa_x, sa_y)

and CA information (ca_x, ca_y) are derived for two modalities simultaneously. More specifically, SA computes intra-modal information using $\{Q_x, K_x, V_x\}$ and $\{Q_y, K_y, V_y\}$ while CA computes inter-modal information using $\{Q_x, K_y, V_y\}$ and $\{Q_y, K_x, V_x\}$. This process realizes the extraction and fusion of the deep-level features in one module. As shown in Fig. 4, the process in the Ada-MBA module can be expressed as:

$$[Q_x, K_x, V_x] = z^x_{n-1} U^x_{qkv}, \tag{8}$$

$$[Q_y, K_y, V_y] = z^y_{n-1} U^y_{qkv}, \tag{9}$$

$$sa_x = \varphi\left(\frac{Q_x K_x^\top}{\sqrt{d}}\right) V_x, \tag{10}$$

$$sa_y = \varphi\left(\frac{Q_y K_y^\top}{\sqrt{d}}\right) V_y, \tag{11}$$

$$ca_x = \varphi\left(\frac{Q_x K_y^\top}{\sqrt{d}}\right) V_y, \tag{12}$$

$$ca_y = \varphi\left(\frac{Q_y K_x^\top}{\sqrt{d}}\right) V_x, \tag{13}$$

where d is the normalization parameter while $\varphi(\cdot)$ and $(\cdot)^\top$ are the softmax function and the matrix transpose operator, respectively. Eq. (12) and Eq. (13) are the mutual-association guidance mechanism. It is observed that a guided fusion of the two modalities is achieved by exchanging the partial matrices.

Next, the following adaptive mechanism is proposed to fuse SA and CA:

$$g^x_n = \lambda^x_{sa} sa_x + \lambda^x_{ca} ca_x, \tag{14}$$

$$g^y_n = \lambda^y_{sa} sa_y + \lambda^y_{ca} ca_y, \tag{15}$$

where λ^x_{sa}, λ^x_{sa}, λ^x_{sa} and λ^x_{sa} are the learnable weighting coefficients to balance the contributions from SA and CA.

Finally, the fused feature maps are enhanced by N_3 SA layers. The procedures from Eqs. (1)-(4) are repeated for N_3 times to enhance the fused feature maps for the VIS branch and the DSM branch, respectively. The final output of FVit denoted as $z_N \in \mathbb{R}^{C_{\text{hid}} \times L}$ is the feature maps derived from the last SA layer. Based on the proposed FVit, the rich contextual information extracted from multimodal data is deeply fused before being fed into the cascaded decoder.

C. Cascaded Decoder

As shown in Fig. 1, the cascaded decoder recovers the hidden fused features for the final segmentation process by exploiting multiple upsampling modules. More specifically, the decoder first reshapes the 2D input sequence z_N into 3D tensors of size $C_{\text{dec}} \times \frac{H}{2^{l-1}} \times \frac{W}{2^{l-1}}$ using the Reconstruction module where C_{dec} is the input channel number of the first block in the decoder. After that, multiple cascaded decoder blocks restore the spatial resolution to $H \times W$ by concatenating skip connections from the corresponding CNN backbone layers. Each decoder block consists of an upsampling operator, a convolution (Conv) layer, and a ReLU layer. Finally, the segmentation head performs the final semantic prediction.

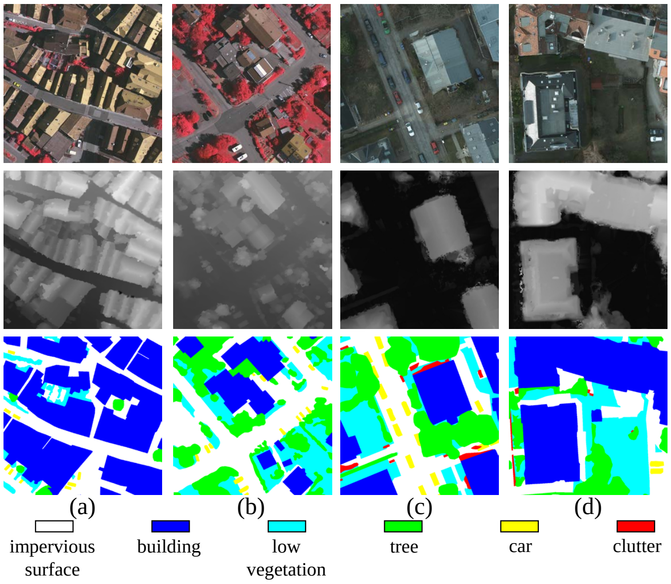

Fig. 5. Data samples of size 1024 × 1024 from Vaihingen (first two columns) and Potsdam (last two columns), respectively. The first row shows the orthophotos with three channels (NIRRG for Vaihingen and RGB for Potsdam). The second and third rows show the corresponding depth information and semantic labels in pixel-wise. These samples present the auxiliary role of DSM modality for classes with consistent elevation and clear boundaries.

IV. Experiments and Discussion

A. Datasets

1) Vaihingen: The Vaihingen dataset consists of 16 very high-resolution True Orthophotos of an average size of 2500×2000 pixels. Each Orthophoto has three channels, namely Near-InfRared, Red and Green (NIRRG), as well as a normalized digital surface model (DSM) of 9 cm ground sampling distance (GSD). The dataset includes five foreground classes, namely *Building (Bui.)*, *tree (Tre.)*, *Low vegetation (Low.)*, *Car*, *Impervious surface (Imp.)* and one background class (*Clutter*). Furthermore, 16 orthophotos are divided into one training set of 12 patches and one test set of 4 patches in which the training set contains those orthophotos indexed by $1, 3, 23, 26, 7, 11, 13, 28, 17, 32, 34, 37$ while the test set $5, 21, 15, 30$.

2) Potsdam: The Potsdam dataset is comprised of 24 very high-resolution True Orthophotos of the size of 6000×6000 pixels for each tile. It provides four multispectral channels, namely InfRared, Red, Green and Blue (IRRGB), as well as a normalized digital surface model (DSM) of 5 cm ground sampling distance (GSD). The foreground classes are the same as in Vaihingen, while the class distribution differs due to the different geographical locations. RGB composites are used in this dataset. 24 orthophotos are divided into 18 patches and 6 patches for training and test, respectively. More specifically, the training set contains orthophotos indexed by $6\_10, 7\_10, 2\_12, 3\_11, 2\_10, 7\_8, 5\_10, 3\_12, 5\_12, 7\_11, 7\_9, 6\_9, 7\_7, 4\_12, 6\_8, 6\_12, 6\_7, 4\_11$ whereas the test $2\_11, 3\_10, 4\_10, 5\_11, 6\_11, 7\_12$.

Some data samples from the two datasets are illustrated in Fig. 5. Unless otherwise specified, a sliding window is used to dynamically collect the training batches, which helps read

TABLE II
EXPERIMENTAL RESULTS ON THE VAIHINGEN DATASET. WE PRESENT THE OA OF FIVE FOREGROUND CLASSES AND THREE OVERALL PERFORMANCE METRICS. BOLD VALUES ARE THE BEST.

Type	Method	OA(%)						mF1(%)	mIoU(%)
		Bui.	Tre.	Low.	Car	Imp.	Total		
CNN-based	ABCNet [61]	94.10	90.81	78.53	64.12	89.70	89.25	85.34	75.20
	PSPNet [62]	94.52	90.17	78.84	79.22	92.03	89.94	86.55	76.96
	MAResU-Net [63]	94.84	89.99	79.09	85.89	92.19	90.17	88.54	79.89
	vFuseNet [15]	95.92	91.36	77.64	76.06	91.85	90.49	87.89	78.92
	FuseNet [14]	96.28	90.28	78.98	81.37	91.66	90.51	87.71	78.71
	ESANet [33]	95.69	90.50	77.16	85.46	91.39	90.61	88.18	79.42
	SA-GATE [64]	94.84	92.56	81.29	87.79	91.69	91.10	89.81	81.27
	CMGFNet [47]	97.75	91.60	80.03	87.28	92.35	91.72	90.00	82.26
Transformer-based	TransUNet [31]	96.48	**92.77**	76.14	69.56	91.66	90.96	87.34	78.26
	CMFNet [19]	97.17	90.82	80.37	85.47	92.36	91.40	89.48	81.44
	UNetFormer [44]	96.23	91.85	79.95	86.99	91.85	91.17	89.85	81.97
	MFTransNet [59]	96.41	91.48	80.09	86.52	92.11	91.22	89.62	81.61
	FTransUNet (Ours)	**98.20**	91.94	**81.49**	**91.27**	**93.01**	**92.40**	**91.21**	**84.23**

large images without cropping in advance. The sliding window size is set to 256×256 with a stride of 256 in the training stage and 32 in the test stage. A smaller stride in the test stage can reduce border effects by averaging the prediction results on the overlapping regions.

B. Evaluation Metrics

To evaluate the segmentation results of the multimodal remote sensing data, the Overall Accuracy (OA), the mean F1 score (mF1) and the mean Intersection over Union (mIoU) are employed. These standard statistical indices can provide fair comparisons of the performance of the proposed FTransUNet and other state-of-the-art methods. More specifically, we compute mF1 and mIoU of the five foreground classes. In addition, we also include *Clutter* or *Background* in the evaluation of OA whose definition is given by:

$$\text{OA} = \frac{TP+TN}{TP+FP+TN+FN}, \quad (16)$$

where TP, TN, FP, and FN denote true positive, true negative, false positive and false negative, respectively. Furthermore, F1 and IoU are calculated for each foreground class indexed by c according to the following formula:

$$\text{F1} = 2 \times \frac{Q_c \times R_c}{Q_c + R_c}, \quad (17)$$

$$\text{IoU} = \frac{TP_c}{TP_c + FP_c + FN_c}, \quad (18)$$

where TP_c, FP_c and FN_c are true positives, false positives, and false negatives for the c-th class, respectively. Finally, Q_c and R_c are given by:

$$Q_c = \frac{TP_c}{TP_c + FP_c}, \quad (19)$$

$$R_c = \frac{TP_c}{TP_c + FN_c}. \quad (20)$$

We derive the mean values, namely mF1 and mIoU, upon obtaining F1 and IoU for the five useful foreground classes according to the definitions above.

C. Implementation details

The experiments were all implemented with PyTorch on a single NVIDIA GeForce RTX 3090 GPU with 24GB RAM. All models were trained using the stochastic gradient descent (SGD) algorithm with a learning rate of 0.01, a momentum of 0.9, a decaying coefficient of 0.0005, and a batch size of 10. Simple data augmentations, e.g., random rotation and flipping, are applied after the sliding window collects samples. The CNN backbone used in the proposed FTransUNet is comprised of two ResNet50 models, each of which consists of four convolutional layers, i.e. $I = 4$, with a hidden size of $C_{\text{hid}} = 768$. FVit has in total $N_1 + N_2 + N_3 = 12$ transformer layers with $N_1 = 3$, $N_2 = 6$ and $N_3 = 3$. The number of heads in each layer is set to $H = 12$ while the channel size is set to $C_{\text{dec}} = 512$. Finally, all transformer backbones and ResNet50 were pretrained on ImageNet [65] for better initialization.

D. Performance Comparison

We benchmarked the performance of the proposed FTransUNet against twelve representative state-of-the-art methods, namely ABCNet [61], PSPNet [62], MAResU-Net [63], vFuseNet [15], FuseNet [14], ESANet [33], TransUNet [31], SA-GATE [64], UNetFormer [44], CMFNet [19], MFTransNet [59] and CMGFNet [47]. In our experiments, ABCNet, PSPNet, MAResU-Net and UNetFormer only considered the main modal information, i.e., the VIS images. These advanced single-modal methods can illustrate the influence of DSM data, in other words, the advantages of multimodal over single-modal. In contrast, other methods took into account crossmodal data. In particular, the following modification was made on the baseline method TransUNet for a fair comparison: We modified the input layer and stacked VIS and DSM data into four channels for network training and test as suggested in [15, 48]. The quantitative results are listed in Table II and Table III.

1) Performance Comparison on the Vaihingen dataset: As presented in Table II, the proposed FTransUNet showed significant improvements in terms of OA, mF1 and mIoU as compared to the baseline TransUNet, which confirmed that the novel FTransUNet successfully fused shallow-level

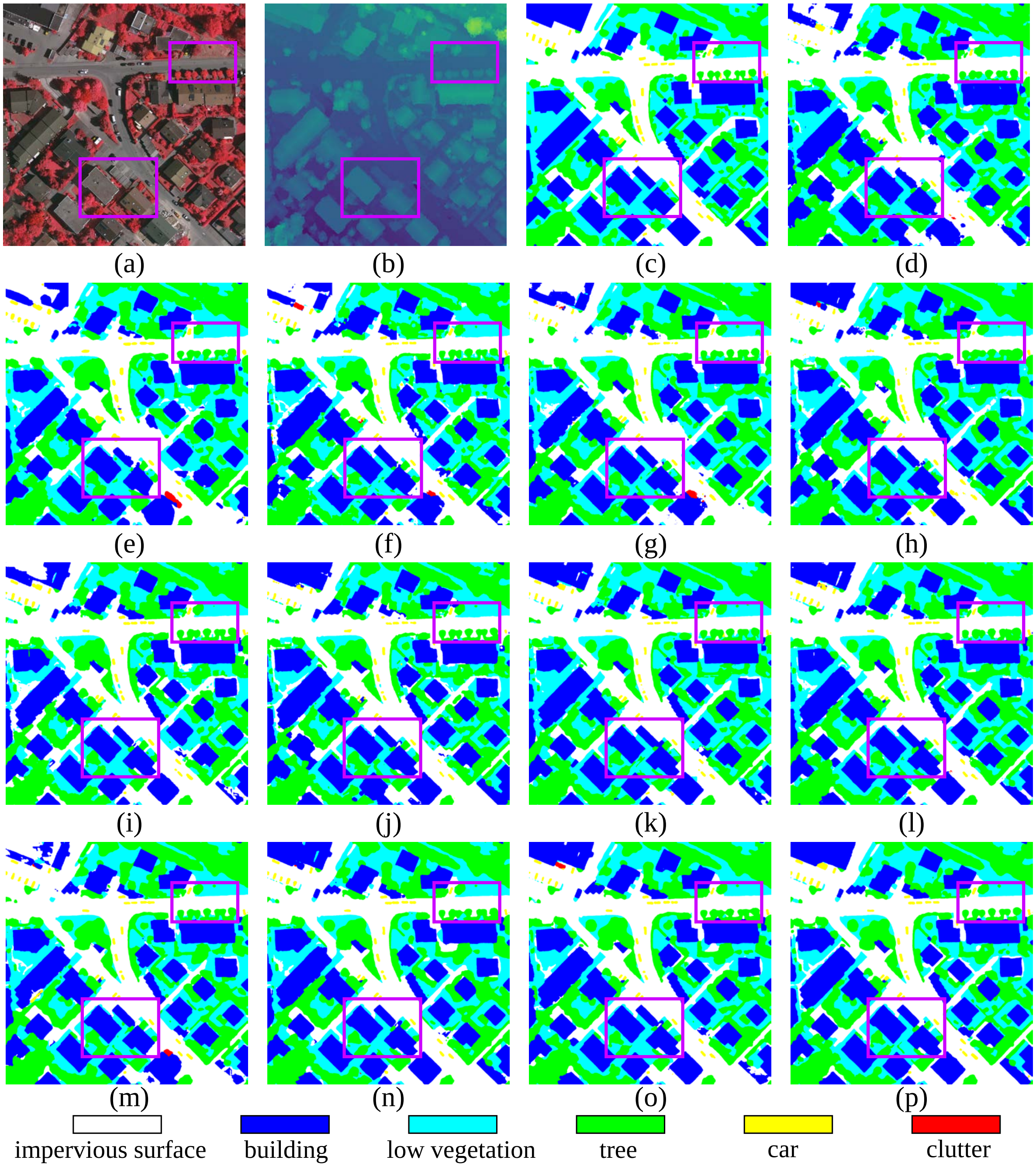

Fig. 6. Qualitative performance comparisons on the Vaihingen test set with the size of 1800 × 1800. (a) NIRRG images, (b) DSM, (c) Ground Truth, (d) ABCNet, (e) PSPNet, (f) MAResU-Net, (g) vFuseNet, (h) FuseNet, (i) ESANet, (j) TransUNet, (k) SA-GATE, (l) CMFNet, (m) UNetFormer, (n) MFTransNet, (o) CMGFNet, (p) The proposed FTransUNet. Two purple boxes are added to all subfigures to highlight the differences.

and deep-level features by extracting complementary information from the assisting modality and effectively deriving robust representations. Compared with existing state-of-the-art models, FTransUNet outperformed on four classes, namely *Building*, *Low vegetation*, *Car* and *Impervious surface*. In particular, on the Vaihingen data set, FTransUNet provided a substantial improvement of 3.48% on the *Car* class as compared to the existing method SA-GATE. Furthermore, the classification accuracy for *Impervious surface* has been enhanced by 0.65% compared to the existing method CMFNet and the classification accuracy for *Building* has been enhanced by 0.45% compared to the existing method CMGFNet. These results can be explained by the fact that FTransUNet can more effectively extract and fuse multilevel multimodal features by exploiting CNN and transformer successively as compared to CMFNet. Furthermore, as CMGFNet was designed for building extraction, global information was not well taken into consideration for multi-category classification tasks. In contrast, the proposed FTransUNet is designed to effectively derive and exploit the depth information from the DSM images. As a

TABLE III
EXPERIMENTAL RESULTS ON THE POSTDAM DATASET. WE PRESENT THE OA OF FIVE FOREGROUND CLASSES AND THREE OVERALL PERFORMANCE METRICS. BOLD VALUES ARE THE BEST.

Type	Method	OA(%)						mF1(%)	mIoU(%)
		Bui.	Tre.	Low.	Car	Imp.	Total		
CNN-based	ABCNet [61]	96.23	78.92	86.40	92.92	88.90	87.52	88.14	79.26
	PSPNet [62]	97.03	83.13	85.67	88.81	90.91	88.67	88.92	80.36
	MAResU-Net [63]	96.82	83.97	87.70	95.88	92.19	89.82	90.86	83.61
	vFuseNet [15]	97.23	84.29	89.03	95.49	91.62	90.22	91.26	84.26
	FuseNet [14]	97.48	85.14	87.31	96.10	92.64	90.58	91.60	84.86
	ESANet [33]	97.10	85.31	87.81	94.08	92.76	89.74	91.22	84.15
	SA-GATE [64]	96.54	81.18	85.35	**96.63**	90.77	87.91	90.26	82.53
	CMGFNet [47]	97.41	86.80	86.68	95.68	92.60	90.21	91.40	84.53
Transformer-based	TransUNet [31]	96.63	82.65	**89.98**	93.17	91.93	90.01	90.97	83.74
	CMFNet [19]	97.63	87.40	88.00	95.68	92.84	91.16	92.10	85.63
	UNetFormer [44]	97.69	86.47	87.93	95.91	92.27	90.65	91.71	85.05
	MFTransNet [59]	97.37	85.71	86.92	96.05	92.45	89.96	91.11	84.04
	FTransUNet (Ours)	**97.78**	**88.27**	88.48	96.31	**93.17**	**91.34**	**92.41**	**86.20**

TABLE IV
EXPERIMENTAL RESULTS ON THE POSTDAM DATASET FOR SPECTRUM ANALYSIS. WE PRESENT THE OA OF FIVE FOREGROUND CLASSES AND THREE OVERALL PERFORMANCE METRICS.

Method	Bands	OA(%)						mF1(%)	mIoU(%)
		Bui.	Tre.	Low.	Car	Imp.	Total		
UNetFormer [44]	RGB	97.69	86.47	87.93	95.91	92.27	90.65	91.71	85.05
	IRRG	97.83	86.96	88.67	96.16	91.38	90.75	91.85	85.22
FTransUNet	RGB	96.63	82.65	89.98	93.17	91.93	90.01	90.97	83.74
	RGB+DSM	97.78	88.27	88.48	96.31	93.17	91.34	92.41	86.20
	IRRG+DSM	97.67	87.77	89.30	96.81	93.24	91.53	92.49	86.32

result, the fused features for classes of distinctive depth values have become more distinguishable. For instance, *Building* and *Impervious surface* generally exhibit higher and lower depth values, respectively. In terms of the overall performance, the proposed FTransUNet achieved OA of 92.40%, mF1 score of 91.21% and mIoU of 84.23%, which stands for an increase of 1.44%, 3.87% and 5.97% as compared to the corresponding performance of the baseline TransUNet, respectively. These results confirmed that the proposed FTransUNet achieved better generalization performance.

Fig. 6 shows a visualization example of the results obtained by all twelve methods under consideration. It is observed that remote sensing images are more complex than natural images as stated in Sec. I: (1) *Building* varies greatly in scale, with neat borders but various shapes. (2) *Tree* and *Low vegetation* are interleaved. Clearly, the proposed FTransUNet can identify complex edges with smoother results, providing a complete and connected object with fewer independent points. Specifically, the shallow-level feature fusion carried out by SE modules in CNN is beneficial for retaining details of objects with various scales and shapes, resulting in accurate edges of *Building* and *Impervious surface* objects. Furthermore, the deep-level feature fusion stage in FVit can more accurately recognize complex long-distance semantic information, which helps identify a complete object with reduced scattered points. These advantages enabled the proposed FTransUNet to achieve more accurate classification as compared to other single-modal or multimodal methods. There are two purple boxes in all subfigures of Fig. 6. It is observed that in the upper right box, FTransUNet clearly segmented the trees arranged in a successive row. In the lower box, FTransUNet effectively identified the low vegetation around the building, providing a more tidy and complete segmentation of the building.

2) Performance Comparison on the Potsdam dataset: Experiments on the Potsdam dataset also demonstrated results similar to those derived from the Vaihingen dataset. As presented in Table III, the classification accuracy for *Building*, *Tree*, *Car* and *Impervious surface* were 97.78%, 88.27%, 96.31% and 93.17%, respectively, which amounts to an increase of 1.15%, 5.62%, 3.14% and 1.24% as compared to the baseline TransUNet. The corresponding OA, mF1 score and mIoU values were 91.34%, 92.41%, 86.20% respectively, which corresponds to increases of 1.33%, 1.44% and 2.46%, respectively, over TransUNet. In particular, substantial improvements were achieved for *Building*, *Tree* and *Impervious surface* as compared to other state-of-the-art methods.

Fig. 7 shows a visualization example from Potsdam for all methods under consideration. It is observed that the proposed FTransUNet could more accurately classify objects like *Building* and small-scale objects like *Car*. This was achieved by the synergy of the CNN fusion and FVit in FTransUNet: The CNN fusion extracts robust representations of the basic features of ground objects while FVit recognizes complex remote sensing image content on the basis of the robust shallow-level features. In sharp contrast, most existing methods misclassified object boundaries, inflicting scattered artifacts. Two purple boxes are added to all subfigures in Fig. 7. The upper box shows *Building* covered by *Tree*. It can be observed that FTransUNet was more accurate in segmenting trees in this complex scene. Meanwhile, the segmentation of the building under the tree was closer to the ground truth. The lower left box shows the segmentation of the *Tree* in the area of *Low vegetation*. Inspection of Fig. 7 suggests that FTransUNet successfully divided the edges of the trees.

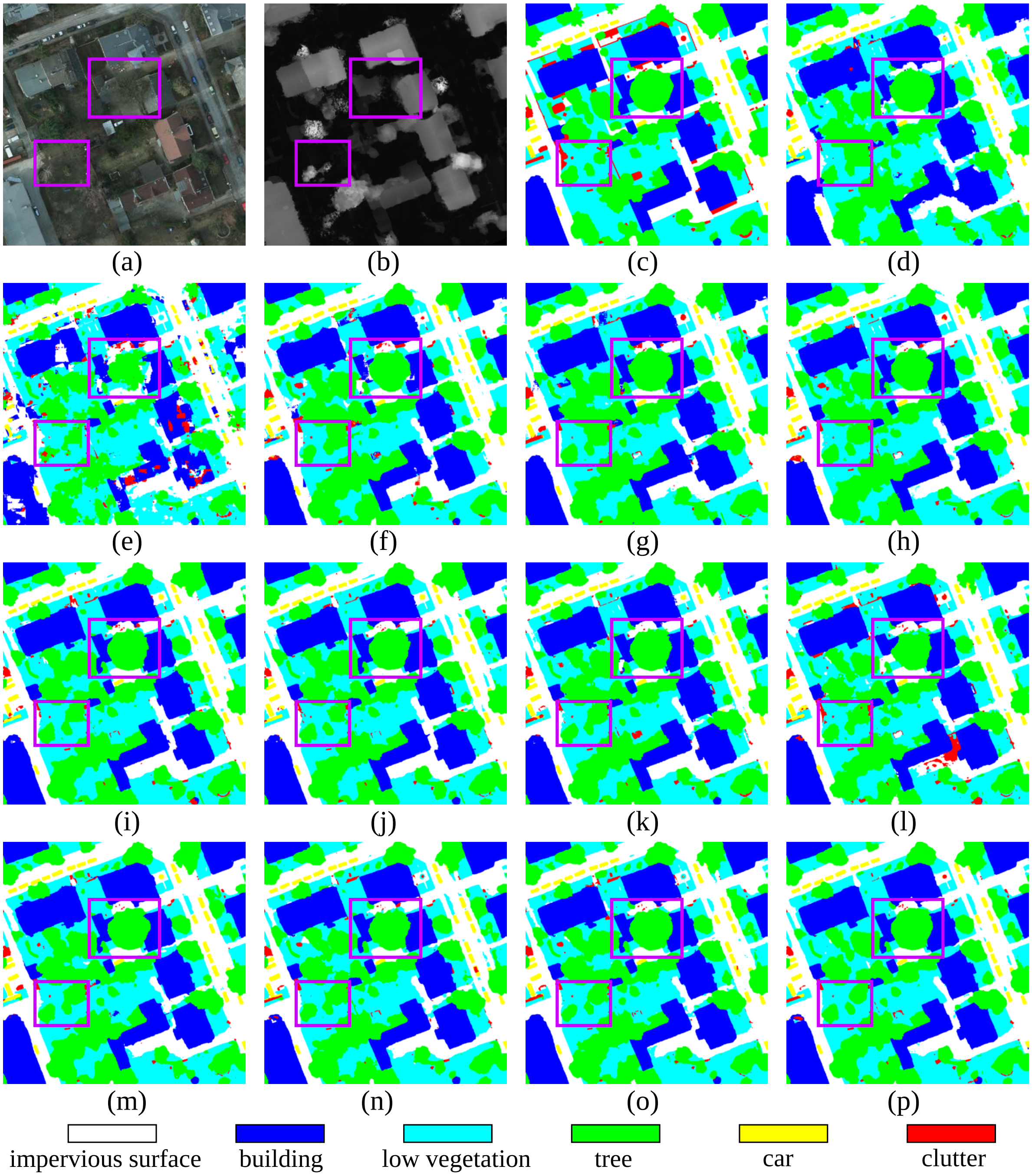

Fig. 7. Qualitative performance comparisons on the Potsdam test set with the size of 2000×2000. (a) RGB images, (b) DSM, (c) Ground Truth, (d) ABCNet, (e) PSPNet, (f) MAResU-Net, (g) vFuseNet, (h) FuseNet, (i) ESANet, (j) TransUNet, (k) SA-GATE, (l) CMFNet, (m) UNetFormer, (n) MFTransNet, (o) CMGFNet, (p) The proposed FTransUNet. Two purple boxes are added to all subfigures to highlight the differences.

Next, we performed spectrum analysis on Postdam dataset of IRRGB four bands. The analysis is designed to compare the difference between IRRG and RGB in single-modal and multimodal segmentation tasks as presented in Table IV. The results in the first and second rows suggest that NIRRG was more informative than RGB for semantic segmentation. This may be due to the fact that NIRRG can better characterize plants (*Tree* and *Low vegetation*). Similar observations were obtained for IRRG and RGB by inspecting the fourth and fifth rows in Table IV. Furthermore, a comparison of the third and fourth rows reveals that the DSM data provided a noticeable improvement in overall segmentation performance, though the accuracy for *Low vegetation* degraded by 1.5%. To shed light on this observation, Table V compared the accuracy for each category using the Vaihingen dataset and the Potsdam dataset. While the overall performance has been substantially improved by the proposed FTransUNet on both datasets, the accuracy for *Low vegetation* and *Tree* exhibited slightly different change patterns on the two datasets. More specifically, the proposed FTransUNet greatly improved the performance of *Low veg-*

TABLE V

EXPERIMENTAL RESULTS ON DSM ANALYSIS. WE PRESENT THE OA OF FIVE FOREGROUND CLASSES AND THREE OVERALL PERFORMANCE METRICS.

Dataset	**Bands**	**OA(%)**				
		Bui.	**Tre.**	**Low.**	**Car**	**Imp.**
Vaihingen	NIRRG	96.48	92.77	76.14	69.56	91.66
	NIRRG+DSM	98.20 (+1.72)	91.94 (-0.83)	81.49 (+5.35)	91.27 (+21.71)	93.01 (+1.35)
Potsdam	RGB	96.63	82.65	89.98	93.17	91.93
	RGB+DSM	97.78 (+1.15)	88.27 (+5.62)	88.48 (-1.50)	96.31 (+3.14)	93.17 (+1.24)

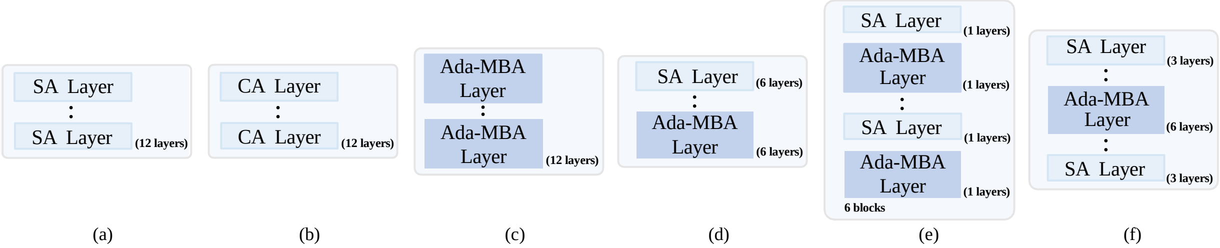

Fig. 8. Illustration of different FVit structures adopted in the experiments. (a) stru1, (b) stru2, (c) stru3, (d) stru4, (e) stru5 and (f) three-stage. We constructed five different structures with 12 attention layers to illustrate the superiority of the three-stage designation.

etation with minor degradation for *Tree* on the Vaihingen dataset. However, the result was opposite for the Potsdam dataset, i.e. improvement for *Tree* but minor degradation for *Low vegetation*. These results may be caused by the fact that the two categories share great similarities in color with highly irregular shapes of uncertain boundaries. As a result, it remains challenging to differentiate *Tree* and *Low vegetation* while improving overall performance.

As evidenced in Table V, the proposed FTransUNet can improve the accuracy for most categories on both datasets by effectively exploiting the additional DSM data. In particular, noticeable improvement on *Building* and *Impervious Surface* was observed as ground objects belonging to these two categories usually possess distinct elevation characteristics. Furthermore, since *Car* is mostly located on *Road*, it was also helpful to identify the boundary of *Car* from *Road* whose height is relatively consistent. Finally, as discussed before, the elevation characteristics of *Tree* and *Low vegetation* are rather similar, the FTransUNet experienced challenges in improving the accuracy for both categories. So said, the proposed FTransUNet was able to achieve significant overall performance improvement on both datasets.

E. Structure analysis

The proposed three-stage FVit structure includes N_1 SA layers for deep-level feature enhancement, N_2 Ada-MBA layers for deep-level feature fusion followed by N_3 SA layers for fusion feature enhancement. To shed light on this structure, in-depth structure analyses were conducted. As shown in Table VI, six sets of structure experiments were performed by varying the numbers and the order of SA layers and Ada-MBA layers. As shown in Fig. 8, the first experiment labeled as “stru1” utilizes FVit comprised of only 12 SA layers, i.e., the deep-level features in two branches are not fused by the Ada-MBA layers. In contrast, the second experiment labeled as “stru2” and the third experiment labeled as “stru3” consider the cases that FVit is comprised of only 12 CA layers and 12 Ada-MBA layers, respectively. Furthermore, the fourth experiment labeled as “stru4” represents 6 SA layers and 6 Ada-MBA layers, while the fifth experiment labeled as “stru5” stands for 6 blocks each of which is one SA layer followed by one Ada-MBA layer. Finally, the sixth experiment employs the proposed three-stage fusion structure, i.e., 3 SA layers followed by 6 Ada-MBA layers and 3 additional SA layers.

TABLE VI

STRUCTURE ANALYSIS OF FVIT. THE RESULT SHOWS THAT THE PROPOSED THREE-STAGE STRUCTURE IS CAPABLE OF DEEP-LEVEL FEATURE ENHANCEMENT, DEEP-LEVEL FEATURE FUSION AND FUSION FEATURE ENHANCEMENT. BOLD VALUES ARE THE BEST.

Structure	**OA(%)**	**mF1(%)**	**mIoU(%)**
stru1	92.13	90.60	83.25
stru2	92.06	90.70	83.39
stru3	92.33	90.82	83.58
stru4	92.35	90.91	83.75
stru5	92.26	90.75	83.46
three-stage	**92.40**	**91.21**	**84.23**

As shown in Table VI, the structure analysis results showed that the first experiment “stru1”, i.e., SA layers only, had the worst performance, suggesting the importance of the cross-modal fusion on deep-level features. Furthermore, Table VI confirmed that the hybrid designs employed in the fourth experiment “stru4”, in the fifth experiment “stru5” and the sixth experiment “three-stage”, i.e. the proposed FVit, outperformed the simple SA layers only “stru1”, the simple CA layers only “stru2” and the Ada-MBA layers only “stru3”. Finally, the design of “stru5” was not as effective as the proposed FVit structure. In summary, the six experiments showed that carefully designed hybrid structures combining SA and Ada-MBA layers effectively enhanced feature extraction and fusion. More specifically, it is evidenced from Table VI that intra-modal feature extraction done through SA and the cross-modal feature fusion achieved by Ada-MBA are critical for performance improvement. In particular, the proposed three-stage FVit structure plays a crucial role in deep-level feature

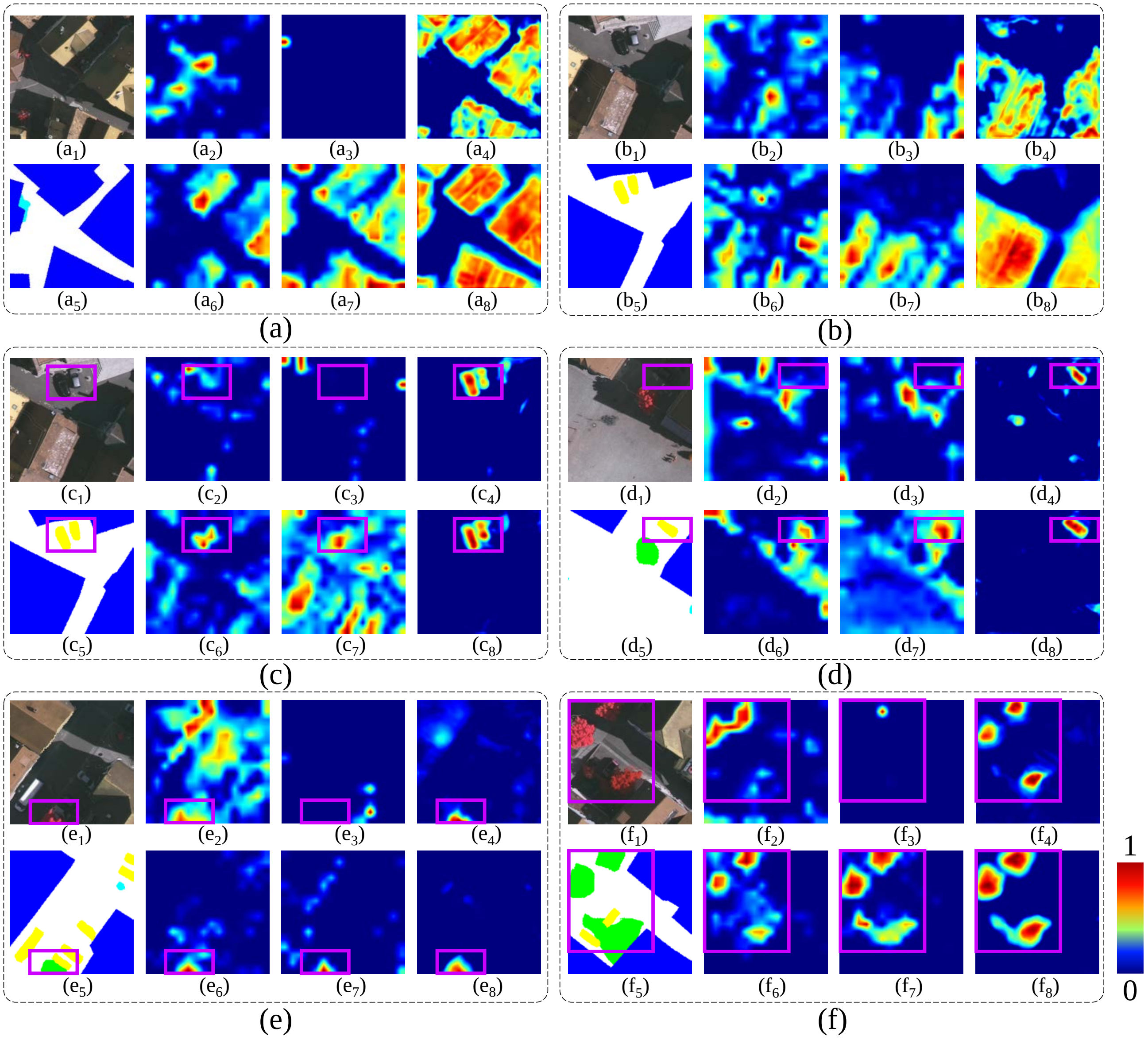

Fig. 9. Six groups of heatmap samples. In subfigure (a): (a_1) NIIRG, (a_2-a_4) three heatmaps from TransUNet, (a_5) Grounb Truth, (a_6-a_8) three heatmaps from FTransUNet. Subfigure (b-f) are organized in the same way. Subfigure (a-b), (c-d) and (e-f) show the two models of how to determine a pixel belongs to *Building*, *Car* and *Tree*, respectively. Compared to *Building*, *Car* and *Tree* have smaller shapes, so we added purple boxes to indicate the areas of concern.

TABLE VII
ABLATION STUDY OF THE PROPOSED APPROACH, WHERE SFF REFERS TO SHALLOW-LEVEL FEATURE FUSION IN CNN BLOCKS AND DFF REPRESENTS DEEP-LEVEL FEATURE FUSION IN FVIT. THE RESULT SHOWS THAT IT IS NECESSARY TO FUSE MULTIMODAL INFORMATION IN SHALLOW-LEVEL AND DEEP-LEVEL FEATURES SIMULTANEOUSLY.

Bold values are the best.

SFF	DFF	OA(%)	mF1(%)	mIoU(%)
✓		92.13	90.60	83.25
	✓	91.73	90.67	83.32
✓	✓	**92.40**	**91.21**	**84.23**

enhancement, deep-level feature fusion and enhancement. This exemplifies that the proposed FTransUNet differs from existing works by leveraging both CNN and transformer to extract and fuse features simultaneously.

F. Ablation Study

To verify the effectiveness of each proposed component in FTransUNet, ablation experiments were carried out by removing specific components while keeping the dual-branch framework. As shown in Table VII, two ablation experiments were designed based on our fusion scheme. In the first experiment, the proposed FVit was disassembled into two single-modal Vit modules, i.e., two independent SA-based transformers, while the SFF modules in CNN blocks are kept as before. In contrast, the second experiment removed the SFF modules in CNN with shallow-level features from two modalities individually extracted by the two branches.

Inspection of Table VII suggests that both shallow-level and deep-level feature fusion modules are critical for the proposed FTransUNet to provide improved performance. More specifi-

TABLE VIII
COMPUTATIONAL COMPLEXITY ANALYSIS MEASURED BY A 256×256 IMAGE ON A SINGLE NVIDIA GEFORCE RTX 3090 GPU. MIOU VALUES ARE THE RESULTS ON THE VAIHINGEN DATASET. BOLD VALUES ARE THE BEST.

Method	**Multimodal**	**FLOPs (G)**	**Parameter (M)**	**Memory (MB)**	**Speed (FPS)**	**MIoU(%)**
ABCNet [61]	N	**3.9**	**13.39**	1598	15.87	75.20
PSPNet [62]	N	49.03	46.72	3124	**66.01**	76.96
MAResU-Net [63]	N	8.79	26.27	1908	10.62	79.89
UNetFormer [44]	N	6.04	24.20	1980	13.89	81.97
vFuseNet [15]	Y	60.36	44.17	2618	16.93	78.92
FuseNet [14]	Y	58.37	42.08	2284	18.92	78.71
ESANet [33]	Y	7.73	34.03	1914	10.42	79.42
TransUNet [31]	Y	32.27	93.23	3028	10.81	78.26
SA-GATE [64]	Y	41.28	110.85	3174	10.00	81.27
CMFNet [19]	Y	78.25	123.63	4058	8.62	81.44
MFTransUNet [59]	Y	8.44	43.77	**1549**	14.88	81.61
CMGFNet [47]	Y	19.51	64.20	2463	11.61	82.26
FTransUNet	Y	45.21	160.88	3463	9.74	**84.23**

cally, shallow-level feature fusion can learn and provide robust representations of primary features of ground objects, such as shapes, boundaries, colors and textures, regardless of the scale variations of ground objects. In addition, deep-level feature fusion helps to distinguish complex remote sensing scenes by exploiting semantic information derived from shallow-level feature fusion. To clearly demonstrate the effectiveness of the multilevel fusion scheme, we present the heatmaps generated by the baseline TransUNet and the proposed FTransUNet in Fig. 9 of six subfigures. Labels have been added in Fig. 9 to describe each subfigure. The three heatmaps from FTransUNet were collected after shallow-level fusion, after deep-level fusion and before segmentation head, respectively. The three heatmaps generated by TransUNet were collected from the corresponding layers. The heatmaps in Fig. 9(a-b), Fig. 9(c-d) and Fig. 9(e-f) show how the two models determined a pixel belonging to *Building*, *Car* and *Tree*, respectively. Firstly, it is easily observed that for all samples, there are more activated (high-value) regions in our method, which means that the FTransUNet can provide more useful global semantics at the overall map scale to facilitate the final prediction with the help of deep-level fusion. Secondly, the contour of our activated regions is closer to the boundaries of actual ground objects, proving that the shallow-level fusion can extract more local details. The visualization results further confirmed that the proposed multilevel fusion scheme extracted and fused multimodal data more effectively, which led to better semantic segmentation performance.

G. Model Complexity Analysis

We evaluate the computational complexity of the proposed FTransUNet using the following evaluation metrics: the floating point operation count (FLOPs), the number of model parameters, the memory footprint and the frames per second (FPS). FLOPs is used to evaluate the model complexity whereas the number of model parameters and the memory footprint are used to evaluate the memory requirement. Finally, FPS is designed to assess the execution speed. Ideally, an efficient model should have smaller values in the first three metrics, but larger FPS value.

Table VIII shows the complexity analysis results of all comparing methods considered in this work. Inspection of Table VIII shows that the proposed FTransUNet exhibited lower FLOPs than conventional FuseNet, vFuseNet and CMFNet despite we have a larger number of parameters. However, it is observed that the proposed FTransUNet demonstrated better performance than CMFNet in terms of memory occupancy and FPS. This is because that FTransUNet used a two-branch encoder and only calculated attention with the global receptive field at the deep-level. Furthermore, compared to light-weight networks, such as UNetFormer, MFTransUNet and CMGFNet, the proposed FTransUNet provided noticeable segmentation performance improvement with slightly increased model complexity.

V. CONCLUSION

In this work, a novel multilevel multimodal fusion scheme, FTransUNet, has been proposed for semantic segmentation of remote sensing data by exploiting the synergy of CNN and Vit-based fusion. In particular, a CNN-based architecture equipped with the SFF modules is designed to extract and fuse detailed shallow-level features across multiple scales, followed by a Fusion Vision transformer (FVit) that performs deep-level semantic feature extraction and fusion. The proposed three-stage FVit can effectively characterize the complex content of remote sensing data by capitalizing on the novel Ada-MBA modules that use SA to extract deep-level features and a mutual-association mechanism to guide the fusion of multimodal deep-level features. Extensive results on two datasets, ISPRS Vaihingen and Potsdam, have confirmed that the proposed FTransUNet can achieve superior performance as compared to state-of-the-art segmentation methods.

There are several extensions of this study that can be further explored. In particular, it remains challenging to distinguish *Tree* and *Low vegetation*. Thus, it is interesting to exploit new strategies for ground objects of similar colors and irregular boundaries. Furthermore, since it is labor-intensive to generate DSM data for high-resolution remote sensing images, it is of great practical interest to explore image-based elevation estimation for downstream remote sensing tasks. Finally, it will be interesting to investigate the incorporation of large-scale models such as segment anything model (SAM) into the semantic segmentation framework in remote sensing.

References

[1] R. Touati, M. Mignotte, and M. Dahmane, "Multimodal change detection in remote sensing images using an unsupervised pixel pairwise-based markov random field model," *IEEE Transactions on Image Processing*, vol. 29, pp. 757–767, 2019.

[2] L. T. Luppino, M. A. Hansen, M. Kampffmeyer, F. M. Bianchi, G. Moser, R. Jenssen, and S. N. Anfinsen, "Code-aligned autoencoders for unsupervised change detection in multimodal remote sensing images," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2022.

[3] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4340–4354, 2020.

[4] Y. Li, Y. Zhou, Y. Zhang, L. Zhong, J. Wang, and J. Chen, "DKDFN: Domain knowledge-guided deep collaborative fusion network for multimodal unitemporal remote sensing land cover classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 186, pp. 170–189, 2022.

[5] Z. Xu, Z. Shen, Y. Li, L. Xia, H. Wang, S. Li, S. Jiao, and Y. Lei, "Road extraction in mountainous regions from high-resolution images based on dsdnet and terrain optimization," *Remote Sensing*, vol. 13, no. 1, p. 90, 2020.

[6] Y. Meng, S. Chen, Y. Liu, L. Li, Z. Zhang, T. Ke, and X. Hu, "Unsupervised building extraction from multimodal aerial data based on accurate vegetation removal and image feature consistency constraint," *Remote Sensing*, vol. 14, no. 8, p. 1912, 2022.

[7] F. Palsson, J. R. Sveinsson, M. O. Ulfarsson, and J. A. Benediktsson, "Model-based fusion of multi-and hyperspectral images using pca and wavelets," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2652–2663, 2014.

[8] Q. Wei, J. Bioucas-Dias, N. Dobigeon, and J.-Y. Tourneret, "Hyperspectral and multispectral image fusion based on a sparse representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3658–3668, 2015.

[9] Y. Shen, J. Chen, L. Xiao, and D. Pan, "Optimizing multiscale segmentation with local spectral heterogeneity measure for high resolution remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 157, pp. 13–25, 2019.

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

[11] X. Lu, J. Zhang, T. Li, and G. Zhang, "Synergetic classification of long-wave infrared hyperspectral and visible images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 7, pp. 3546–3557, 2015.

[12] L. Gao, J. Li, M. Khodadadzadeh, A. Plaza, B. Zhang, Z. He, and H. Yan, "Subspace-based support vector machines for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 2, pp. 349–353, 2014.

[13] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," *Advances in neural information processing systems*, vol. 24, pp. 1–9, 2011.

[14] 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*, pp. 213–228, 2016.

[15] 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.

[16] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, "A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[17] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–10, 2021.

[18] 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.

[19] 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.

[20] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[21] X. Zhang, B. Zhang, W. Yu, and X. Kang, "Federated deep learning with prototype matching for object extraction from very-high-resolution remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[22] L. Gao, H. Liu, M. Yang, L. Chen, Y. Wan, Z. Xiao, and Y. Qian, "STransFuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 10990–11003, 2021.

[23] 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.

[24] 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.

[25] J. Zhu, Y. Guo, G. Sun, L. Yang, M. Deng, and J. Chen, "Unsupervised domain adaptation semantic segmentation of high-resolution remote sensing imagery with invariant domain-level prototype memory," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–18, 2023.

[26] 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.

[27] 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.

[28] 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.

[29] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[30] 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*, pp. 234–241, 2015.

[31] 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.

[32] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-Unet: Unet-like pure transformer for medical image segmentation," 2022.

[33] 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)*, pp. 13525–13531, 2021.

[34] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1090–1099, 2022.

[35] 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.

[36] D. Hong, L. Gao, R. Hang, B. Zhang, and J. Chanussot, "Deep encoder–decoder networks for classification of hyperspectral and lidar data," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2020.

[37] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1341–1360, 2020.

[38] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot, "Multimodal fusion transformer for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[39] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[40] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 135, pp. 158–172, 2018.

[41] K. Nogueira, M. Dalla Mura, J. Chanussot, W. R. Schwartz, and J. A. Dos Santos, "Dynamic multicontext segmentation of remote sensing images based on convolutional networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 7503–7520, 2019.

[42] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7262–7272, 2021.

[43] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12077–12090, 2021.

[44] 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.

[45] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "Ds-transunet: Dual swin transformer u-net for medical image segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–15, 2022.

[46] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "Uctransnet: rethinking the skip connections in U-net from a channel-wise perspective with transformer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, pp. 2441–2449, 2022.

[47] 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.

[48] 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.

[49] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, pp. 689–696, 2011.

[50] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[51] Q. He, X. Sun, W. Diao, Z. Yan, F. Yao, and K. Fu, "Multimodal remote sensing image segmentation with intuition-inspired hypergraph modeling," *IEEE Transactions on Image Processing*, vol. 32, pp. 1474–1487, 2023.

[52] W. Zhou, J. Jin, J. Lei, and L. Yu, "CIMFNet: Cross-layer interaction and multiscale fusion network for semantic segmentation of high-resolution remote sensing images," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 666–676, 2022.

[53] J. Ma, W. Zhou, J. Lei, and L. Yu, "Adjacent bi-hierarchical network for scene parsing of remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[54] 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*, pp. 10012–10022, 2021.

[55] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, 2019.

[56] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7077–7087, 2021.

[57] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "Swinfusion: cross-domain long-range learning for general image fusion via swin transformer," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 7, pp. 1200–1217, 2022.

[58] L. Yan, J. Huang, H. Xie, P. Wei, and Z. Gao, "Efficient depth fusion transformer for aerial image semantic segmentation," *Remote Sensing*, vol. 14, p. 1294, 2022.

[59] 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.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2016.

[61] 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.

[62] 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*, pp. 2881–2890, 2017.

[63] 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.

[64] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, and G. Zeng, "Bi-directional cross-modality feature propaga-

tion with separation-and-aggregation gate for RGB-D semantic segmentation," in *European Conference on Computer Vision*, pp. 561–577, 2020.

[65] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, IEEE, 2009.

Xianping Ma received his Bachelor in Geographical Information Science, Wuhan University, China, in 2019. He is currently pursuing the PhD degree at the Chinese University of Hong Kong, Shenzhen, China. His research interests include remote sensing image processing, deep learning, multimodal learning and unsupervised domain adaptation.

He has reviewed for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing and APSIPA Transactions on Signal and Information Processing.

Xiaokang Zhang (Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from The School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, in 2018. From 2019 to 2022, he was a Postdoctoral Research Associate with The Hong Kong Polytechnic University, Hong Kong, and The Chinese University of Hong Kong, Shenzhen, China. Since 2023, he has been a specially appointed Professor with the School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan, China. He has authored or coauthored more than 30 scientific publications in international journals and conferences. His research interests include remote sensing image analysis, computer vision and machine learning.

Dr. Zhang is currently a Reviewer for more than 20 renowned international journals, such as Remote Sensing of Environment, ISPRS Journal of Photogrammetry and Remote Sensing, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS and the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Man-On Pun (Senior Member, IEEE) received his BEng. degree in Electronic Engineering from the Chinese University of Hong Kong (CUHK) in 1996, the MEng. degree in Computer Science from University of Tsukuba, Japan in 1999 and the Ph.D. degree in Electrical Engineering from University of Southern California (USC) in Los Angeles, U.S.A in 2006, respectively. He was a postdoctoral research associate at Princeton University from 2006 to 2008.

Currently, he is an associate professor at the School of Science and Engineering, the Chinese University of Hong Kong, Shenzhen (CUHKSZ). Prior to joining CUHKSZ in 2015, he held research positions at Huawei (USA) in New Jersey, Mitsubishi Electric Research Labs (MERL) in Boston and Sony in Tokyo, Japan.

Prof. Pun's research interests include AI Internet of Things (AIoT) and applications of machine learning in communications and satellite remote sensing. Prof. Pun has received best paper awards from IEEE VTC'06 Fall, IEEE ICC'08 and IEEE Infocom'09. He served as associate editor for the IEEE Transactions on Wireless Communications in 2010 - 2014. He is the founding chair of the IEEE Joint SPS-ComSoc Chapter, Shenzhen.

Ming Liu graduated from Harbin Institute of Technology, majoring in Computer Science and Technology, and has been working in the technology field for more than 15 years, with rich successful experience in radio resource management, algorithm innovation and design, product implementation and field validation for wireless GSM, UMTS, LTE / 5G-NR networks. Prior to founding MizarVision Technology, he worked at Huawei's wireless department in Shanghai for a long time.

During his time at Huawei, Ming Liu, as the LTE wireless solution area leader and wireless system simulation area leader, was involved in the research and development of Huawei Research Institute in North America to explore and apply machine learning and AI technologies in the wireless network operation, optimization and design space. As a principal, he led the joint North America/Shanghai/Chengdu team to complete the system simulation modeling and system simulation platform software and hardware design and development of 3G/4G/5G wireless network communication protocols/wireless channel propagation/wireless receivers and other key technologies in the wireless network simulation space. In addition, Ming Liu has extensive experience in leading/managing teams of technologists, managing multi-team/cross-organizational multi-disciplinary teamwork across geographic locations, and a broad background in pre-sales RFP / RFI preparation/response/strategy/defense.