Deep Multimodal Fusion Network for Semantic Segmentation Using Remote Sensing Image and LiDAR Data

Yangjie Sun, Zhongliang Fu, Chuanxia Sun, Yinglei Hu, and Shengyuan Zhang

***Abstract*— Extracting semantic information from very-high-resolution (VHR) aerial images is a prominent topic in the Earth observation research. An increasing number of different sensor platforms are appearing in remote sensing, each of which can provide corresponding multimodal supplemental or enhanced information, such as optical images, light detection and ranging (LiDAR) point clouds, infrared images, or inertial measurement unit (IMU) data. However, these current deep networks for LiDAR and VHR images have not fully utilized the complete potential of multimodal data. The stacked multimodal fusion network (MFNet) ignores the structural differences between the modalities and the manual statistical characteristics within the modalities. For multimodal remote sensing data and its corresponding carefully designed handcrafted features, we designed a novel deep MFNet that can use multimodal VHR aerial images and LiDAR data and the corresponding intramodal features, such as LiDAR-derived features [slope and normalized digital surface model (NDSM)] and imagery-derived features [infrared–red–green (IRRG), normalized difference vegetation index (NDVI), and difference of Gaussian (DoG)]. Technically, we introduce the attention mechanism and multimodal learning to adaptively fuse intermodal and intramodal features. Specifically, we designed a multimodal fusion mechanism, pyramid dilation blocks, and a multilevel feature fusion module. Through these modules, our network realized the adaptive fusion of multimodal features, improved the receptive field, and enhanced the global-to-local contextual fusion effect. Moreover, we used a multiscale supervision training scheme to optimize the network. Extensive experimental results and ablation studies on the ISPRS semantic dataset and IEEE GRSS DFC Zeebrugge dataset show the effectiveness of our proposed MFNet.**

***Index Terms*— Aerial images, attention mechanism, convolutional neural network (CNN), multimodal fusion, semantic labeling.**

I. Introduction

THE breakthrough of remote sensing sensor technology, such as multispectral sensors, hyperspectral sensors and light detection and ranging (LiDAR) systems, has made the acquisition of multimodal data increasingly convenient. These different types of sensors or devices can form multimodal cross mapping and perception. For example, in areas occluded by trees or shadows, advanced LiDAR systems could provide excellent complementary information for very-high-resolution (VHR) images. Semantic segmentation consists of assigning a label to each pixel in an image. This has been applied to precision agriculture, urban planning, and natural resources inventory [1]. Developing a segmentation network using rich multimodal spatial information is a current remote sensing research frontier. For example, using a combination of multispectral data and the corresponding LiDAR point cloud data, the shadow areas of buildings or cars can be segmented more accurately.

However, the complexity of VHR images and the multimodality between images and point clouds bring several challenges. For example, in urban building areas of the VHR image, some building roofs are complex and diverse. This is a typical problem, where the same roofs have different spectra. At the same time, they are also affected by occlusion and shadow, and their accurate labeling is still a challenge for VHR aerial image segmentation. On the other hand, most of the current semantic segmentation algorithms simply stack multimodal data together as multichannel inputs. Thus, in addition to traditional three-channel RGB images, how to jointly use complementary multimodal remote sensing information to maximize labeling accuracy is a new problem.

Generally speaking, for VHR image segmentation, previous excellent algorithms used supervised training classifiers and designed the corresponding handcrafted features to obtain semantic information. Simultaneously, researchers have also tried unsupervised methods. For example, Ye and Wang [2] first used graph theory-based methods to segment VHR remote sensing images. Fu *et al.* [3] used superpixels and graph theory for VHR image multiscale segmentation. However, these methods have obvious disadvantages where these handcrafted features usually come from image appearance information, such as color, texture, and gradient, and it is difficult to express deeper and more abstract features. In addition, the coefficients of the handcrafted feature descriptor and the parameters of the optimal segmentation require significant manual tuning [4], [5]. Selection of the features may be an effective proxy [6]–[8], which is data-driven, but it is still not a complete learning-based method. As shown in Fig. 1, these features are typically complementary, including handcrafted intramodal features, such as spectrum, geometry, and texture,

Manuscript received March 17, 2021; revised July 18, 2021; accepted August 7, 2021. Date of publication September 23, 2021; date of current version January 26, 2022. This work was supported in part by the Research on Highway Video Surveillance and Perception Complete Technology Based on Big Data Analysis, Highway Administration Bureau of Henan Transportation Department, under Grant 2019G1. *(Corresponding authors: Zhongliang Fu; Yangjie Sun.)*

Yangjie Sun, Zhongliang Fu, and Shengyuan Zhang are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: yjs@whu.edu.cn; fuzl@whu.edu.cn; zshengyuan@126.com).

Chuanxia Sun and Yinglei Hu are with the Highway Administration Bureau, Henan Transportation Department, Zhengzhou 450046, China (e-mail: 801715@qq.com; 462869327@qq.com).

Digital Object Identifier 10.1109/TGRS.2021.3108352

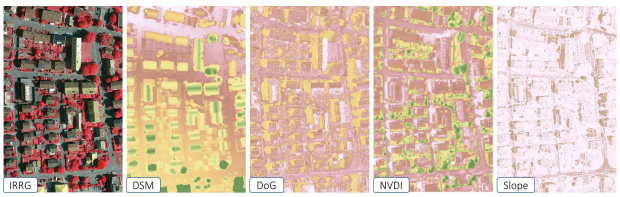


Fig. 1. Illustration of some handcrafted feature maps.

and intermodal features from different modalities, e.g., RGB images, LiDAR point clouds, infrared images, or inertial measurement unit (IMU) data. Clearly, these traditional methods have not exploited the full potential of multimodal data.

Fully convolutional networks (FCNs) [9], [10] from the field of computer vision (CV) may provide some solutions to these problems. Different from traditional convolutional neural networks (CNNs) [11], FCNs change the fully connected (FC) layer to the convolutional layer to perform the pixel-by-pixel classification, which significantly improves the segmentation accuracy of the network. This simple and efficient structure of FCNs soon became the basic framework of the semantic segmentation task. However, the original FCN had multiple downsampling structures, and the output map was too rough. In this regard, researchers have made many improvements, such as SegNet and DeconvNet, which improved FCNs to a symmetric encode–decode structure and added a multilevel feature fusion (MFF) module [12], [13]. These designs have significantly improved the segmentation results. At the same time, in the 3-D vision and robotics community, researchers embed multimodal data collected by multiple sensors into the FCNs framework for a semantic analysis. Guo *et al.* [14] used the parallel FCNs structure for semantic segmentation and merged two modal RGB-D data in the middle layer of the network. Hazirbas *et al.* [15] used a parallel SegNet structure to perform semantic segmentation of RGB-D data, called FuseNet.

Despite their success, there are still some areas for improvement. In this article, we systematically consider the following issues.

1) For multimodal remote sensing data, most algorithms only combine the intermodal features extracted from a deep network and do not integrate well-designed handcrafted features into the network. Can a deep framework simultaneously exert the advantages of the intramodel and intermodal features?
2) Due to the information contained in different modal data not being equivalent for each semantic scene and because the current networks using the strategy of direct stack-fusion cannot sufficiently combine multimodal information (such as RGB and LiDAR) [16], [17], can we design an architecture that is able to yield a satisfactory performance via adaptive fusion of different modal features?
3) Notice that most of these existing methods are less effective at utilizing the complementarity between multiscale contexts, especially for small objects. Thus, how should we use the complementarity between multiscale features for accurate labeling?
4) Currently, most FCNs adopt a fusion strategy of directly stacking high- and low-level features in order to obtain a fine final feature map. However, the complementarity and dependencies between these high- and low-level features are ignored. Can we model the interrelationship between these features through the attention mechanism and adaptively enhance the representation ability of the feature map?

In fact, the questions mentioned above are where the motivation of our study lies. We thus attempt to build a common network framework to adaptively utilize multimodal VHR aerial images and LiDAR data. Our major contributions are summarized as follows.

1) A novel multimodal fusion network (MFNet) is proposed for adaptive multimodal remote sensing data fusion by interleaving intramodality and intermodality features. As far as we know, this is the first time that the soft-attention mechanism is utilized in the multimodal feature fusion framework with intramodality and intermodality for VHR aerial images segmentation tasks.
2) Two multimodal fusion blocks are proposed to adaptively fuse intramodality and intermodality features extracted from multimodal remote sensing data and to learn cross-modal interdependencies and complementary contextual information.
3) A pyramid dilation (PD) module is proposed to effectively combine features from different scales by integrating multiscale receptive fields step-by-step. This greatly improves the labeling results of confusing man-made objects.
4) A skip-architecture MFF module is proposed, which can model the relationship between multiple levels of contextual information and obtain a fine resolution feature map. Extensive experiments and ablation studies show that the proposed MFNet can achieve the state-of-the-art segmentation performance.

The remaining sections of this article are organized as follows. In Section II, the attention mechanism and multimodal research are briefly introduced. Section III introduces the details of MFNet. In Section IV, we show the experiments and ablation studies, as well as a detailed analysis of MFNet. Conclusions and other discussions are given in Section V.

II. Related Work

A. Attention Mechanism

Attention mechanism stems from visual signal processing. When looking at a scene, attention will make our brain focus on the target area and ignore other insignificant information. Attention first appeared in the CV field in the 1990s. In 2014, the Google mind team [18] used it on the RNN and achieved satisfactory performance. Subsequently, Bahdanau *et al.* [19] were the first to use the attention mechanism for natural language processing. In 2017, the Google team proposed a self-attention structure in the paper "Attention is all you need." This caused an enormous response, making the attention mechanism a prominent topic in recent research, which has achieved success in various CV and NLP tasks. Benefiting

from the squeeze-and-excitation (SE) module, SENet [20] won the title of the ImageNet 2017 competition classification task. This module introduces the attention mechanism on the channel dimension, which can pay more attention to the channel features with more information and suppress the channel features with less information. Wang *et al.* [21] used the idea of nonlocality in neural networks to achieve an excellent performance in multiple CV tasks. The idea behind it is actually the generalized expression of positional attention and self-attention. Since then, a wave of research on the attention mechanism has been triggered, such as CBAM [22], DANet [23], SKNet [24], and CCNet [25].

B. Multimodal Learning

Multimodal learning refers to learning between different modalities, such as image, video, audio, and semantics. Since the 1970s, it has gone through several stages of development, entering the deep learning phase in 2010. In 2012, Srivastava and Salakhutdinov [26] proposed to extend the deep Boltzmann machine (DBM) structure to the multimodal learning domain. Through multimodal DBM, we can learn the multimodal joint probability distribution. Eitel *et al.* [27] combined multimodal RGB-D images for object recognition using a two-stream network. Similarly, Hazirbas *et al.* [15] designed a parallel SegNet structure for RGB-D semantic segmentation, namely FuseNet. Wang *et al.* [28] presented deeply fused nets to discuss the significant advantages of deep multimodal fusion in detail. Chen *et al.* [17] used multimodal data as the input and predictive 3-D targets via a multiview 3-D (MV3D) object recognition network.

C. Multimodal Classification Methods Within the Remote Sensing Field

Recently, similar to the RGB-D tasks in CV, various multimodal networks, multibranch, or fusion networks have been used for multisource remote sensing data. Indeed, different Earth observation sensors have favorable complementarity and can provide multimodal remote sensing information for the same scene. Therefore, Paisitkriangkrai *et al.* [29] combined CNN and handcrafted designed features to build a semantic segmentation model from both RGB images and the digital surface model (DSM). Since then, Sherrah [10] and Volpi and Tuia [30] both proposed a semantic segmentation fusion network based on the VHR image and LiDAR data. Similar to Paisitkriangkrai *et al.* [29], Liu *et al.* [31] used the CRF framework to fuse the deep features obtained by training and the handcrafted features. Audebert *et al.* [16] studied the basis of FuseNet and deeply investigated two multimodal fusion methods, namely early fusion and late residual correction fusion. The research results show that both fusion structures can use the advantages of multimodal data. Sun *et al.* [32] proposed the concept of intermodal and intramodal features and designed a multifilter CNN network to aggregate multimodal features. Finally, the traditional multiresolution segmentation (MRS) method was used to further improve the segmentation results. In DFC 2018, Xu *et al.* [33], based on FCN and postclassification strategy, proposed fusion-FCN using multispectral and video data and achieved the best classification results. In short, these fusion networks designed for multimodal remote sensing data have attempted to take advantage of LiDAR, DEM, DSM, hyperspectral, video, and VHR images for segmentation or classification tasks. Despite their success, there are still some areas for improvement. We have listed the detailed descriptions of the four issues in Section I.

Inspired by attention network and multimodal learning, this article attempted to propose a novel MFNet framework, which is the first to use the attention mechanism and intramodality and intermodality fusion framework for semantic labeling of VHR remote sensing images. We integrate the multisource remote sensing data and traditional handcrafted features into a multimodal network. First, the MFNet encodes the intermodality features and intramodality features through two multimodal fusion modules to aggregate more related information and suppress unrelated information. Second, the PD module combines multilevel features from different receptive fields, and then, the global fusion scheme makes the global-to-local contexts sequentially aggregated, forming finer feature maps. Finally, like Jiang *et al.* [34], we utilize a multiscale supervision strategy to optimize our MFNet and evaluate our method on the ISPRS segmentation dataset and IEEE GRSS DFC Zeebruge dataset.

III. Methods Description

As shown in Fig. 2, the MFNet is an FCNs framework that contains five main ingredients: a two branches encoder–decoder backbone network, two kinds of multimodal attention fusion blocks, a PD attention module, a multilevel feature attention module, and a multiscale supervision training scheme.

A. Encoder–Decoder Backbone Network

The encoder–decoder backbone network is a residual FCN architecture [34] that has two convolutional branches, i.e., the RGB branch and the LiDAR branch. The backbone network is specially designed for the diversity of remote sensing data. It can receive multimodal data from a variety of remote sensing sensors and the corresponding intramodal features, such as LiDAR-derived features [intensity, elevation, slope angle, and normalized DSM (NDSM)], and imagery-derived features [infrared–red–green (IRRG) bands data, normalized difference vegetation index (NDVI), and scale-invariant feature transform (SIFT)]. These multimodal data and intramodal features are used as the input of the encoder network.

The encoder part of each branch is largely based on ResNet-50 [35]. This consists of a series of residual blocks, the input of which is added into the output using a skip connection. We use different colors to represent different blocks in Fig. 2. The first encoder block is composed of a 7×7 convolutional layers with stride 2 and a 3×3 max-pooling layer with stride 2, followed by a batch-normalized (BN) layer and an rectified linear unit (ReLU) layer. This can make our network flexible to deal with different

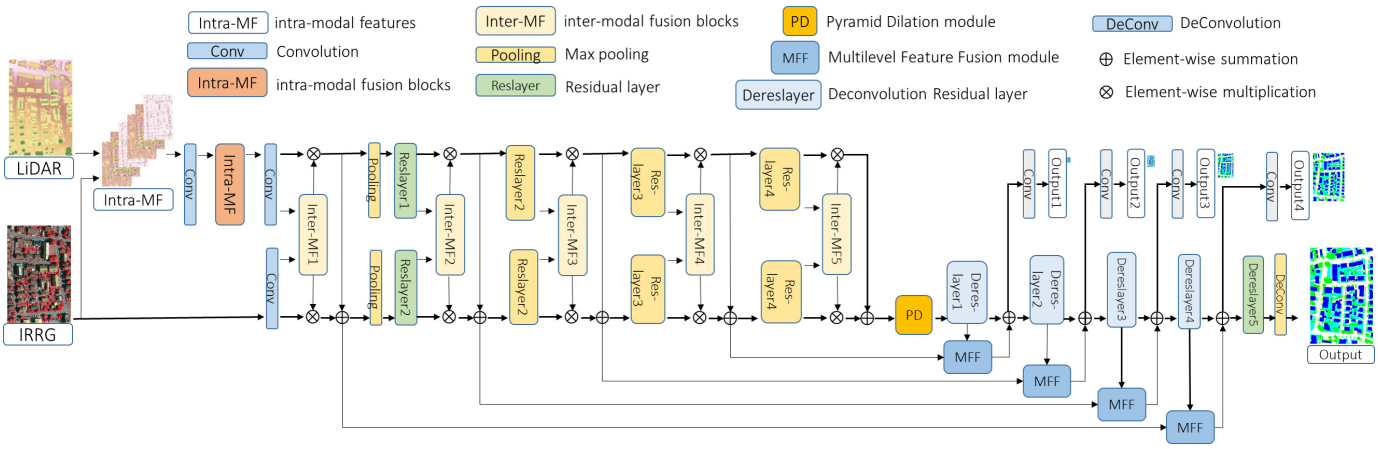


Fig. 2. Overview of the MFNet.

bands input and adaptively cope with the multimodal input data.

Notice that each convolution operation is followed by a BN layer before the ReLU function, and it is omitted in Fig. 2 for simplification. After that, four residual layers are sequentially added. In addition to the first one, the other three residual layers have one residual unit that downsamples the feature map and increases the feature channel by a factor of 2. Thus, within the encoder network, the final output resolution is 1/32 of the original input data. The two branches' features are elementwise summation fused into one branch on the five layers.

The structure of the decoder part is symmetrical with respect to the encoder part. First, four deconvolution layers with 2× upsampling residual units are sequentially added at the end of the encoder part. Similar to the encoder part, the fifth deconvolution layer of the decoder network does not have an upsample unit. Second, the final layer is a single 2 × 2 deconvolution layer and the size of the final output is equal to the input. We add these encoders' features into the decoder features using the skip concatenation function to enable the decoder network to obtain finer feature maps. Finally, our network has a typical FCNs output and four side outputs.

B. Multimodal Fusion Blocks

Inspired by the work of SENet [20] and SKNet [24], we have specially designed two multimodal fusion blocks to deal with the multimodal features' fusion.

1) Intramodal Fusion Block: Diverse remote sensing data have a good complementarity and contains diverse information in different urban areas. In order to selectively learn more useful features from these remote sensing data, we borrowed the global average pooling mechanism [20] and designed the intramodal fusion ($\text{IMA}_{\text{intra}}$) block, which enables our network to efficiently focus on the intramodal features fusion. The $\text{IMA}_{\text{intra}}$ block begins with intramodal fusion, which gathers features derived from different modal data, such as LiDAR-derived features (elevation and NDSM) and imagery-derived features (IRRG, NDVI, and SIFT). Then, as shown in Fig. 3, given these multimodal feature maps, $F \in \mathrm{R}^{n\times h\times w}$ as inputs. We first perform global average pooling to obtain the global

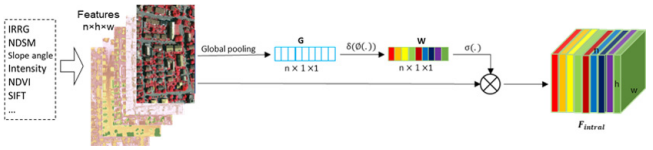


Fig. 3. Intramodal fusion ($\text{IMF}_{\text{intra}}$) block.

multimodal information in a channel descriptor $G \in \mathrm{R}^{n\times 1\times 1}$, where n is the channels number and w and h are the width and height of F, respectively. Formally, G can be calculated by

$$G_i = F_{\text{gp}}(F_i) = \frac{1}{h \times w}\sum_{p=1}^{h}\sum_{q=1}^{w} F_i(p,q). \tag{1}$$

Next, G_c is obtained by performing a 1 × 1 convolution followed by a BN, which can fully capture the multimodal channelwise dependencies. Then, G_c is activated by a softmax function and we can obtain the intramodal weights W. Finally, the output is F_{intra}, and these calculations are expressed as follows:

$$G = F_{\text{gp}}(F) \tag{2}$$

$$G_c = \delta(\emptyset(G)) \tag{3}$$

$$W = \sigma(G_c) \tag{4}$$

$$F_{\text{intra}} = W \otimes F \tag{5}$$

where F_{gp} is the global average pooling function, $\otimes$ denotes the elementwise multiplication, σ denotes the softmax function, δ is the ReLU function, and $\emptyset$ refers to the 1 × 1 convolution. In this regard, the $\text{IMA}_{\text{intra}}$ blocks intrinsically introduce dynamics conditioned on the intramodal features, helping to improve feature discriminability.

2) Intermodal Fusion Block: The two branches of the encoder network can learn the characteristics of the different modalities separately. We designed the intermodal fusion ($\text{IMA}_{\text{inter}}$) block to focus on the fusion of these intermodal features. Different from $\text{IMA}_{\text{intra}}$, $\text{IMA}_{\text{inter}}$ is designed with multiple branches to aggregate information on multiple paths. At the same time, an FC layer was added to mimic the effects of stimuli in multimodal fusion, enabling $\text{IMA}_{\text{inter}}$ to make precise and adaptive selection features and reduce the dimensions to increase the efficiency. As shown in Fig. 4,

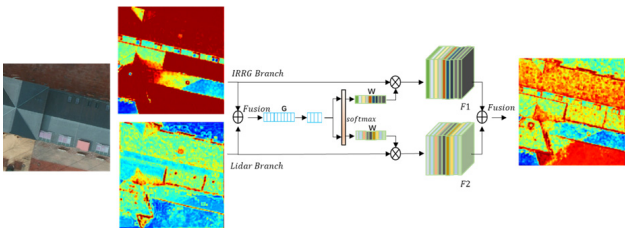


Fig. 4. Intermodal fusion ($\text{IMF}_{\text{inter}}$) block.

we first fuse features from two branches via an elementwise summation. Like Fig. 3, we use global average pooling F_{gp} to obtain channelwise multimodal feature statistics $G \in \text{R}^{n\times 1\times 1}$. Then, a compact feature $U \in \text{R}^{d\times 1\times 1}$ is created by an FC layer. These specific formulas are as follows:

$$G = F_{\text{gp}}(\text{Fusion}) \tag{6}$$

$$U = F_{\text{fc}}(G) = \delta(\text{Bn}(w \cdot G)) \tag{7}$$

$$d = \max\left(\frac{n}{r}, l\right) \tag{8}$$

where δ refers to the ReLU function, Bn denotes the batch normalization, F_{fc} is the FC layer, $w \in \text{R}^{d\times n}$ is the weight vector of F_{fc}, r is the ratio of dimension reduction, l is the minimum dimension, and $l = 32$. Finally, U is activated by the softmax operator on the two branches channelwise digits

$$w_1 = \frac{e^{aU}}{e^{aU} + e^{bU}}, \quad w_2 = \frac{e^{bU}}{e^{aU} + e^{bU}} \tag{9}$$

$$F_1 = w_1 \cdot \text{IRRGBranch} \tag{10}$$

$$F_2 = w_2 \cdot \text{LidarBranch2} \tag{11}$$

$$w_1^i + w_2^i = 1 \tag{12}$$

where a and b are the weight vectors and w_1 and w_2 denote the new weight vector after the softmax operator for IRRGBranch and LidarBranch, respectively. F_1 and F_2 are the output features of the two branches, respectively, and w_1^i and w_2^i are the ith row elements of w_1 and w_2, respectively.

C. PD Module

The objects of the VHR remote sensing images often have multiple scales, and their scale intervals are enormous. It is difficult to obtain appropriate information using only single-scale features. For the large gap between the scales and receptive field, the identification of large-scale objects is incomplete, and small-scale objects introduce too much irrelevant information, which often leads to an inability to identify them. Thus, a novel PD module is designed by combining the pyramid structure and dilation convolution. Inspired by [36] and [37], we use cascade dilation convolution in a pyramid structure with the prior global context.

As shown in Fig. 5(a), the receptive field of the convolution operation can be increased by defining a kernel gap. Specifically, when the gap is equal to 2, one pixel is skipped, and when the gap is equal to 4, three pixels are skipped. More details can be viewed here [38]. The PD module contains dilated convolution in the pyramid structure with multiple levels of receptive fields. In Fig. 5(b), from top to bottom, we use the dilation rate of 1, 2, 4, 8, and 16 in the PD module. Thus, the receptive fields of each layer will be 3, 7, 15, 31, and 63, respectively. The pyramid structure integrates multiscale receptive fields step-by-step and enhances the network's ability to perceive multiscale objects. At the same time, we also designed a 1×1 convolution and global average pooling branch to the PD module. Then, three output features of the PD are fused by pixelwise multiplication and concatenation. The PD module can obtain more global and multiscale coding features without losing the original features and greatly improve the labeling results of confusing man-made objects.

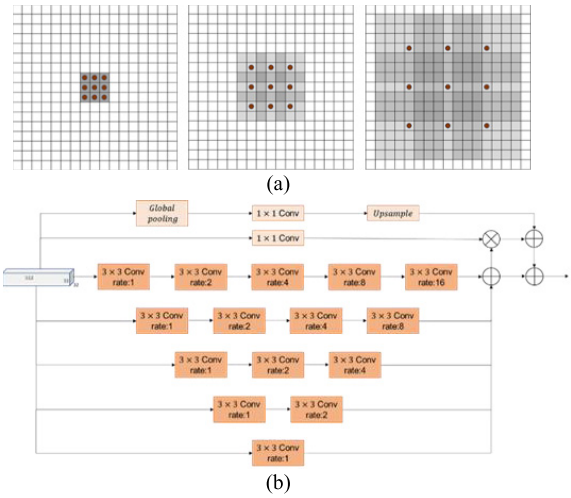


Fig. 5. (a) Red dots are the inputs of a 3×3 filter, and the gray area is the receptive field. (Left) One-dilated convolution; the receptive field is 3×3. (Middle) Two-dilated convolution; the receptive field is 7×7. (Right) Four-dilated convolution; the receptive field is 15×15. (b) PD module.

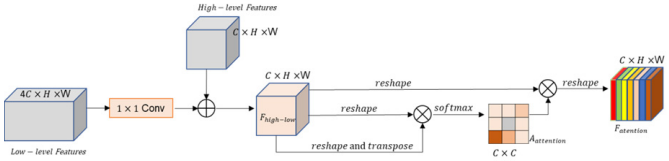


Fig. 6. MFF module.

D. MFF Module

Since there are multiple downsamplings in the FCN, the final feature map is coarser and often produces "blobby" labeling results. The current semantic segmentation method directly combining high- and low-level features is not sufficiently efficient. In particular, high-level features have rich semantic information and shallow features can bring important spatial information. Li *et al.* [39] proposed a global average pooling module to obtain global features as guidance for the shallow feature information. This mechanism is not sufficiently comprehensive because it only utilizes the semantic information to guide low-level feature empowerment, and the local spatial information (e.g., corners and edges) of low-level features is equally important.

Inspired by DANet [23], we use a skip-architecture MFF module to simultaneously fuse global to local information

from high- and low-level features. In Fig. 6, first, the low-level feature is projected through a 1×1 convolution feature map to the same channel number of the high-level feature, allowing them to have a pixelwise addition. Thus, we can obtain the fusion feature map $F_{\text{high-low}}$, and then, we reshape it to obtain $F_{\text{reshape}} \in \mathrm{R}^{C \times N}$ and transpose F_{reshape} to obtain $F_{\text{reshape-transpose}} \in \mathrm{R}^{N \times C}$. Second, we obtain F_w by the matrix multiplication of F_{reshape} and $F_{\text{reshape-transpose}}$ is, and then, the feature fusion map $A_{\text{fusion}} \in \mathrm{R}^{C \times C}$ is obtained through a softmax layer. A_{fusion} is calculated as follows:

$$a_{ji} = \frac{\exp\left(F_w^i \cdot F_w^j\right)}{\sum_{i=1}^{C} \exp\left(F_w^i \cdot F_w^j\right)} \tag{13}$$

where F_w^i and F_w^j are the ith and jth feature vectors of F_w, respectively, and a_{ji} indicates the impact of the ith feature vector of F_{reshape} on the jth feature vector. Finally, the transposition of A_{fusion} and F_{reshape} are subjected to matrix multiplication and then multiplied by a factor β to obtain the output feature F_{fusion}. Specifically,

$$F_{\text{fusion}}^j = \beta \sum_{i=1}^{C} \left(a_{ji} \cdot F_{\text{reshape}}^i\right) \tag{14}$$

where β is initialized to 0 and gradually learns to a larger weight, reshaping F_{output} to $F_{\text{fusion}} \in \mathrm{R}^{C \times H \times W}$. It can be seen from (14) that each feature of F_{fusion} is the weighted sum of all the features of the original feature maps $F_{\text{high-low}}$. Thus, the MFF module further highlights the global semantic information and local spatial information from high- and low-level features. By using the MFF module to model the relationship between multiple levels of features, we can obtain a fine resolution and a highly discriminative feature map.

E. Multiscale Supervision Training Scheme

We use the multiscale supervision training scheme of RedNet [34] to optimize the network. As shown in Fig. 2, our decoding network has four side outputs: Out4, Out3, Out2, and Out1, whose heights and widths are 1/2, 1/4, 1/8, and 1/16 of the final output, respectively. Then, each output will calculate its loss by the above formula. Therefore, the total cross-entropy loss is the joint calculation of all five outputs

$$\text{loss} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{k} -y_j^i \log\left(\frac{\exp\left(z_j^i\right)}{\sum_{l=1}^{k} \exp\left(z_l^i\right)}\right) \tag{15}$$

where N is the total pixel numbers, k is the class number, z_j^i is the predictions of a pixel at index i for the jth class, y_j^i denotes the label, and $(z_1^i, z_2^i, \ldots, z_k^i)$ is the output prediction vector.

Considering the class unbalanced problem of ISPRS and IEEE DFC datasets, we also use median frequency balancing [40] to calculate the class weights and assign them into loss function to reduce the impact of class imbalance.

IV. Experiments and Analysis

Qualitative and quantitative comparisons, ablation experiments, and the corresponding data and settings are described in this section.

A. Dataset Description

The ISPRS Vaihingen and Potsdam 2-D semantic labeling contest dataset [41] are used to validate our method. The datasets depict six different semantic classes, namely, impervious surfaces (white), buildings (blue), low vegetation (cyan), trees (green), cars (yellow), and clutter/background (red). The Vaihingen dataset consists of 33 VHR images with a resolution of 9 cm/pixel and a size of 2500×2000 pixels. Each image contains IRRG bands, as well as a LiDAR point cloud and corresponding NDSM data [42]. Since the ISPRS 2-D semantic labeling contest dataset has been made public, we are consistent with the normal settings of the competition [43]. Specifically, the 17 public ground truth images are still used as the test set, five images (ID 11, 15, 28, 30, and 34) are used as the verification set, and the remaining 11 images are used as the training set. The Potsdam dataset consists of 38 VHR images with a resolution of 5 cm/pixel and a size of 6000×6000 pixels. Each image contains IRRG bands, as well as a LiDAR point cloud and corresponding NDSM data [42]. Similar to the Vaihingen dataset, the original ground truth images are still used as the test set, and the remaining five images (ID 7_7, 7_8, 7_9, 7_11, and 7_12) are used as the verification set. Because it has error annotations [43], we did not use it to train the network in our experiments.

We also conducted experiments on the IEEE GRSS DFC Zeebrugge dataset [44] to verify the generalization ability of MFNet. The Zeebrugge dataset has eight different semantic classes, namely, impervious surface (white), buildings (blue), low vegetation (cyan), trees (green), cars (yellow), clutter (red), boats (pink), and water (dark blue). It consists of seven VHR images (ID 1, 2, 3, 4, 5, 6, and 7) with a resolution of 5 cm/pixel and a size of $10\,000 \times 10\,000$ pixels. Each image contains RGB bands, as well as a LiDAR 3-D point cloud data containing X (latitude), Y (longitude), Z (elevation), and I (intensity) and a corresponding max 5000×5000 pixel-sized DSM data at a 10-cm resolution (Lagrange.2015). The five images (ID 1, 2, 4, 5, and 7) are used as the training set and two images (ID 3 and 6) are used as the testing set. The two test images can be uploaded to the IEEE GRSS DFC website to obtain accuracy and F-measures.

The input of MFNet first receives multimodal remote sensing data and corresponding intramodal features. For a LiDAR point cloud, which reflects a wealth of geometric information, we use the elevation, slope angle, and intensity as point cloud handcrafted features. For the images, there are many excellent handcrafted feature descriptors. Here, we choose the difference of Gaussian (DoG) feature based on two different Gaussian kernels, which is consistent with [32]. The DoG descriptor is insensitive to noise and preserves accurate boundary localization while smoothing the image [45]. Then, these multimodal features are shown in Table I.

B. Experimental Setting

1) Implementation Details: All the experiments were conducted on two 11-GB Nvidia 1080ti graphics cards and the PyTorch framework. In the experiment, the parameters of the encoder part of our model use ResNet-50 [35] to initialize,

TABLE I
MULTIMODAL FEATURES OF HIGH-RESOLUTION IMAGERY AND LiDAR DATA

Category	Name	Meaning	Formula
Raw data	RGB	Red, Green and Blue bands	
	IR	Infrared band	
	DSM	Digital surface model	
Intra-modal features	Slope angle		$\frac{\sum_{i=1}^{k} \arctan\left(\frac{\lvert z - z_i \rvert}{\sqrt{(x-x_i)^2 + (y-y_i)^2}}\right)}{\mathrm{k}}$
	NDSM	Normalized digital surface model	
	DoG	Difference of Gaussian	$\frac{1}{\sqrt{2\pi}}\left(\frac{1}{\sigma_1}\exp\left(-\frac{x^2+y^2}{2\sigma_1^2}\right) - \frac{1}{\sigma_2}\exp\left(-\frac{x^2+y^2}{2\sigma_2^2}\right)\right)$
	NDVI	Normalized difference vegetation index	$(\mathrm{IR}-\mathrm{R})/(\mathrm{IR}+\mathrm{R})$
Inter-modal features	Encoder features	/	
	Decoder features	/	

In the formula for the Slope angle, x, y, and z are the spatial positions and elevation of the LiDAR point, k denotes the neighborhood number of the current point, and x_i, y_i, and z_i are the spatial positions of the i-th neighborhood point. For the DoG formula, x and y denote the position of neighborhood points, and σ_1 and σ_2 are the standard deviations of the two Gaussian kernels, respectively. For the NDVI formula, IR refers to infrared and R denotes red bands.

while the parameters on the other layers are initialized by He *et al.* [46]. The input size of MFNet is 256 × 256. For the four side outputs of multiscale supervision, we downsample the ground truth map to four resolutions of 128 × 128–16 × 16. The input and ground truth map are then enhanced by applying random scaling and cropping, horizontal and vertical flipping, and a counterclockwise 90° rotation. Simultaneously, we calculated the mean and standard deviation of each dataset to normalize each input value. The total params of MFNet are 115.94 MB. Table II shows the detailed configurations of MFNet when using ResNet-50 as the encoder. The layer or module name is self-explanatory. For example, Conv1 denotes the first layer of convolution, and here, 7 × 7 is the size of the convolution filters, 64 is the number of output feature channels, the stride is equal to 2, and the output size is 128 × 128. $\mathrm{IMA}_{\mathrm{inter}}1$ denotes the first intermodal fusion ($\mathrm{IMF}_{\mathrm{inter}}$) module, whose input and output are both 128 × 128, and the corresponding size of the weight matrices is 1 × 1 × 64.

During training, we use the SGD algorithm for optimization. Its momentum parameter is 0.9, the weight decay is 0.0005, and the initial learning rate is 0.002, which is reduced by 0.8 times every 100 epochs. It took approximately 230 epochs for our network to converge. When the inputs of the two branches are (1, 3, 255, and 255) and (1, 4, 255, and 255), the inference time of MFNet is approximately 128 ms.

2) Evaluation Metrics: We evaluated the network using the default ISPRS 2-D test set and the corresponding benchmark metrics, the confusion matrix, the F1 score, and the overall accuracy (OA). In the confusion matrix [41], P and N are the number of positive and negative instances, respectively. The true positive (TP) value represents the corresponding diagonal element of the confusion matrix, TN is the true negative number, FP is the false positive number, and FN is the false negative number. The following formula can be used to calculate the OA and F1 indicators

$$A = \frac{\mathrm{TP} + \mathrm{TN}}{P + N} \tag{16}$$

$$\mathrm{F1} = 2 \times \frac{\mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{17}$$

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad \mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{18}$$

C. ISPRS 2-D Semantic Dataset Test Results

We selected a series of methods from the ISPRS 2-D semantic benchmark competitors and state-of-the-art deep semantic models from the CV field for comparative experiments to fully verify the segmentation accuracy of MFNet. They are SVL_3 [42], UZ_1 [30], ADL_3 [29], DST_2 [10], RIT_L7 [31], DLR_8 [47], ONE_7 [48], CASIA2 [49], SWL_2 [50], UFMG_4 [51], TreeUNet (BUCTY5) [43], FCN-8s [9], Segnet [12], Unet [52], PSPNet [53], RDFNet [54], RedNet [34], LANet [55], CCANet [56], CF-Net [57], HRNet [62], DeeplabV3 [58], and Segformer [59]. The details of the benchmark competition methods have been described on the challenge evaluation website or in the corresponding references. It is worth noting that the basic settings of all deep semantic models are the same and use the same dataset and data enhancement algorithm. These settings ensure that the experimental results are only affected by the network structure. We have summarized the above methods (including our methods) in Table III.

1) Comparison With Deep Models: We compared MFNet with six state-of-the-art semantic segmentation deep models from CV and 11 benchmark methods from ISPRS 2-D semantic benchmark competitors, as listed in Table III. The six deep models used ResNet-50 as their backbone and all experimental settings were consistent.

a) Potsdam validation set: As shown in Fig. 7, all the deep model methods have achieved fine segmentation results on the Potsdam 7–7 image in the first line, especially for buildings, trees, and roads. However, in terms of local details, the six deep models are not sufficiently robust and effective for confusing artifacts, shadow-blocked roads, and low vegetation. As shown in the last three line details pictures, neither Segnet nor PSPNet correctly segmented shadow-blocked roads between buildings. In addition to our MFNet, FCN-8s, Unet, RDFNet, and REDNet do not distinguish well between confusing artifacts and impervious surfaces. Although there

TABLE II
CONFIGURATIONS OF MFNET FOR THE ISPRS 2-D SEMANTIC LABELING CONTEST

Layer/Module	Output size	Encoder	Layer/Module	Output size	Decoder
IMA_{intra}	256×256	$\begin{bmatrix}\text{conv},1\times1,8\\ \text{conv},1\times1,16\\ \text{conv},1\times1,16\\ \text{fc},[1,16]\end{bmatrix}$	Conv2	8×8	$\text{conv},1\times1,512,\text{stride }1$
Conv1	128×128	$\text{conv},7\times7,64,\text{stride }2$	PD	8×8	$\begin{bmatrix}1\times(\text{conv},3\times3),512,[1]\\ 2\times(\text{conv},3\times3),512,[1,2]\\ 3\times(\text{conv},3\times3),512,[1,2,4]\\ 4\times(\text{conv},3\times3),512,[1,2,4,8]\\ 5\times(\text{conv},3\times3),512,[1,2,4,8,16]\end{bmatrix}$
$\text{IMF}_{inter}1$	128×128	$\begin{bmatrix}\text{conv},1\times1,64\\ \text{conv},1\times1,32\\ \text{conv},1\times1,64\\ \text{IMF}_{inter}[\text{r}=16,\text{l}=32]\end{bmatrix}$	Dereslayer1	16×16	$[\text{conv},3\times3,256,\text{stride }2]\times6$
Pooling1	64×64	$3\times3,\max pool,\text{stride }2$	MFF1	16×16	$\begin{bmatrix}\text{conv},1\times1,256,\text{stride }1\\ MFF,256\end{bmatrix}$
Reslayer1	64×64	$\begin{bmatrix}\text{conv},1\times1,64\\ \text{conv},3\times3,64\\ \text{conv},1\times1,256\end{bmatrix}\times3$	Dereslayer2	32×32	$[\text{conv},3\times3,128,\text{stride }2]\times4$
$\text{IMF}_{inter}2$	64×64	$\begin{bmatrix}\text{conv},1\times1,256\\ \text{conv},1\times1,32\\ \text{conv},1\times1,256\\ \text{IMF}_{inter}[\text{r}=16,\text{l}=32]\end{bmatrix}$	MFF2	32×32	$\begin{bmatrix}\text{conv},1\times1,128,\text{stride }1\\ MFF,128\end{bmatrix}$
Reslayer2	32×32	$\begin{bmatrix}\text{conv},1\times1,256\\ \text{conv},3\times3,256\\ \text{conv},1\times1,512\end{bmatrix}\times4$	Dereslayer3	64×64	$[\text{conv},3\times3,64,\text{stride }2]\times3$
$\text{IMF}_{inter}3$	32×32	$\begin{bmatrix}\text{conv},1\times1,512\\ \text{conv},1\times1,32\\ \text{conv},1\times1,512\\ \text{IMF}_{inter}[\text{r}=16,\text{l}=32]\end{bmatrix}$	MFF3	64×64	$\begin{bmatrix}\text{conv},1\times1,64,\text{stride }1\\ MFF,64\end{bmatrix}$
Reslayer3	16×16	$\begin{bmatrix}\text{conv},1\times1,512\\ \text{conv},3\times3,512\\ \text{conv},1\times1,1024\end{bmatrix}\times6$	Dereslayer4	128×128	$[\text{conv},3\times3,64,\text{stride }2]\times3$
$\text{IMA}_{inter}4$	16×16	$\begin{bmatrix}\text{conv},1\times1,1024\\ \text{conv},1\times1,64\\ \text{conv},1\times1,1024\\ \text{IMF}_{inter}[\text{r}=16,\text{l}=32]\end{bmatrix}$	MFF4	128×128	$\begin{bmatrix}\text{conv},1\times1,64,\text{stride }1\\ MFF,64\end{bmatrix}$
Reslayer4	8×8	$\begin{bmatrix}\text{conv},1\times1,1024\\ \text{conv},3\times3,1024\\ \text{conv},1\times1,2048\end{bmatrix}\times3$	Dereslayer5	128×128	$[\text{conv},1\times1,64,\text{stride }1]\times3$
$\text{IMF}_{inter}5$	8×8	$\begin{bmatrix}\text{conv},1\times1,2048\\ \text{conv},1\times1,128\\ \text{conv},1\times1,2048\\ \text{IMF}_{inter}[\text{r}=16,\text{l}=32]\end{bmatrix}$	FinalConv	256×256	$\text{conv},3\times3,6,\text{stride }2$

are some flaws in our labeling results, this can mark the boundary and annotations more precisely. Table IV shows a quantitative comparison between the seven deep models. Except for Segnet, all models achieve high precision. As for OA and the mean F1 score, the deep multimodal fusion strategy has greatly improved the performance of our MFNet, which exceeds all other models, especially for impervious surfaces, buildings, and low vegetation.

b) Vaihingen validation set: In Fig. 8, similar to the results of the Potsdam validation set, the six models, including Segnet, PSPNet, FCN-8s, Unet, RDFNet, and REDNet, are not sufficiently accurate to identify the artificial structure. They are also sensitive to the effects of occlusion and shadowing, such as artificial buildings and roads (impervious surfaces), as well as low vegetation and clutter, which can also affect the accuracy of segmentation. Our MFNet can handle these difficulties well, resulting in more complete boundaries and more accurate positioning, especially for fine structures, such as cars. As shown in Table V, compared with the other six state-of-the-art deep models, our MFNet is significantly better than the second place by 1.6% and 2.1% in OA and meanF1 scores, respectively, especially for impervious surfaces, low vegetation, and cars. This proves the effectiveness of our network structure and the deep multimodal fusion strategy.

2) Comparison With Published Methods on Benchmark Sets: We performed a comparative experiment with these published ISPRS 2-D benchmark leaderboard methods. All the methods used the same settings and data.

a) Potsdam benchmark test set: Fig. 9 shows the segmentation results of the eight ISPRS leaderboard methods, and Table VI shows their corresponding quantitative indicators. The OA and mean F1 score of our MFNet were 91.8 and 92.9, respectively, both of which exceeded all other methods, even though SWJ_2, CASIA2, and the other methods used a deeper Resnet-101 network, which is more difficult to train. For each class, our method achieved the highest F1 scores in the four classes of buildings, low vegetation, tree, and car. The F1 scores of the impervious surfaces class are second and only lower than SWJ_2. Specifically, as shown in Fig. 9, we show the comparison of eight methods in three complete images

TABLE III
DETAILS OF THE EXPERIMENTAL COMPARISON METHODS

Category	Name	Methods	Reference
Benchmark	SVL_3	SVL-features + DSM + Boosting + CRF	Gerke (2015)
	UZ_1	CNN + NDSM + Deconvolution	Volpi and Tuia (2017)
	ADL_3	CNN + DSM + NDSM + RF + CRF	Paisitkriangkrai et al. (2016)
	DST_2	FCN + DSM + RF + CRF	Sherrah (2016)
	RIT_L7	FCN-8s + Hand-crafted features + CRF	Liu et al. (2017)
	DLR_8	HED·H + FCN·H + SEG·H + CRF	Marmanis et al. (2017)
	ONE_7	SegNet + DSM + NDSM	Audebert et al. (2016)
	CASIA2	ResNet + pretrain + cascade	Liu et al. (2018).
	SWL_2	Resnet + multi-scale + shortcut block	Wang et al. (2018)
	UFMG_4	DilatedNet + Multicontext	Nogueira et al. (2019)
	TreeUNet (BUCTY5)	DeepUNet + Tree-CNN block	Yue et al. (2019)
	CCANet	Resnet +CCA-loss	Deng et al. (2021)
	CF-Net	Resnet +cross fusion + small scale	Peng et al. (2021)
	LANet	Resnet + Local Attention	Ding et al. (2021)
	HSV+Dgrad/SVM	Superpixels and classified with linear SVM	Campos-Taberner et al.(2015)
	RGBD$^+$ VGG/SVM	Linear SVM trained on VGG-16	Lagrange et al. (2015)
	RGBD$^+$ trained AlexNet	AlexNet on RGB and a more precise DSM	Campos-Taberner et al.(2015)
	CoFsn	RGB and DSM composite fusion	Piramanayagam et al. (2018)
	L3Fsn	RGB and DSM fusion after layer 3	Piramanayagam et al. (2018)
Deep model	FCN-8s	ResNet + NDSM + Single branch	Long et al. (2015)
	Segnet	ResNet + NDSM + Single branch	Badrinarayanan et al. (2015)
	Unet	ResNet + NDSM + Single branch	Ronneberger et al. (2015)
	PSPNet	ResNet + NDSM + Single branch	Zhao et al. (2017)
	RDFNet	ResNet + NDSM + two branches	Seong-Jin Park et al. (2017)
	RedNet	ResNet + NDSM + two branches	Jiang et al. (2018)
	DeepLabV3	ResNet + NDSM + Single branch	Lecun et al. (2017)
	HRNet	ResNet + NDSM + Single branch	Wang et al. (2021)
	Segformer	ResNet + NDSM + Single branch	Xie et al. (2021)
Ours	MFNet	ResNet + Hand-crafted features +two branches	Sun et al. (2021)

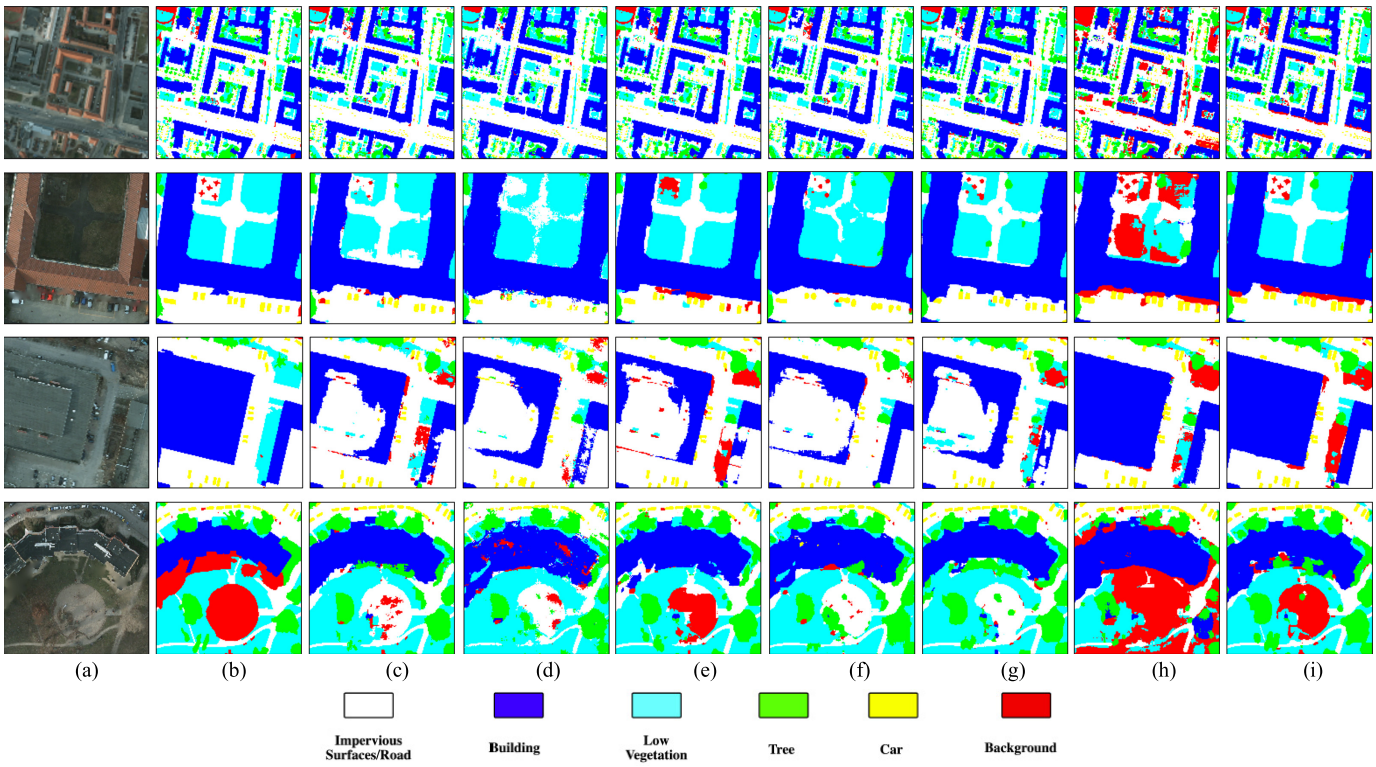


Fig. 7. Qualitative comparison with deep models on the ISPRS Potsdam Validation set. (a) Image. (b) Ground truth. (c) FCN-8s (d) Segnet. (e) Unet. (f) PSPNet. (g) RDFNet. (h) REDNet. (i) MFNet (ours).

and four partial patches. Almost all methods have achieved good segmentation results, and our methods have obtained more accurate boundaries for confusing road surfaces. Also, for buildings affected by shadows and small areas, such

TABLE IV

QUANTITATIVE COMPARISON WITH DEEP MODELS ON THE ISPRS POTSDAM VALIDATION SET

Name	F1 Score					OA	Mean F1
	Imp surf	Building	Low veg	Tree	Car		
FCN-8s (Resnet50)	91.1	96.0	80.9	83.5	94.5	89.3	89.20
Segnet (Resnet50)	84.8	92.4	78.4	78.7	83.6	83.3	83.58
Unet (Resnet50)	91.3	95.7	82.2	**85.1**	**94.8**	89.7	89.88
PSPNet (Resnet50)	89.8	94.3	81.1	81.8	91.4	87.2	87.68
RDFNet (Resnet50)	90.3	93.4	80.0	82.2	94.0	87.0	87.98
REDNet (Resnet50)	88.5	93.3	78.6	77.6	89.8	85.4	85.58
MFNet(Resnet50)	**91.9**	**96.8**	**82.5**	**85.1**	94.2	**90.2**	**90.08**

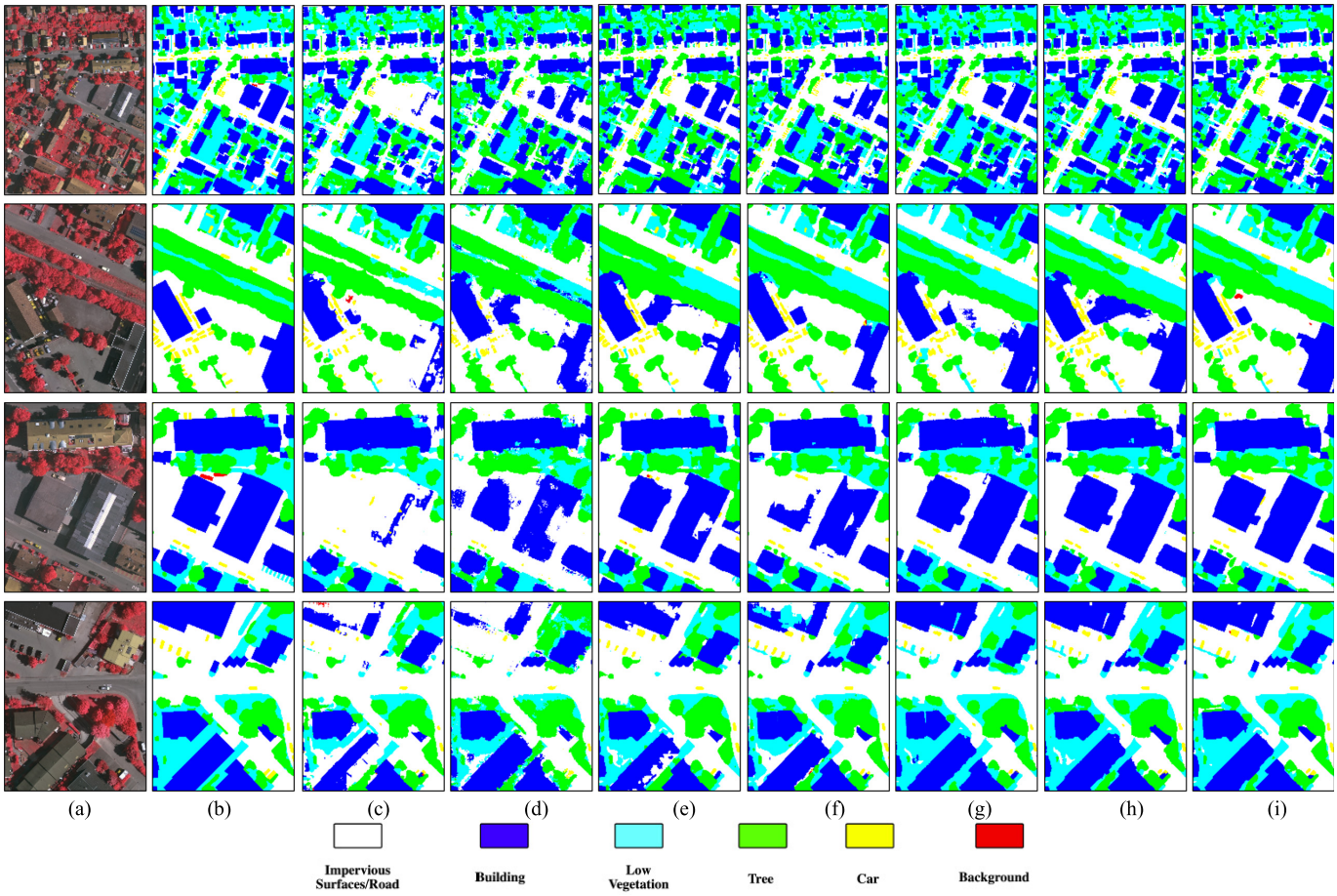


Fig. 8. Qualitative comparison with deep models on the ISPRS Vaihingen Validation set. (a) Image. (b) Ground truth. (c) FCN-8s. (d) Segnet. (e) Unet. (f) PSPNet. (g) RDFNet. (h) REDNet. (i) MFNet (ours).

TABLE V

QUANTITATIVE COMPARISON WITH DEEP MODELS ON THE ISPRS VAIHINGEN VALIDATION SET

Name	F1 Score					OA	Mean F1
	Imp surf	Building	Low veg	Tree	Car		
FCN-8s (Resnet50)	82.9	86.3	72.9	83.5	73.2	81.5	79.8
Segnet (Resnet50)	86.5	90.5	72.5	84.9	47.8	84.0	76.4
Unet (Resnet50)	90.9	94.8	78.6	87.2	85.4	88.3	87.4
PSPNet (Resnet50)	90.4	94.1	80.0	88.6	84.7	88.6	87.6
RDFNet (Resnet50)	90.8	95.2	79.1	86.5	78.4	87.9	86.0
REDNet (Resnet50)	90.8	95.65	79.6	87.0	85.5	88.4	87.7
DeeplabV3	90.1	91.1	77.9	85.5	71.8	86.3	\
HRNet	90.0	92.3	79.6	86.8	85.4	89.8	\
Segformer	90.2	91.4	79.1	86.2	85.5	89.4	\
MFNet(Resnet50)	**92.2**	**96.6**	**81.7**	**89.1**	**89.7**	**90.2**	**89.9**

as the comparison on the fourth, fifth, and sixth lines of Fig. 9, our approach shows more accurate boundaries and better robustness. It is worth noting that in the last line of Fig. 9, we have listed an ambiguous building category with dense low vegetation attached on its top. In the label dataset, it is classified as a building class. There are some similar phenomena in the entire dataset. In our algorithm, DST_2 and RIT_L7 classify its outline as a building and the inner part

TABLE VI
QUANTITATIVE COMPARISON WITH OTHER PUBLISHED METHODS ON THE ISPRS POTSDAM TEST SET

Name	F1 Score					OA	Mean F1
	Imp surf	Building	Low veg	Tree	Car		
SVL_3	84.0	89.8	72.0	59.0	69.8	77.2	74.9
UZ_1	89.3	95.4	81.8	80.5	86.5	85.8	86.7
RIT_L7	91.2	94.6	85.1	85.1	92.8	88.4	89.8
DST_2	91.8	95.9	86.3	87.7	89.2	89.7	90.2
CASIA2	93.3	97.0	87.7	88.4	96.2	91.1	92.5
TreeUNet	93.1	97.3	86.8	87.1	94.1	90.6	92.0
SWJ_2	**94.4**	97.4	87.8	87.6	94.7	91.7	92.4
MFNet	93.9	**97.6**	**88.2**	**89.2**	**95.5**	**91.8**	**92.9**

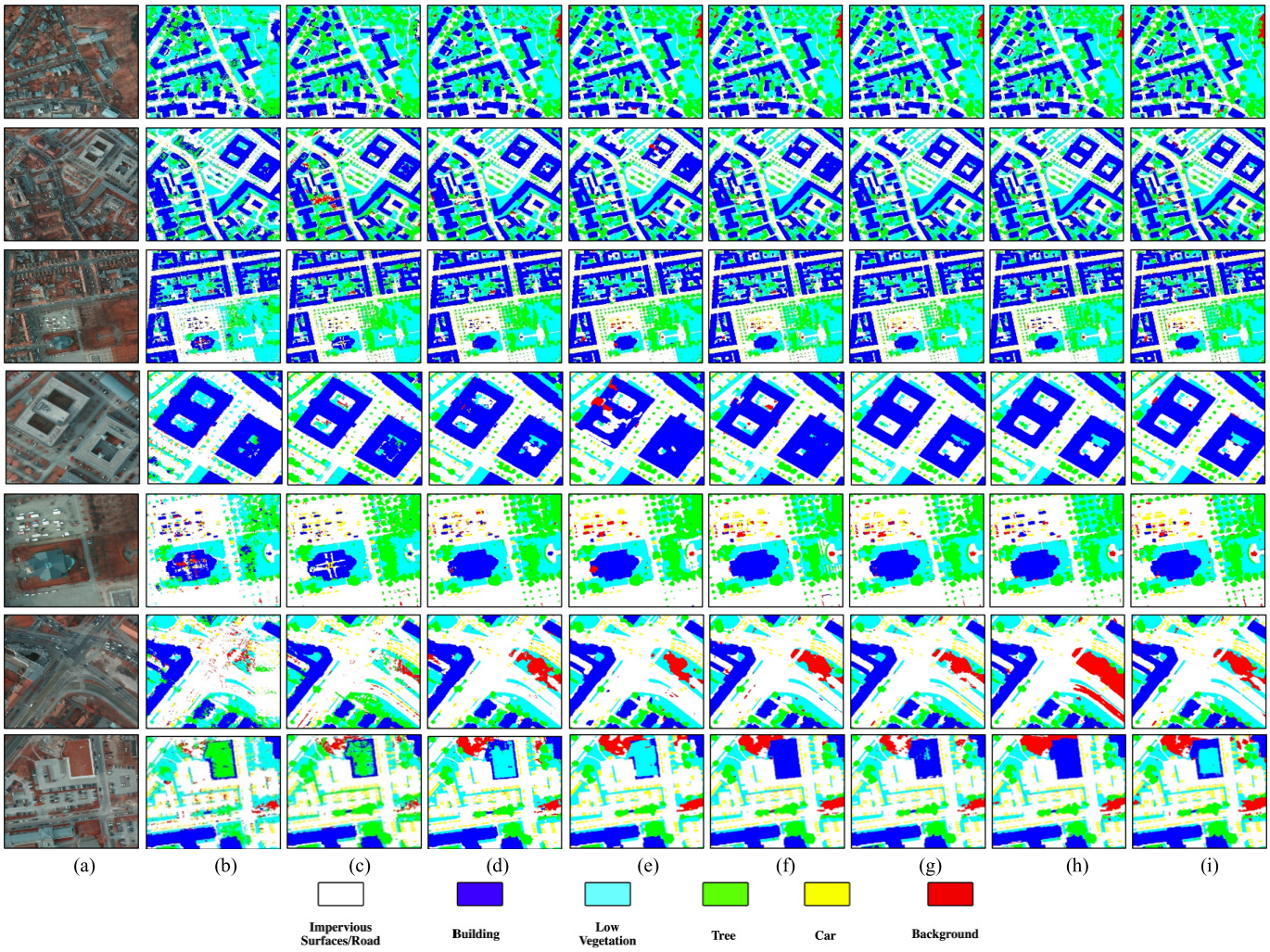


Fig. 9. Qualitative comparison with the officially published methods on the ISPRS Potsdam Test set. (a) Image. (b) SVL_3. (c) UZ_1. (d) RIT_L7. (e) DST_2. (f) CASIA2. (g) TreeUNet. (h) SWJ_2. (i) MFNet (ours).

as low vegetation. This will obviously reduce the OA of the algorithm, but it shows its better discriminative power and robustness.

b) Vaihingen benchmark test set: Table VII and Fig. 10 show a comparative experiment of the eight methods on the Vaihingen benchmark test set. Just like the previous analysis of the Potsdam test set, our method achieved the best OA and F1 score. Since our MFNet can aggregate the context information of multiscale and multimodality, it is very effective for labeling artificial and small objects. As shown in Fig. 10, MFNet has achieved more accurate boundaries for confusing buildings and fine structured cars. Table VII shows that the quantitative indicators of our methods are superior to other methods in all categories. Obviously, this is because we have made full use of the existing multimodal remote sensing data, and our network can make more efficient use of this multiscale and multimodal information.

D. GRSS DFC Zeebrugge Dataset Test Results

In order to verify the generalization ability of our proposed MFNet, we conducted further experiments in the GRSS DFC dataset. At the same time, we omitted the IMA_{intra} module,

TABLE VII
QUANTITATIVE COMPARISON WITH THE OFFICIALLY PUBLISHED METHODS ON THE ISPRS VAIHINGEN TEST SET

Name	F1 Score					OA	Mean F1
	Imp surf	Building	Low veg	Tree	Car		
SVL_3	86.6	91.0	77.0	85.0	55.6	84.8	79.0
ADL_3	89.5	93.2	82.3	88.2	63.3	88.0	83.3
DLR_8	90.4	93.6	83.9	89.7	76.9	89.2	86.9
UFMG_4	91.1	94.5	82.9	88.8	81.3	89.4	87.7
ONE_7	91.0	94.5	84.4	**89.9**	77.8	89.8	87.5
CASIA2	93.2	**96.0**	84.7	**89.9**	86.7	91.1	90.1
TreeUNet	92.5	94.9	83.6	89.6	85.9	90.4	89.3
CCANet	93.3	94.3	82.0	88.6	86.6	91.1	/
LANet	92.4	94.9	82.9	88.9	81.3	89.8	88.1
CF-Net	91.4	95.1	80.3	88.8	89.1	89.3	90.0
MFNet	**93.5**	**96.0**	**82.9**	88.9	**88.5**	**91.3**	**90.6**

so all comparison methods only used the RGB and DSM data to directly test the effectiveness of MFNet. It is worth mentioning that in our experiments on the ISPRS dataset, all methods used the same handcrafted features derived from RGB and LiDAR.

1) Comparison With Deep Models: We compared MFNet with four state-of-the-art semantic segmentation deep models from CV, as listed in Table III. They are Segnet [12], Unet [52], RDFNet [54], and RedNet [34].The four deep models used ResNet-50 as their backbone and all experimental settings were consistent.

Fig. 11 shows the segmentation results and the corresponding RGB and DSM images of the GRSS DFC Zeebrugge test data, while Table VIII summarizes the performance measures for each class and the overall obtained from the IEEE GRSS benchmark website. Compared with the other four deep models, the OA and Kappa values of our MFNet are significantly better than the second place (REDNet) by 7.24% and 9%, respectively, especially for buildings, boats, and cars. On the contrary, Segnet, Unet, and RDFNet failed to segment the car, and the RDFNet and REDNet are not sufficiently effective for buildings. This proves the effectiveness of our network structure and the deep multimodal fusion strategy. As shown in Fig. 11, neither Segnet nor Unet correctly segmented the tree and cars. In addition to our MFNet, Segnet, Unet, RDFNet, and REDNet do not distinguish well between buildings and impervious surfaces. Although REDNet has achieved a better accuracy in the impervious surfaces and clutter, the results of other classes, especially buildings, are not ideal. Both the quantitative and qualitative experiments show that our MFNet achieves the best segment results and the deep multimodal fusion strategy can greatly improve the performance of our method.

2) Comparison With Published Methods: We also performed a comparative experiment with seven published methods on the GRSS DFC Zeebrugge dataset. They are HSV+Dgrad/SVM [44], RGBD$^+$ VGG/SVM [60], RGBD$^+$ trained AlexNet [44], CCANet [56], DeeplabV3 [58], CoFsn [61], and L3Fsn [61]. All seven published methods are from the remote sensing community, as listed in Table III.

Table VIII also shows the quantitative results for the Zeebruges test images of the seven published methods. The OA and Kappa values of our MFNet exceeded all other methods. For each class, our method achieved the highest precision in the three classes of building, car, and water.

V. DISCUSSION

A. Model Effectiveness Analysis

1) Ablation Experiment: To evaluate the effectiveness of our model, we chose REDNet [34] as a baseline and made two ablation experiments by adding different modules progressively and separately. Then, we further analyzed their impact on the model and experimental accuracy.

The first ablation experiment was conducted on the ISPRS Potsdam Validation set. In Table IX, the Potsdam part lists the quantitative experimental results of the corresponding models. Compared with the baseline, when the intramodal fusion (IMF_{intra}) module is added, the experimental accuracy is increased by 0.43%. The IMF_{intra} module can flexibly process different modal features as input, such as DSM and NVDI. It efficiently and adaptively encodes multisource remote sensing data into multimodal features through the deep fusion mechanism, which can easily embed a segmentation network. For a further comparison, when the intermodal fusion (IMF_{inter}) module is continuously added in the encoder stage, the mean F1 score increases by 1% compared to the baseline. This is because the IMF_{inter} module can efficiently integrate intermodal features from two encode branches and learn to redistribute the weights for features of different modalities, enhance useful information, and suppress useless information. In addition, when the PD module is added into the final stage of the encoder, the cascaded dilation convolution of the pyramid structure efficiently captures information with a larger field of view and the performance is further improved, especially for cars and trees. Finally, after adding the MFF module, which is the proposed MFNet, the F1 score of each category is improved, and the mean F1 score is increased by 2.14% compared to the baseline. These well-designed improvements further prove the effectiveness of our MFNet through this ablation experiment.

The second ablation experiment was conducted on the GRSS DFC Zeebrugge dataset. Different from adding the modules progressively, the second ablation experiment compares the experimental effects of different modules separately. As with Section IV-D (experimental setup), we omitted the

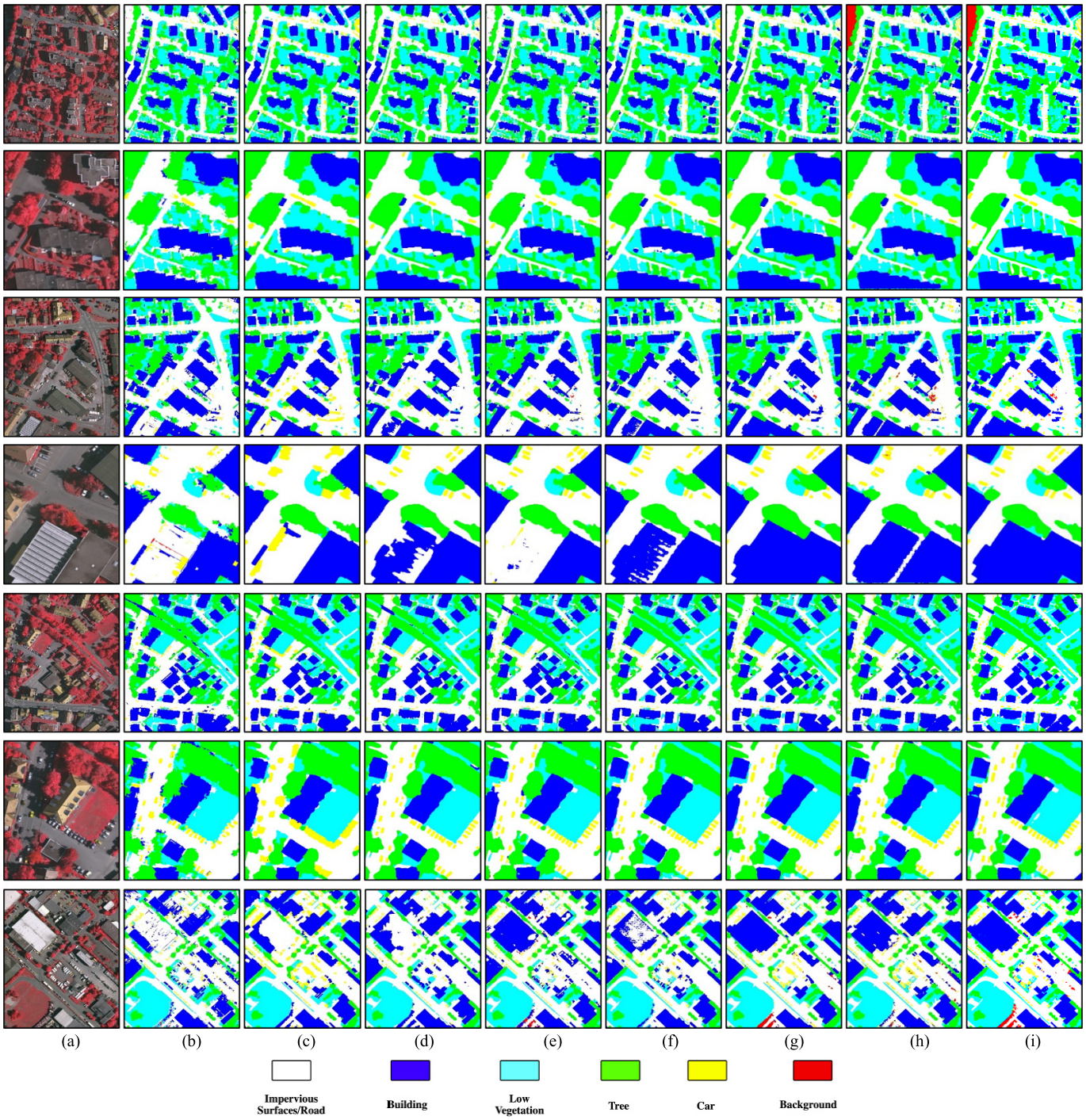


Fig. 10. Qualitative comparison with the officially published methods on the ISPRS Vaihingen Test set. (a) Image. (b) SVL_3. (c) UZ_1. (d) RIT_L7. (e) DST_2. (f) CASIA2. (g) TreeUNet. (h) SWJ_2. (i) MFNet (ours).

IMA_{intra} module. Through this ablation experiment, we can clearly analyze the role of the PD and IMA_{inter} modules. In Table IX, the Zeebrugge part also lists the quantitative experimental results of the corresponding models. Our PD module is similar to the FPA module proposed by Li *et al.* [39] in their pyramid attention network. Compared with the REDNet baseline, the PD module improves the network performance more than the FPA module. Except for the water, imp surf, and buildings classes, the segmentation accuracy of all classes exceeds the baseline. For the multimodal fusion mechanism, we only use the baseline with the intermodal fusion (IMF_{inter}) module for a comparison experiment. The segmentation accuracy of the baseline + IMF_{inter} has been significantly improved, and the OA and Kappa values are significantly better than the baseline by 5.03% and 7%, respectively. This second ablation experiment shows the effectiveness of the IMF_{inter} and PD modules, whose performance surpasses that on the ISPRS dataset. In this regard, we will

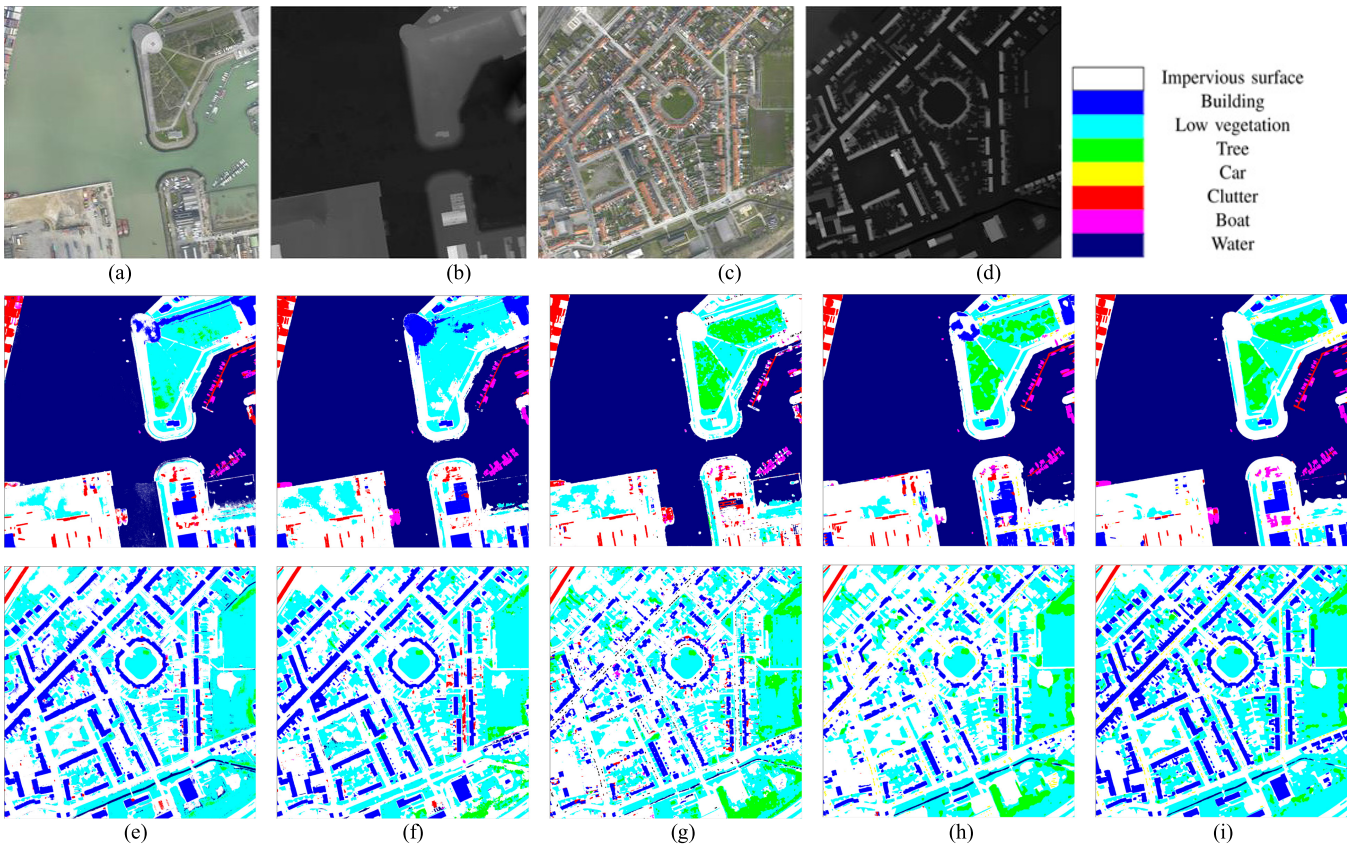


Fig. 11. Qualitative comparison with deep models on the GRSS DFC Zeebrugge dataset. (a) and (b) DSM 1. (c) and (d) DSM 2. (e) Segnet. (f) Unet. (g) RDFNet. (h) REDNet. (i) MFNet (ours).

TABLE VIII
QUANTITATIVE COMPARISON WITH PUBLISHED METHODS ON THE GRSS DFC ZEEBRUGGE DATASET

Name	Imp surf	Building	Low veg	Tree	Car	Clutter	Boat	Water	OA	Kappa
Segnet	89.62	62.61	76.81	12.05	0	58.56	28.43	99.01	80.82	0.74
Unet	87.01	62.58	75.18	11.88	0	55.05	51.83	98.66	79.72	0.73
RDFNet	89.99	27.06	76.45	76.45	0	56.11	31.13	98.33	78.54	0.71
REDNet	91.80	37.06	78.14	66.23	86.39	63.52	51.12	98.97	81.63	0.76
DeeplabV3	84.92	80.03	80.00	79.37	80.61	49.76	48.55	98.62	87.32	0.77
HSV+Dgrad/SVM	73.30	70.85	68.75	0.17	0	17.11	0	92.37	73.60	0.65
RGBD$^+$ VGG/SVM	67.66	72.70	68.38	78.77	33.92	45.60	56.10	96.50	76.56	0.70
RGBD$^+$ trained AlexNet	79.10	75.60	78.00	79.50	50.80	63.40	44.80	98.20	83.32	0.78
CoFsn	81.26	76.96	74.67	77.95	82.08	57.47	50.81	96.50	76.56	0.79
L3Fsn	84.8	83.93	84.24	80.17	83.13	62.83	55.77	98.97	87.91	0.84
CCANet	87.11	82.95	85.68	80.36	88.05	61.45	70.25	98.75	88.83	0.85
MFNet	86.95	**84.55**	81.77	77.52	**88.65**	61.72	69.29	**99.03**	**88.87**	**0.85**

continue to study the performance of these modules on other remote sensing datasets.

2) Computational Metrics: We tested the GPU memory, the number of model parameters, giga floating-point operations per second (GFLOPS), and forward inference time of these deep learning models and the variants of MFNet on a single Nvidia GTX1080ti GPU. It can be seen that the parameters and GPU memory of RDFNet are both the highest, and our MFNet has the longest inference time in Table X. Our model has undergone a comprehensive transformation based on the baseline model. Although it achieved the highest segmentation accuracy, the GPU memory and model parameters also increased by 41.8% and 41.5%, respectively. Therefore, in the future, we will attempt to study a light version of MFNet.

B. Effects of the Deep Multimodal Fusion Strategies

To further demonstrate the effectiveness of our fusion strategy, we visualized three sets of multimodal feature fusion processes. In particular, the first 16 feature maps from layer 2 of our MFNet are selected to explain how these complementary multimodal features are fused together. As shown in Fig. 12, the Modality1 features are from IRRG, and the Modality2 features are LiDAR-derived features (slope and NDSM) and imagery-derived features (NDVI and DoG). Each

TABLE IX

ABLATION EXPERIMENT (%) ON THE GRSS DFC ZEEBRUGGE DATASET AND ISPRS POTSDAM VALIDATION SET. FPA: FEATURE PYRAMID ATTENTION MODULE. IMF_{intra}: INTRAMODAL FUSION MODULE. IMF_{inter}: INTERMODAL FUSION MODULE. PD: PYRAMID DILATION MODULE. MFF: MULTILEVEL FEATURE FUSION MODULE

Name (Zeebrugge)	Clutter	Boat	Water	Imp surf	Building	Low veg	Tree	Car	OA	Kappa
Baseline	63.52	51.12	98.97	91.80	37.06	78.14	66.23	86.39	81.63	0.76
Baseline + FPA	64.78	64.58	99.11	92.47	9.62	81.92	84.03	79.52	79.86	0.73
Baseline + PD	63.72	65.79	98.31	90.23	30.31	86.12	80.76	90.21	82.50	0.77
Baseline + IMF_{inter}	63.53	52.08	99.16	85.42	83.80	81.04	64.94	90.62	86.66	0.83
Name (Potsdam)				Imp surf	Building	Low veg	Tree	Car	OA	Mean F1
Baseline				90.83	95.65	79.64	87.04	85.54	88.38	87.74
+ IMF_{intra}				91.57	95.83	80.10	87.39	88.29	88.81	88.64
+ IMF_{intra}+ IMF_{inter}				91.84	96.06	80.36	87.30	88.13	89.37	88.74
+ IMF_{intra}+ IMF_{inter}+ PD				92.22	96.59	81.12	88.52	89.89	89.84	89.67
+ IMF_{intra}+ IMF_{inter}+ PD + MFF(our-MFNet)				92.24	96.63	81.71	89.07	89.69	90.20	89.88

TABLE X

COMPARISONS OF THE NETWORK EFFICIENCY AND THE VARIANTS OF MFNET. THE TEST DATA WERE 256 × 256 IMAGE PATCHES

Name	GPU Memory (MB)	The number of Parameters (Mega)	GFLOPS (Giga)	Forward Time (ms)
Segnet	112.34	29.45	40.4	19
Unet	563.86	147.81	53.85	47
RDFNet	**1403.4**	**367.89**	**145.52**	105
Baseline	312.61	81.94	21.16	75
Baseline + PD	348.6	91.38	21.77	88
Baseline + IMA_{inter}	406.3	106.51	26.38	108
MFNet	443.31	115.94	27.05	**128**

group of feature maps with their corresponding weights is represented by a 4 × 4 matrix, and then, the fusion features are obtained by pixelwise multiplication and addition. Specifically, it can be seen that Fig. 12(a) is a building covered in shadows and the features of Modality1 and Modality2 cannot completely express the entire building area. However, in Fig. 12(a), the two modalities show outstanding complementarity, such as the feature maps in the third line marked by black boxes. The features of different modalities are well integrated through the weights learned by the deep fusion network. The multimodal fusion feature can fully express this building area and has an excellent discrimination with surrounding low vegetation. Fig. 12(b) is a mixed area of buildings, roads, and cars. Just like Fig. 12(a), through the fusion process, more accurate and discriminative features are obtained. At the same shown by the black box feature map in the first and third lines, the characteristics of small-scale objects have also been well expressed, such as cars and sunroofs. Fig. 12(c) shows a mixed area of vegetation, roads, and cars. As shown in these black boxes of third line, before the fusion, the features of Modality1 could not distinguish between road and vegetation and the features of Modality2 could not correctly represent vegetation, but accurately characterized the road area. After learning the adaptive weight parameters through the deep fusion network, the multimodal fusion features can accurately and completely express the road, cars, and vegetation areas.

C. Effects of the Multiscale Supervision Training Scheme

Many studies have tested the benefits of multiscale inputs on network performance. In contrast, we use multiscale outputs strategies with supervised training. As shown in Fig. 2, our MFNet has four side outputs: Out4, Out3, Out2, Out1, and a final output, while the height and width of Out4, Out3, Out2, and Out1 are 1/2, 1/4, 1/8, and 1/16 of the final output, respectively. All these five outputs are used for loss calculation. To test the multiscale supervision scheme, we designed a comparison experiment of one-scale supervision (Final Out), three-scale supervision (Out4, Out2, and Final Out), and five-scale supervision (MFNet) on the ISPRS Potsdam dataset. As shown in Table XI (Potsdam part), the one-scale supervised training scheme achieved a fine accuracy on the Vaihingen validation set. At the same time, the three-scale supervised training scheme further improves the segmentation results. In the end, the five-scale supervised training scheme adopted by our MFNet achieved an optimal segmentation accuracy in all categories, except for the road. This shows that our multiscale supervision scheme is sufficient to effectively improve the network performance.

For the problem of computational complexity caused by a multiscale supervision scheme, we test the GPU memory, training time, and GFLOPs of the model under the above three scales on the GRSS DFC Zeebrugge dataset. As shown in Table XI (Zeebrugge part), the GFLOPs of the three multiscale supervision methods are almost the same and equal to 27. The GPU memory of the three strategies is also very close. The training times of each Epoch (10 000 samples) are 573, 585, and 650 s. Under the three multiscale supervision methods, the number of epochs required for network convergence is 100, 150, and 230. It can be seen that multiscale supervision scheme hardly increases the computational complexity of the model. However, it reduces the convergence speed of the network and takes longer to train the model. Compared with the one-scale supervision method, the five-scale supervision method requires 2.3 times the number of epochs to achieve network convergence. Although it requires more training time, the five-scale supervision method can achieve a higher segmentation accuracy and at the same time avoid the network from prematurely falling into a local optimal situation. For example, in the GRSS DFC Zeebrugge dataset segmentation

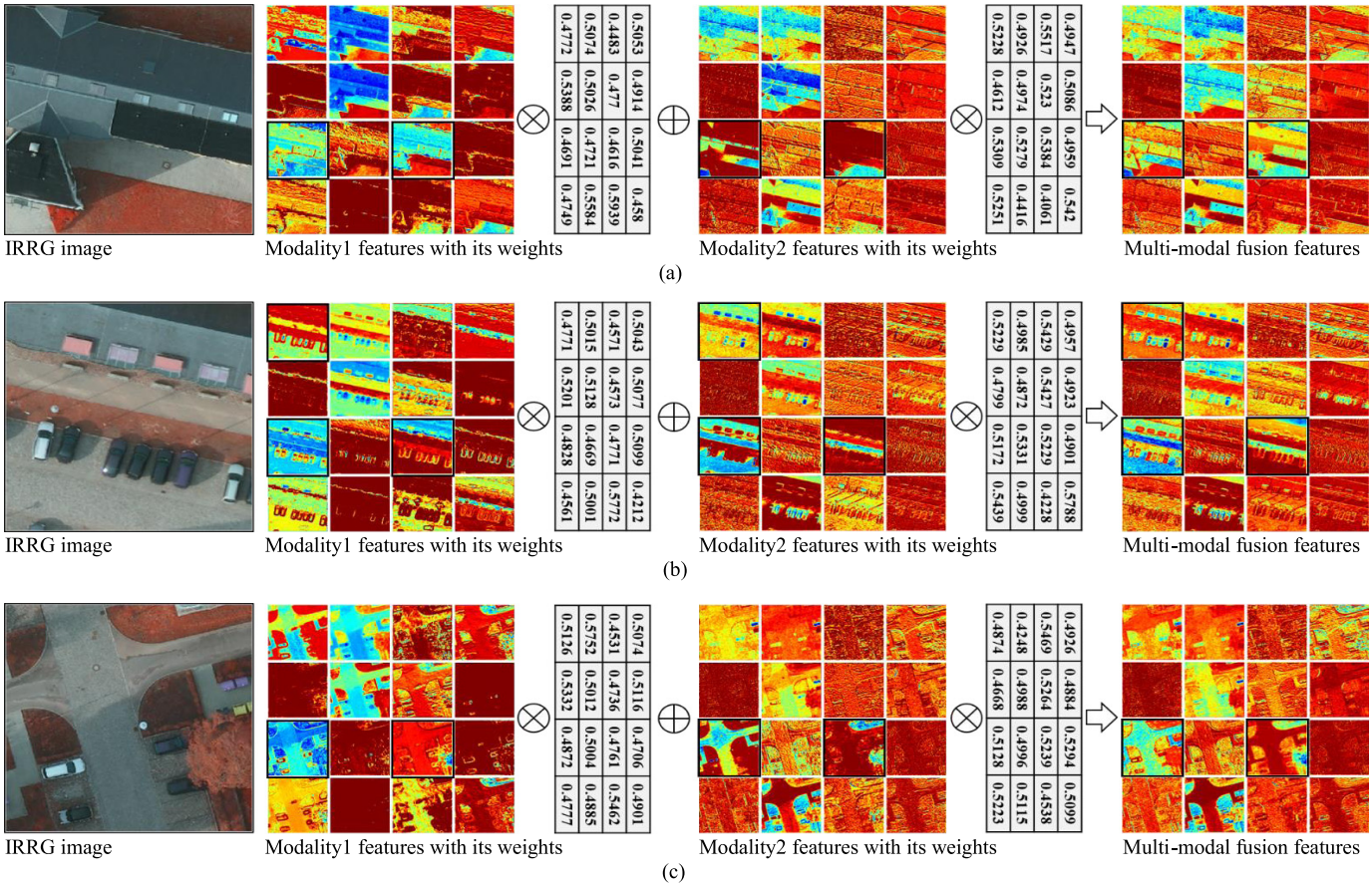


Fig. 12. Features fusion using the multimodal fusion strategies. ⊗ denotes an elementwise product and ⊕ denotes an elementwise addition. The sum of the weights from Modality1 and Modality2 is one. (a) Multimodal fusion on a patch of the Potsdam dataset. (b) Multimodal fusion on a patch of the Potsdam dataset. (c) Multimodal fusion on a patch of the Potsdam dataset.

TABLE XI

QUANTITATIVE COMPARISON (%) BETWEEN ONE-SCALE SUPERVISION (FINAL OUT), THREE-SCALE SUPERVISION (OUT4, OUT2, AND FINAL OUT), AND FIVE-SCALE SUPERVISION (MFNET) ON THE GRSS DFC ZEEBRUGGE DATASET AND ISPRS VAIHINGEN VALIDATION SET

Methods (Potsdam)	Imp surf	Building	Low veg	Tree	Car	OA	Mean F1
1-scale supervision (Final Out)	91.81	96.18	81.67	88.66	88.57	89.56	89.38
3-scale supervision (Out4, Out2 and Final Out)	92.66	96.57	81.66	88.49	89.27	90.07	89.73
5-scale supervision (MFNet)	92.24	96.63	81.71	89.07	89.69	**90.20**	**89.88**
Methods (Zeebrugge)	GPU Memory (MB)	One Epoch Time (s)		GFLOPS (G)		OA	Kappa
1-scale supervision (Final Out)	442.29	573		36.98		84.48	0.78
3-scale supervision (Out4, Out2 and Final Out)	443.31	585		27.05		86.42	0.80
5-scale supervision (MFNet)	443.31	650		27.05		**88.87**	**0.85**

results shown in Table XI, the accuracy of five-scale supervision is 4.39% higher than that of the one-scale supervision.

VI. CONCLUSION

A deep MFNet was proposed in this study, which can use multimodal VHR aerial images and LiDAR data and the corresponding intramodal features, such as LiDAR-derived features (slope and NDSM) and imagery-derived features (IRRG, NDVI, and DoG). The proposed MFNet achieves excellent segmentation performance through three key aspects.

1) Two multimodal fusion modules are proposed: $\text{IMF}_{\text{intra}}$ and $\text{IMF}_{\text{inter}}$. They can learn complementary features and cross-modal interdependencies by interleaving intramode and intermode features. To the best of our knowledge, this is the first VHR segmentation framework that uses a soft attention mechanism for intramodality and intermodality feature fusion.
2) A PD module is proposed, which can obtain more global and multiscale coding features without losing the original features. It greatly improves the labeling results of confusing artificial objects.
3) An MFF module is proposed that to fuse high- and low-level features. By using the deep fusion mechanism to model the relationship between these multiple levels features, we can obtain fine resolution and highly distinguishable feature maps. Quantitative and qualitative

comparative experiments and ablation studies illustrated that MFNet has a competitive performance in comparison with state-of-the-art VHR segmentation approaches.

ACKNOWLEDGMENT

The authors would like to thank the ISPRS and IEEE Geoscience and Remote Sensing Society for providing the data and owe great appreciation to the editors and anonymous reviewers for their valuable comments.

REFERENCES

[1] L. Matikainen and K. Karila, "Segment-based land cover mapping of a suburban area—Comparison of high-resolution remotely sensed datasets using classification trees and test field points," *Remote Sens.*, vol. 3, no. 8, pp. 1777–1804, Aug. 2011, doi: 10.3390/rs3081777.

[2] W. Ye and Y. Wang, "MST image segmentation based on Mumford-Shah theory," *J. Comput.-Aided Des. Comput. Graph.*, vol. 21, no. 8, pp. 1127–1133, 2009.

[3] Z. Fu, Y. Sun, L. Fan, and Y. Han, "Multiscale and multifeature segmentation of high-spatial resolution remote sensing images using superpixels with mutual optimal strategy," *Remote Sens.*, vol. 10, no. 8, p. 1289, Aug. 2018.

[4] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Advances in spectral-spatial classification of hyperspectral images," *Proc. IEEE*, vol. 101, no. 3, pp. 652–675, Mar. 2013.

[5] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "*TextonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 3951, May 2006, pp. 1–15.

[6] P. Tokarczyk, J. D. Wegner, S. Walk, and K. Schindler, "Features, color spaces, and boosting: New insights on semantic classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 280–295, Jun. 2014.

[7] D. Tuia, N. Courty, and R. Flamary, "Multiclass feature learning for hyperspectral image classification: Sparse and hierarchical solutions," *ISPRS J. Photogramm. Remote Sens.*, vol. 105, pp. 272–285, Jul. 2015.

[8] D. Tuia, M. Volpi, M. D. Mura, A. Rakotomamonjy, and R. Flamary, "Automatic feature learning for spatio-spectral image classification with sparse SVM," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6062–6074, Oct. 2014.

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

[10] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," 2016, *arXiv:1606.02585*. [Online]. Available: http://arxiv.org/abs/1606.02585

[11] Y. L. Cun, B. Boser, J. S. Denker, D. Henderson, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 1990, vol. 2, no. 2, pp. 396–404.

[12] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[13] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

[14] H. Guo, G. Wang, and X. Chen, "Two-stream convolutional neural network for accurate RGB-D fingertip detection using depth and edge information," 2016, *arXiv:1612.07978*. [Online]. Available: http://arxiv.org/abs/1612.07978

[15] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, vol. 10111, Nov. 2016, pp. 213–228.

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

[17] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6526–6534.

[18] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 3, 2014, pp. 1–9.

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: http://arxiv.org/abs/1409.0473

[20] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[21] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[22] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[23] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[24] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.

[25] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.

[26] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. Neural Inf. Process. Syst. (NIPS)*, vol. 1, 2012, p. 2.

[27] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 681–687.

[28] J. Wang, Z. Wei, T. Zhang, and W. Zeng, "Deeply-fused nets," 2016, *arXiv:1605.07716*. [Online]. Available: http://arxiv.org/abs/1605.07716

[29] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. Van-Den Hengel, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 36–43.

[30] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 881–893, Feb. 2017.

[31] Y. Liu, S. Piramanayagam, S. T. Monteiro, and E. Saber, "Dense semantic labeling of very-high-resolution aerial imagery and LiDAR with fully-convolutional neural networks and higher-order CRFs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1561–1570.

[32] Y. Sun, X. Zhang, Q. Xin, and J. Huang, "Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and LiDAR data," *ISPRS J. Photogramm. Remote Sens.*, vol. 143, pp. 3–14, Sep. 2018.

[33] Y. Xu *et al.*, "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1709–1724, Jun. 2019.

[34] J. Jiang, L. Zheng, F. Luo, and Z. Zhang, "RedNet: Residual encoder-decoder network for indoor RGB-D semantic segmentation," 2018, *arXiv:1806.01054*. [Online]. Available: http://arxiv.org/abs/1806.01054

[35] K. He, X. Zhang, S. Ren, and S. Jian, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[36] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 182–186.

[37] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3085–3094.

[38] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: http://arxiv.org/abs/1511.07122

[39] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," 2018, *arXiv:1805.10180*. [Online]. Available: http://arxiv.org/abs/1805.10180

[40] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2650–2658.

[41] M. Gerke, F. Rottensteiner, J. D. Wegner, and G. Sohn, "ISPRS semantic labeling contest," in *Proc. ISPRS*, Leopoldshöhe, Germany, 2014. [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/semanticlabeling.html

[42] M. Gerke, "Use of the stair vision library within the ISPRS 2D semantic labeling benchmark (Vaihingen)," Univ. Twente, Enschede, The Netherlands, Tech. Rep., 2015.
[43] K. Yue, L. Yang, R. Li, W. Hu, F. Zhang, and W. Li, "TreeUNet: Adaptive tree convolutional neural networks for subdecimeter aerial image segmentation," *ISPRS J. Photogramm. Remote Sens.*, vol. 156, pp. 1–13, Oct. 2019, doi: 10.1016/j.isprsjprs.2019.07.007.
[44] M. Campos-Taberner *et al.*, "Processing of extremely high-resolution LiDAR and RGB data: Outcome of the 2015 IEEE GRSS data fusion contest–Part A: 2-D contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 12, pp. 5547–5559, Dec. 2016.
[45] W. Davidson and M. Abramowitz, "Molecular expressions microscopy primer: Digital image processing-difference of Gaussians edge enhancement algorithm," Olympus Amer. Florida State Univ., Tallahassee, FL, USA, Tech. Rep., 2006.
[46] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[47] 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 J. Photogramm. Remote Sens.*, vol. 135, pp. 158–172, Jan. 2018.
[48] N. Audebert, B. L. Saux, and S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2016, pp. 180–196.
[49] Y. Liu, B. Fan, L. Wang, J. Bai, S. Xiang, and C. Pan, "Semantic labeling in very high resolution images via a self-cascaded convolutional neural network," *ISPRS J. Photogram. Remote Sens.*, vol. 145, pp. 78–95, Nov. 2018.
[50] (2018). *Method Description for Potsdam 2D Labelling Challenge*. [Online].Availabel:http://ftp.ipi.unihannover.de/ISPRS_WGIII_website/ISPRSIII_4_Test_results/papers/Report_SWJTU2.pdf
[51] K. Nogueira, M. D. Mura, J. Chanussot, W. R. Schwartz, and J. A. dos Santos, "Dynamic multicontext segmentation of remote sensing images based on convolutional networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7503–7520, Oct. 2019.
[52] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.
[53] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.
[54] S. Lee, S.-J. Park, and K.-S. Hong, "RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4980–4989.
[55] L. Ding, H. Tang, and L. Bruzzone, "LANet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 426–435, Jan. 2021.
[56] G. Deng, Z. Wu, C. Wang, M. Xu, and Y. Zhong, "CCANet: Class-constraint coarse-to-fine attentional deep network for subdecimeter aerial image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, early access, Feb. 12, 2021, doi: 10.1109/TGRS.2021.3055950.
[57] C. Peng, K. Zhang, Y. Ma, and J. Ma, "Cross fusion net: A fast semantic segmentation network for small-scale semantic information capturing in aerial scenes," *IEEE Trans. Geosci. Remote Sens.*, early access, Jan. 29, 2021, doi: 10.1109/TGRS.2021.3053062.
[58] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587
[59] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," 2021, *arXiv:2105.15203*. [Online]. Available: http://arxiv.org/abs/2105.15203
[60] A. Lagrange *et al.*, "Benchmarking classification of earth-observation data: From learning explicit features to convolutional networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 4173–4176.
[61] S. Piramanayagam, E. Saber, W. Schwartzkopf, and F. Koehler, "Supervised classification of multisensor remotely sensed images using a deep learning framework," *Remote Sens.*, vol. 10, no. 9, p. 1429, Sep. 2018.
[62] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

Yangjie Sun received the M.S. degree in cartography and geography information system from Central South University, Changsha, China, in 2016, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2021.

His research interests include remote sensing image segmentation, deep learning, 3-D scene segmentation, and 3-D geographic information system (GIS).

Zhongliang Fu received the B.S., M.S., and Ph.D. degrees in photogrammetry and remote sensing from Wuhan Technical University of Survey and Mapping, Wuhan, China, in 1985, 1988, and 1996, respectively.

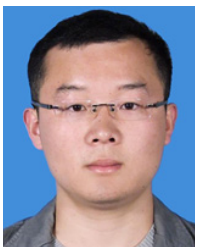
He is currently a Professor with the School of Remote Sensing and Information Engineering, Wuhan University. He is also the Director of the Geographic Information System Department. His research interests include spatial data management and update, remote sensing image processing and analysis, map scanning image recognition, vehicle license plate recognition, and geographic information engineering technology.

Chuanxia Sun received the M.S. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2007.

He is currently a Professorate Senior Engineer with Henan Provincial Department of Transportation Highway Pipeline Bureau, Zhengzhou, China. His research interests include remote sensing and transportation.

Yinglei Hu received the B.S. degree in Software Engineering from Henan University, Zhengzhou, China, in 2007.

He is a senior engineer with the Highway Administration Bureau, Henan Transportation Department, Zhengzhou, China. His research interest includes remote sensing information technology in transportation engineering.



Shengyuan Zhang received the M.S. degree in computer technology from Hebei University of Technology, Tianjin, China, in 2014. He is currently pursuing the Ph.D. degree in remote sensing science and technology with the School of Remote Sensing and Information Engineering, Wuhan University.

His research interests include deep learning and remote sensing information technology in transportation engineering.