HR and LiDAR Data Collaborative Semantic Segmentation Based on Adaptive Cross-Modal Fusion Network

Zhen Ye, Zhen Li, Nan Wang, Yuan Li, and Wei Li, *Senior Member, IEEE*

***Abstract*—Semantic segmentation using cross-modal data is a hot topic in the field of Earth observation. Compared with single-modal strategies, cross-modal networks fuse multiaspect information and yield higher segmentation accuracy, which is widely used in urban planning, environmental monitoring and so on. In this study, an end-to-end adaptive cross-modal fusion network (ACFNet) is proposed for semantic segmentation task using high resolution and light detection and ranging images, because of the difference of sensor resolution, different modal data have different abilities of ground object expression. Therefore, multimodal data fusion should consider the features with different spatial scales, while most existing methods simply use the same spatial scale features for fusion. In this work, we first design an adaptive scale fusion module that can automatically choose the features with optimal spatial scales, making full use of the representation properties of ground object details. Second, the important feature guidance module is designed, which can evaluate the influence weights of deep semantic features and shallow spatial detailed features, achieving adaptive deep and shallow feature fusion, and reducing the semantic-spatial information dilution caused by layer-by-layer up and down sampling. Finally, we introduce a divide Fourier context learning (DFCL) module to transform the feature maps from spatial domain to frequency domain. Compared to the limited perception of current spatial convolution kernels, the DFCL module can easily model the contextual dependencies of cross-modal features, which will improve the segmentaion accuracy for complex ground objects of cities, especially for occlusion. To demonstrate the generalisation performance of our module, we conduct extensive experiments and ablation studies on three datasets: Potsdam, Vaihingen, and IEEE GRSS DFC 2018. Results show that the proposed ACFNet is effective in semantic segmentation.**

***Index Terms*—Adaptive learning, aerial imagery, cross-modal fusion, semantic segmentation.**

I. Introduction

HIGH resolution (HR) images have meter-level or even submeter-level spatial resolution [1], [2], which can clearly express the spatial structure and surface texture characteristics of ground targets [3], [4], [5], [6]. Semantic segmentation in remote sensing is to interpret HR images into pixel-level semantic annotations [7]. Currently, semantic segmentation tasks involve in precision agriculture, urban planning and natural resource inventory [8], [9], [10], [11], [12], [13]. Although significant progress has been made in semantic segmentation for HR images, their accuracy will be constrained by some factors, such as shadows, noise, obstacles, geometric distortion and height displacement of tall buildings. Light detection and ranging (LiDAR) images can complement the elevation information of objects [14], so the joint application of HR and LiDAR is considered to obtain more refined segmentation results [15], [16], [17].

The development of sensor technology has led to the emergence of cross-modal data that can express the ground surface in various aspects. This has significantly improved the interpretation accuracy of Earth observation and has become a research hotspot, especially in the deep learning community. For example, Yuan et al. [16] proposed a fully convolutional network (FCN) based on residual structure to jointly extract target features from HR and LiDAR images to improve the detection accuracy of buildings. Piramanayagam et al. [18] introduced a deep convolutional neural network (DCNN) based on FCN-32 for merging multi-sensor features for semantic segmentation, which can obtain better segmentation performance than traditional CNN-based network. Hazirbas et al. [19] proposed FuseNet algorithm, which uses SegNet architecture proposed by Badrinarayanan et al. [20] to segment cross-modal data by incorporating a cross-fusion algorithm into the encoder part. Zhang et al. [21] proposed a feature-level fusion network based on hybrid attention-aware fusion network (HAFNet), in which an attention-aware fusion block is designed to better fuse cross-modal information.

The results of these previous studies show that the combination of cross-modal feature extraction with DCNN can improve the accuracy of semantic segmentation [22], [23], [24], but there are still some room for improvement. The following issues can be discussed.

1) Due to the differences in sensor resolution, the abilities of geophysical information expression of different modal data are different. When fusing cross-modal data, it is important to consider features with different spatial scales, instead of considering single spatial scale as most DCNNs

Manuscript received 31 January 2024; revised 5 May 2024; accepted 18 June 2024. Date of publication 24 June 2024; date of current version 12 July 2024. This work was supported by the National Key Research and Development Program of China under Grant 2020YFC1512002. *(Corresponding author: Nan Wang.)*

Zhen Ye, Zhen Li, and Yuan Li are with the School of Electronics and Control Engineering, Chang'an University, Xi'an 710064, China (e-mail: yezhen525@126.com; 2021232078@chd.edu.cn; ly612592@163.com).

Nan Wang and Wei Li are with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: wang_nan@bit.edu.cn; liwei089@ieee.org).

Data is available online at https://github.com/Zhenzhen98/LI-ZHEN.

Digital Object Identifier 10.1109/JSTARS.2024.3418387

do. Can we design an architecture to dynamically select the appropriate spatial scale features of HR and LiDAR for fusion?

2) Many DCNNs used for semantic segmentation are based on encoder–decoder architecture. In the baseline, down-sampling layer by layer will lead to the continuous attenuation of spatial details, and up-sampling layer by layer will lead to the continuous dilution of semantic information. Can we design a network to capture spatial-semantic information at each layer of the network?
3) Current DCNNs use spatial convolution kernel strategy in decode stage, which only focuses on local pixels and has a limited perception. Therefore, the global dependencies of the multiscale features are overlooked. Can we construct a model to obtain global context information and then enhance the representation of feature mapping?

In response to the above questions, we try to give the following answers. Since encoding operation yields multiscale features on cross-modal convolutional layers, the joint application of these features should be divided into two cases. On the one hand, when the features of different modals at the same scale (that is, the cross-modal features from the same coding layer) are fused, it is considered to fully extract the spatial details of the cross-modal data. On the other hand, in order to effectively retain important semantic-spatial information dilution caused by layer-by-layer up and down sampling, multiscale feature reuse (MFR) and information selection can be used to fine-grained semantic-spatial information dilution caused by layer-by-layer up and down sampling with different scales (that is, cross-modal features from different coding layers). In addition, traditional CNNs adopt spatial convolution kernels to decode the final result layer-by-layer. However, the local perception of that could not exactly decode the context information, which influences the accuracy of results under complex ground objects of cities, especially for occlusion. This problem can be solved by decoding global information of multiscale features layer by layer. Thus, we study a common network for semantic segmentation using HR and LiDAR images. The primary contributions of this article are as follows.

1) We propose an adaptive cross-modal fusion network (ACFNet). In order to fully extract the complementary and interactive information of HR and LiDAR images, a cascade network is designed with multiple skip connections in encoder–decoder network.
2) Adaptive scale fusion (ASF) modules, each of which is implemented by an adaptive scale block and a residual block, are constructed to fuse the cross-modal features from the corresponding encoders. Adaptive scale block adaptively learns cross-modal features from HR and LiDAR data with different spatial scales, and residual block retains more spatial details in the form of upsampling and residual convolution.
3) Important feature guided (IFG) module contains a MFR block and an information selection block. IFG can evaluate the influence weights of deep semantic features and shallow spatial detailed features, achieving adaptive deep and shallow feature fusion, and reducing the semantic-spatial information dilution caused by layer-by-layer up and down sampling.
4) The divide Fourier context learning (DFCL) module uses two odd–even 2-D fast Fourier transforms (FFT) to transfer the feature mapping from the spatial domain to the frequency domain. In addition, for the frequency domain information, we adjust the phase and amplitude to learn the context features layer by layer. Then, the global dependencies from semantic to spatial features can be decoded exactly and accurately.

The rest of this article is organized as follows. Section II describes the proposed ACFNet and its constituent components in detail. An overview of the experimental data and setup is presented in Section III. In Section IV, we show the experimental results, ablation studies, as well as a detailed analysis of ACFNet. Finally, Section V concludes this article.

II. Proposed Method

In this section, we describe the structure of the proposed ACFNet, an overview of which is shown in Fig. 1. First, HR-derived red–green–blue (RGB) image and LiDAR-derived digital surface model (DSM) image are input into the dual-branch network, capitalizing on the ResNet-34 network [25] as the backbone of encoder, in which the ASF module is designed to fuse different spatial scale features of different modal data on the premise of retaining enough spatial details. Then, the important feature guided (IFG) module is designed to learn the multiscale fine-grained features and evaluate the influence weights of deep semantic features and shallow spatial detailed features, achieving adaptive deep and shallow feature fusion, and reducing the semantic-spatial information dilution caused by layer-by-layer up and down sampling. Finally, the DFCL module is used to further decode the global features through the frequency domain layer-by-layer and generate the final prediction map.

A. Network Architecture

As mentioned above, coding operations are performed on a HR branch and a LiDAR branch with ResNet-34 as the backbone network, shown as Fig. 1(a). The encoding process for both the RGB-branch and the DSM-branch follows the same setup. The LiDAR-branch sets single-channel input (for a single image in DSM) and the HR-branch sets three-channel input (for three images in RGB). The specific structure and layer sizes of the encoder are shown in Fig. 1(b) and Table I, respectively. The first encoder block involves a 7×7 convolutional layer with a stride of 2, followed by a batch normalization (BN) layer and a rectified linear unit (ReLU) layer. The second encoder block includes a 3×3 maximum pooling layer with a stride of 2 and the initial residual layer of ResNet-34. Subsequently, the remaining three residual layers are added successively. With the exception of the first residual layer, all three residual layers incorporate a residual unit that downsamples the feature map and doubles the feature channels. Thus, the resolution of the final output produced by the encoding is 1/32 of that of the original input data. For the decoder, as shown in Fig. 1(a), three $2 \times$ and a $4 \times$ upsampling operations are performed, in conjunction with

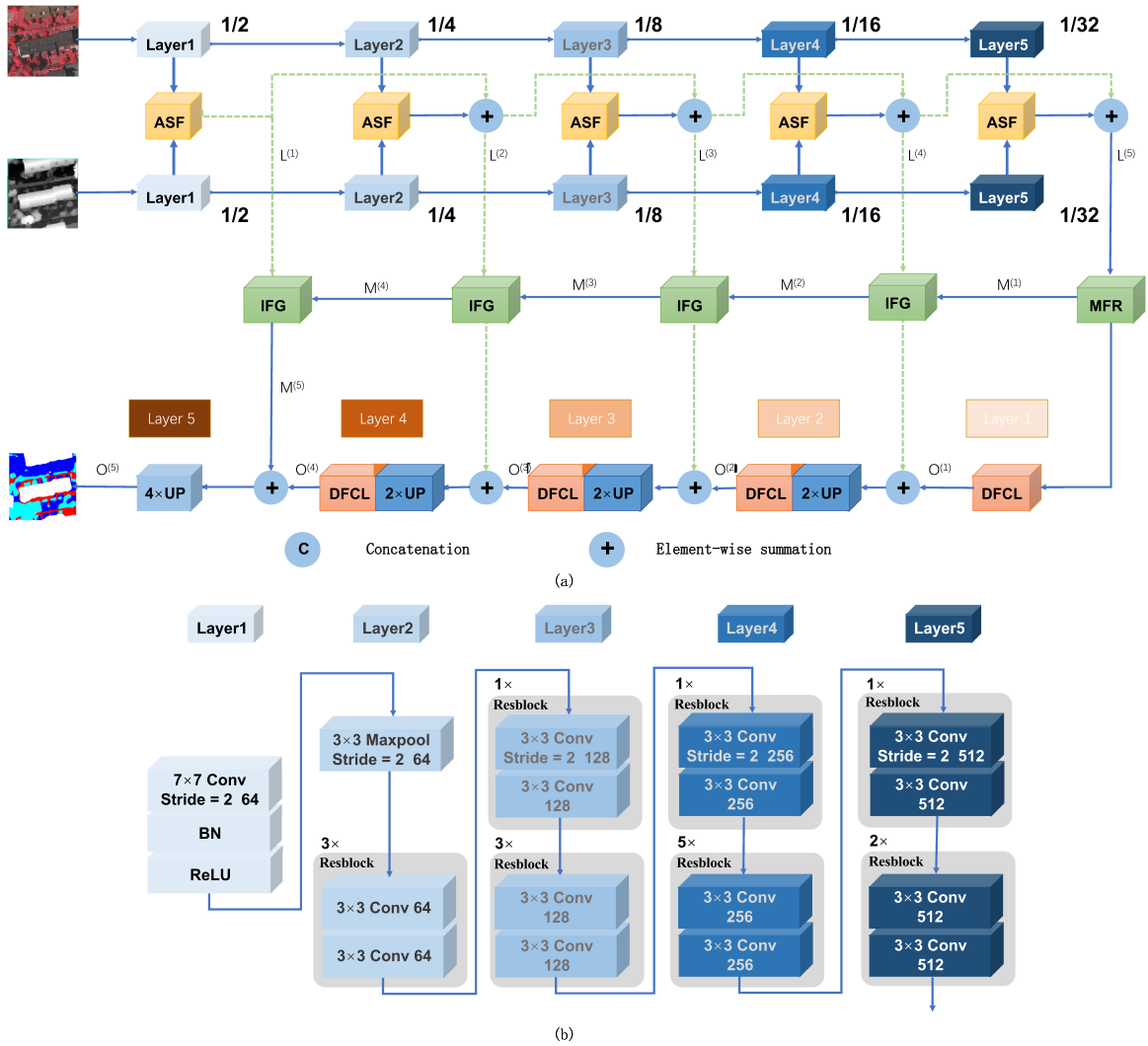


Fig. 1. (a) Overall framework of the proposed ACFNet. (b) Specific structure of encoder.

the FCL modules to ensure that the size of final output matches that of the original input. In addition, to the backbone encoder–decoder structure described above, this section also covers the three designed modules.

Because of the difference of sensor resolution, different modal data have different abilities of ground object expression. Therefore, multimodal data fusion should consider the features with different spatial scales, while most existing methods simply use the same spatial scale features for fusion. To solve this problem, we propose an ASF module adaptively learns cross-modal features from HR data and LiDAR data in different spatial scales. Let $L^{(m)}$ denote the output of the mth layer of ASF, which can be expressed as

$$L^{(m)} = \begin{cases} \text{ASF}\left(E_{\text{rgb}}^{(m)}, E_{\text{dsm}}^{(m)}\right), & m = 1 \\ \text{ASF}\left(E_{\text{rgb}}^{(m)}, E_{\text{dsm}}^{(m)}\right) + L^{(m-1)}, & m = 2, \ldots\ldots, 5 \end{cases} \tag{1}$$

where ASF denotes the implementation of ASF; $E_{\text{rgb}}^{(m)}$ denotes the output feature of the mth encoder in the RGB branch; and $E_{\text{dsm}}^{(m)}$ denotes the output feature of the mth encoder in the DSM branch.

In the task of semantic segmentation, most CNN-based networks are based on encoder–decoder architecture. Due to the up-sampling and down-sampling process in this architecture, the features extracted in deep layers contain high-level semantic information, but the spatial details are insufficient. On the contrary, features of shallow layers have sufficient spatial details, but contain low-level semantic information. In order to ensure that each layer retains important spatial and semantic information and weakens unimportant information, this article proposes an important feature guided (IFG) module to obtain important information from cross-modal fusing features. For the output of fifth encoder, only MFR is conducted instead of IFG module. The output information of the first decoder can be expressed as

$$M^{(1)} = \text{MFR}(L^{(5)}) \tag{2}$$

where $L^{(5)}$ is the output of the fifth layer of encoder. Then, for layers 2–5 of the encoder, each layer will contain two inputs (the

TABLE I
LAYER SIZES OF THE ENCODER

Layername	Network configuration
Layer1	7×7 Conv, 64, stride=2 BN ReLU
Layer2	3×3 Max pool, stride=2 $\begin{bmatrix} 3\times3 & \text{Conv}, 64 \\ 3\times3 & \text{Conv}, 64 \end{bmatrix} \times 3$
Layer3	$\begin{bmatrix} 3\times3 & \text{Conv}, 128, stride=2 \\ 3\times3 & \text{Conv}, 128 \end{bmatrix}$ $\begin{bmatrix} 3\times3 & \text{Conv}, 128, stride=2 \\ 3\times3 & \text{Conv}, 128 \end{bmatrix} \times 3$
Layer4	$\begin{bmatrix} 3\times3 & \text{Conv}, 256, stride=2 \\ 3\times3 & \text{Conv}, 256 \end{bmatrix}$ $\begin{bmatrix} 3\times3 & \text{Conv}, 256, stride=2 \\ 3\times3 & \text{Conv}, 256 \end{bmatrix} \times 5$
Layer5	$\begin{bmatrix} 3\times3 & \text{Conv}, 512, stride=2 \\ 3\times3 & \text{Conv}, 512 \end{bmatrix}$ $\begin{bmatrix} 3\times3 & \text{Conv}, 512, stride=2 \\ 3\times3 & \text{Conv}, 512 \end{bmatrix} \times 2$

outputs of ASF of this layer and the last output information of IFG / MFR). The output of MFR/IFG module is

$$M^{(n)} = \text{IFG}(L^{(m)}, M^{(n-1)}), \qquad n = 2, \ldots\ldots, 5 \tag{3}$$

$$n = 6 - m \tag{4}$$

where n denotes the order of decoder layers; and $L^{(m)}$ represents the output of the mth encoder layer of cross-madal fusion.

Since global context information is crucial for semantic segmentation, we employ DFCL module to learn the global information of cross-modal data and to model the context dependencies in different layers of the decoder stage. DFCL module adds the features of $F^{(n)}$ to each layer of the decoder, enabling the decoder network to obtain a more fine-grained feature mapping. Combined with upsampling, DFCL module transforms the extracted multiscale features in frequency domain to extract deep semantic information while retaining shallow spatial context information, thus improving the global awareness and decoding accuracy of the decoder. The FCL operation in the first layer of decoder can be expressed as

$$O^{(1)} = \text{DFCL}(M^{(1)}) \tag{5}$$

where $O^{(1)}$ denotes the output of layer 1 of the decoder. On the layers 2–4 of the decoder, the input of the FCL modules is the output of decoder from the previous layer and $M^{(n)}$, so as to realize the fusion of semantic information and spatial details. The output of layers 2–4 of decoder can be expressed as

$$O^{(n)} = \text{DFCL}(\text{UP}(M^{(n)} + O^{(n-1)})), \qquad n = 2, \ldots\ldots, 4 \tag{6}$$

where $O^{(n)}$ denotes the output of the nth layer of the decoder; UP denotes $2\times$ upsampling; and FCL represents the implementation of Fourier context learning. The last layer of decoder, which only uses $4\times$ upsampling to ensure the size of final output matching that of the original input without FCL module for feature learning, can be expressed as

$$O^{(5)} = \text{UP}(O^{(4)} + M^{(5)}). \tag{7}$$

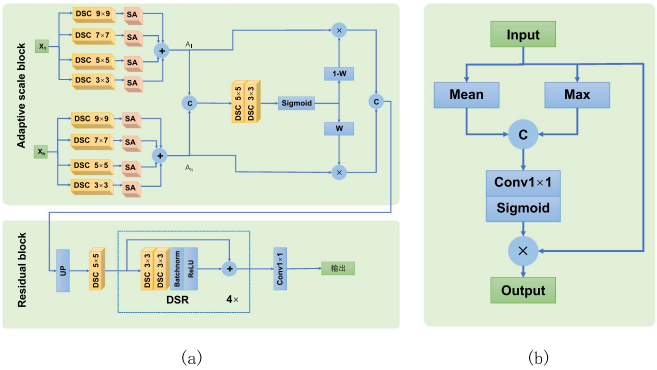
Algorithm 1: Pseudocode for our Method Training.

Input: HR data X_h, LiDAR data X_l, Ground truth Y, number of training epochs.
Output: Classification map O.
Parameter setting and weights initialization.
for $epoch < epochs$ **do**
Extract cross-modal fusion features $L^{(m)}$ of HR data and LiDAR data according to Eq. (1), where m represents layer order of encoder.
Concatenate $L^{(m)}$ and the output of IFG to obtain $M^{(n)}$ by Eq. (3), where n represents the layer order of decoder. Next, feed $M^{(n)}$ into DFCL and upsampling moduls. Then, classification map will be obtained by Eqs. (5)–(7).
Update parameters and weights via Adamax optimizer.
end for
Obtain the probability distribution and classification map.

B. Cross-Modal Information Fusion

Due to the differences in sensor resolution, the abilities of geophysical information expression of different modal data are different. When fusing cross-modal data, it is important to consider features with different spatial scales, instead of considering single spatial scale as most DCNNs do. This module can be divided into two stages: adaptive scale block and residual block, as shown in Fig. 2.

In adaptive scale block, X_h and X_l are considered to represent the outputs of the feature maps of the RGB and DSM encoders, respectively, with consistent number of feature channels for dual branches. First, the output features of LiDAR branch and HR branch are conducted by parallel 3×3, 5×5, 7×7, and 9×9 depth separable convolution (DSC) followed by spatial attention (SA). Next, add operation is used for multiscale fusion, and concatenation operation is used for cross-modal fusion. Then, a 5×5 DSC and a 3×3 DSC are employed to calculate the correlation of the features. Subsequently, the sigmoid activation is employed to obtain the weighted probability matrix W, and the LiDAR-branch adaptive features A_l and HR-branch adaptive features A_h are multiplied with $(1-W)$ and (W), respectively.



(a) (b)

Fig. 2. (a) Architecture of the proposed ASF module including an adaptive scale block and a residual block. (b) SA.

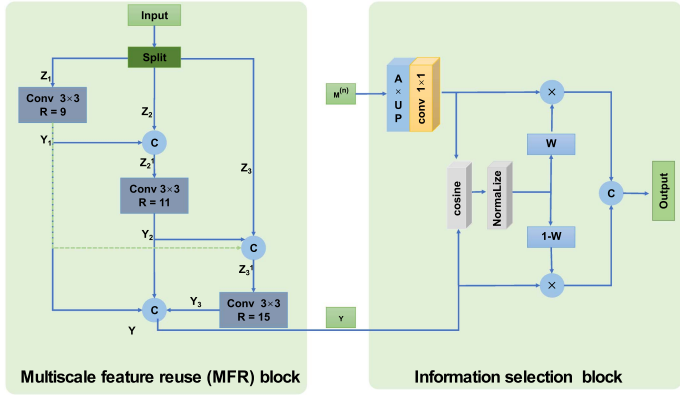


Fig. 3. Architecture of the proposed IFG module including a MFR block and an information selection block.

Finally, these features are concatenated. The process can be expressed as

$$A_l = \text{ADD}(\text{SA}(D_3(X_l)), \text{SA}(D_5(X_l)), \text{SA}(D_7(X_l)), \text{SA}(D_9(X_l))) \quad (8)$$

$$A_h = \text{ADD}(\text{SA}(D_3(X_h)), \text{SA}(D_5(X_h)), \text{SA}(D_7(X_h)), \text{SA}(D_9(X_h))) \quad (9)$$

$$\text{SA} = S(f_1(C[\text{Max}(X), \text{Mean}(X)])) \times X \quad (10)$$

$$F = C[A_l, A_h] \quad (11)$$

$$W = \text{Sigmoid}(D_3(D_5(F))) \quad (12)$$

$$X_f = C[(W) \odot A_h, (1 - W) \odot A_l] \quad (13)$$

where D_3, D_5, D_7, and D_9 denote 3×3, 5×5, 7×7, 9×9 DSC operation, respectively; W denotes the obtained weight matrix; C denotes concatenation; and $\odot$ denotes the Hadamard product.

In residual block, first, the fused features are upsampled and the number of channels is increased using a 5×5 DSC in order to preliminarily extract cross-modal features; second, four identical depth-wise separable residual (DSR) modules are used to fully extract the cross-modal information; finally, a 1×1 standard convolution is used to ensure the consistency of the output and input feature sizes. The process can be expressed as

$$X_{\text{fusion}} = f_1(D_5(\text{UP}(X_f)) + 4 \times \text{DSR}(D_5(\text{UP}(X_f)))) \quad (14)$$

$$\text{DSR} = \text{ReLU}(\text{BN}(D_3(D_3))) \quad (15)$$

where f_1 denotes 1×1 standard convolution; UP denotes $2\times$ upsampling; and BN denotes BN.

The proposed ASF module is important from two points of view. On the one hand, parameters and calculation cost of a DSC perform ($1/n + 1/k^2$) times lesser than a standard convolution [26], where k denotes the kernel size and n denotes the number of the output channels for DSC. On the other hand, an adaptive scale block can adaptively choose the features with optimal spatial scales, making full use of the representation properties of ground object details.

C. Multiscale Feature Extraction

As described in Section II-A, in order to fully extract the complementary and interactive information of HR and LiDAR images to achieve high-precision segmentation and keep important features during encoding and decoding operations, important feature guidance (IFG) module is divided into two blocks: MFR block and information selection block.

Inspired by Res2Net [27] and EPSANet [28], MFR enlarges the receptive fields in a multi-scale way by using feature reuse to extract fine-grained multi-scale feature maps. Each MFR layer splits the input feature map into three partitions along the channel dimension. As an example, Fig. 3 illustrates the framework of MFR. That is, in Fig. 3, the input feature map Z is splited by function $S(\cdot)$; i.e.,

$$\{Z_1, Z_2, Z_3\} = S(Z) \quad (16)$$

such that the a channels of Z are partitioned into Z_1, Z_2, and Z_3, with $a/3$ channels each, where a should be divisible by 3. Then, Z_1 is conducted by a 3×3 dilated convolution, yielding the spatial feature map

$$Y_1 = H(Z_1) \quad (17)$$

where $H(\cdot)$ represents the dilated convolution.

Next, we concatenate Y_1 and Z_2 generating cascade features map Z_2^1, which effectively enables feature reuse. Z_2^1 is then performed the dilated convolution obtaining the multi-scale spatial features Y_2. That is

$$Z_2^1 = C[Y_1, Z_2] \quad (18)$$

$$Y_2 = H(Z_2^1) \quad (19)$$

where $C[\cdot]$ is channel-wise concatenation.

We then concatenate Y_1, Y_2, and Z_3, generating cascade features Z_3^1. Z_3^1 is conducted by a dilated convolution again to obtain Y_3, which can be described as

$$Z_3^1 = C[Y_1, Y_2, Z_3] \quad (20)$$

$$Y_3 = H(Z_3^1). \quad (21)$$

Finally, these feature maps from different channels are concatenated by

$$Y = C[Y_1, Y_2, Y_3]. \tag{22}$$

In this process, we employ feature reuse to increase the information interaction between different partitions. That is, consider the output Y_1 of the first partition, on the one hand, the output Y_1 of the first partition is cascaded to the input Z_2 of the second partition, while, on the other hand, Y_1 is also concatenated to the output of the other partitions to obtain the final output Y; thus, the network uses feature Y_1 multiple times. The MFR employs similar feature reuse for Y_2 as well.

Furthermore, the MFR enlarges the receptive field in a multiscale dilated convolution manner. More specifically, traditional multiscale feature extraction methods are often designed as networks of multiple parallel branches, each with a CNN of a fixed kernel size, such that by using different kernel sizes in the different branches, extraction of features with multiple scales is accomplished. In contrast, in the MFR as proposed here, each partition uses the same kernel size of dilated convolution. The multiscale nature of MFR arises instead from the fact that the dilated convolution is used, which expands the effective receptive field. Take the first partition as an example: kernel size is 3×3 and dilation rate *(R)* is 9, while the effective receptive field of dilated convolution is 19×19 (each output value addresses 361 values in the original input feature map) when stride is set to 1. The receptive field of dilated convolution is larger than that of standard convolution even though the kernel itself is the same size, this conclusion also applies to the second and the third partitions. Thus, the receptive field is enlarged without the cost of actually using the dilated convolution with a larger kernel.

Inspired by GRRNet [29], this work also constructs an information selection block, which preserves the important information between upper and lower layers and removes the unimportant information by calculating the cosine similarity of two neighboring layers.

For information selection block, the output of MFR and the output of IFG in the previous stage are employed to calculate the cosine similarity. Thus, the information with high similarity will be kept, and the information with low similarity will be removed. As shown in Fig. 3, a A-upsampling and a 1×1 convolution are performed to match the channel numbers between the upper and lower stages. Since MRF features (Y) and IFG features ($\mathrm{M}^{(n)}$) have different importance, we design a gating mechanism to allocate adaptive weight (W). The mathematical expression of this block is as follows:

$$S^{'(m)} = f_1(\mathrm{UP}(M^{(n)})) \tag{23}$$

$$\mathrm{similarity} = \mathrm{cosine}(S^{'(m)}, Y, \mathrm{dim} = 1) \tag{24}$$

$$\mathrm{weight} = (\mathrm{similarity} + 1)/2 \tag{25}$$

$$S^{(m+1)} = C[S^{'(m)} \times (\mathrm{weight}), Y \times (1 - \mathrm{weight}))] \tag{26}$$

where $\mathrm{M}^{(n)}$ represents the output of MFR / IFG of the previous layer; f_1 represents 1×1 convolution; Y represents the output of MFR of this layer; and cosine represents cosine similarity calculation.

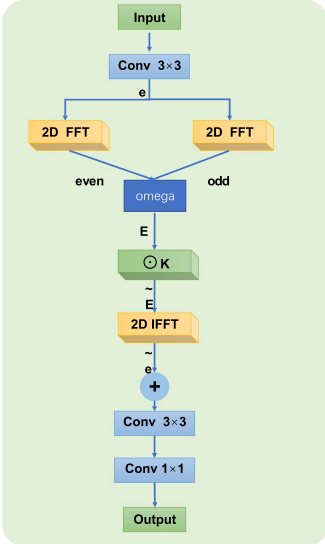


Fig. 4. DFCL module.

D. Global Contextual Learning

Learning global contextual information of HR and LiDAR images is beneficial to improve the performance of joint HR and LiDAR segmentation networks. Inspired by GFNet [30], the use of frequency-domain information can extract global features. Based on this, we propose the DFCL module, which uses parametric learnable filters for FFT to capture global context information in frequency domains to improve the network's performance.

The DFCL module consists of two 2-D FFT, a Hadamard product and a 2-D inverse FFT (IFFT), shown as Fig. 4. Given a feature mapping from the backbone, and feed it into the convolution layer to generate $e \in \mathcal{R}^{H\times W\times C}$, where $\mathcal{R}$, H, W, and C represent real number set, image height, image width, and image channel number, respectively. Then, we perform 2-D FFT in odd–even form for guided features and use phase modulation omega to combine the modulated odd–even frequency information. Assume transforming factor is e, the designed 2-D FFT strategy can be described as

$$\mathrm{even} = \mathrm{FFT}(e[0 :: 2]) \tag{27}$$

$$\mathrm{odd} = \mathrm{FFT}(e[1 :: 2]) \tag{28}$$

$$t = \mathrm{omega} * *i * \mathrm{odd}[i] \tag{29}$$

$$E[i] = \mathrm{even}[i] + t \tag{30}$$

$$E[i + n//2] = \mathrm{even}[i] - t \tag{31}$$

where E contains all frequency components of the feature map e. Note E is a complex tensor and represents the frequency characteristics of e, which is then multiplying parameter learnable filter $K \in \mathcal{C}^{H\times W\times C}$, where $\mathcal{C}$ represents the category to which the image belongs, which is able to cover all the frequencies to

E and we get result

$$\widetilde{E} = E \odot K \tag{32}$$

where $\odot$ denotes the Hadamard product. Then, we adopt the IFFT to transform the modulated $\widetilde{E}$ back to the spatial domain by

$$\widetilde{e} \leftarrow F^{-1} \in \mathcal{R}^{H \times W \times C}. \tag{33}$$

Element by element, * represents multiplication, and e and $\widetilde{e}$ are summed for fusing spatial and frequency information followed by a 3×3 convolution. Finally, a 1×1 convolution is used to adjust the dimensionality of the channels. From the above operations, FCL uses global context information to augment the input feature mapping.

III. Experimental Data and Setup

A. Datasets

1) Potsdam[1]: The Potsdam dataset consists of 38 sets of images. Each set contains a RGB image and a corresponding DSM image. The spatial size of Potsdam dataset is 6000×6000, while the spatial resolution is 0.05 m. There are six main semantic classes, namely, *impervious surfaces, buildings, low vegetation, trees, car, and clutter*, which are described in details shown as Fig. 10.

2) Vaihingen[2]: The Vaihingen dataset consists of 33 sets of images. Each set also contains a RGB image and a corresponding DSM image. The spatial size of the Vaihingen dataset is 2500×2000 with 0.09 m spatial resolution, and it has the same identified classes as that of the Potsdam dataset. The RGB, DSM, and ground-truth map are shown in Fig. 11.

3) DFC2018[3]: The IEEE GRSS DFC2018 dataset has 14 RGB images, and a united DSM image, from which we obtain 14 cropped DSM images corresponding to HR scenes. For this dataset, the spatial size is 1192×1202 and the spatial resolution is 0.05 m. There are 21 classes in original dataset, namely, *unclassified buildings, residential buildings, nonresidential buildings, stadium seating, roads, pavements, pedestrian crossings, main arterials, highways, railways, trains, water, evergreens, deciduous trees, healthy grass, stressed grass, artificial turf, paved car parks, unscored car parks, cars, and bare soil*. The images used in this article covers 9 classes, shown as Fig. 12.

B. Experimental Setup

1) Implementation Details: All the experiments were conducted on a single NVIDIA GeForce RTX 3090 GPU and PyTorch framework. Due to the GPU memory limitation, we crop these images using a 640×640 sliding window with no overlap. For the Potsdam dataset, 50% of the cropped data was selected as the training set, 20% as the validation set, and 30% as the test set. For the Vaihingen dataset, 80% of the cropped data was selected as the training set, 10% as the validation set and 10% as the test set. For the DFC2018 dataset, 80% of the cropped data were selected as the training set, 10% as the validation set and 10% as the test set.

All methods presented in following experiments have the same batch size and epoch. Since data enhancement technique can solve the overfitting problem and artificially increase the training sample when reading them from memory in each epoch [25], we implement data enhancement by randomly rotated the input images vertically and horizontally for all methods.

In training phrase, we use the AdaMax [31] algorithm for optimization. The weight decay is set to 0.0009. A *"poly"* policy [32] is used to optimize the learning rate. In all networks, the initial learning rate is set to 0.001 and is multiplied by $(1 - \frac{\text{iter}}{\text{max}-\text{iter}})^{\text{power}}$, where the *"power"* is set to 0.3, and the *"max-iter"* can be computed by multiplying the number of epochs with whole batches.

2) Evaluation Metrics: The networks presented in this article are evaluated using three popular metrics: overall accuracy (OA), mean intersection over union (MIoU), and frequency weighted intersection over union (FWIoU).

OA: The total number of samples correctly classified are divided by the total number of samples.

MIoU: It is a standard measure for semantic segmentation by calculating the mean value of intersection over unions (IoUs) of all classes.

FWIoU: This is an improved version of MIoU, which assigns different weights to classes according to their occurence frequencies. That is, FWIoU can be obtained by multiplying the IoUs of different classes with their corresponding weights and then summing the resulting products.

3) Parameter Setting: There are two parameters need to be set. The first parameter is the number of layers in DSR blocks, in Fig. 2. Taking the DFC2018 data as an example, the Fig. 5 shows the experimental results of the proposed method for binary-classification semantic segmentation with different layers in DSR blocks. The number of layers is set to four in DSR blocks yielding optimal segmentation results. The second parameter is the dilation rate (R) in Fig. 3. Table II shows the experimental

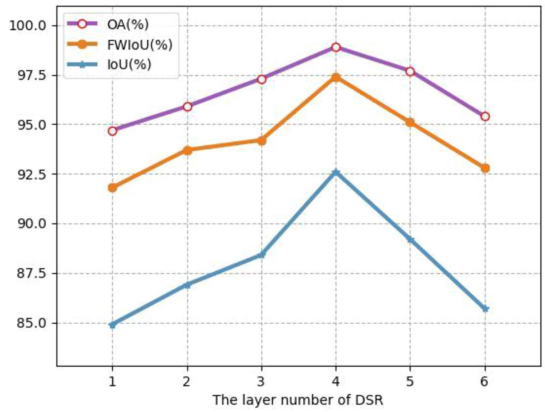


Fig. 5. Binary-classes semantic segmentation results of DSR block with different number of layers on DFC2018 dataset.

[1][Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-Sem-Label-Potsdam.aspx

[2][Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-Sem-Label-Vaihingen.aspx

[3][Online]. Available: https://hyperspectral.ee.uh.edu/?page_id=1075

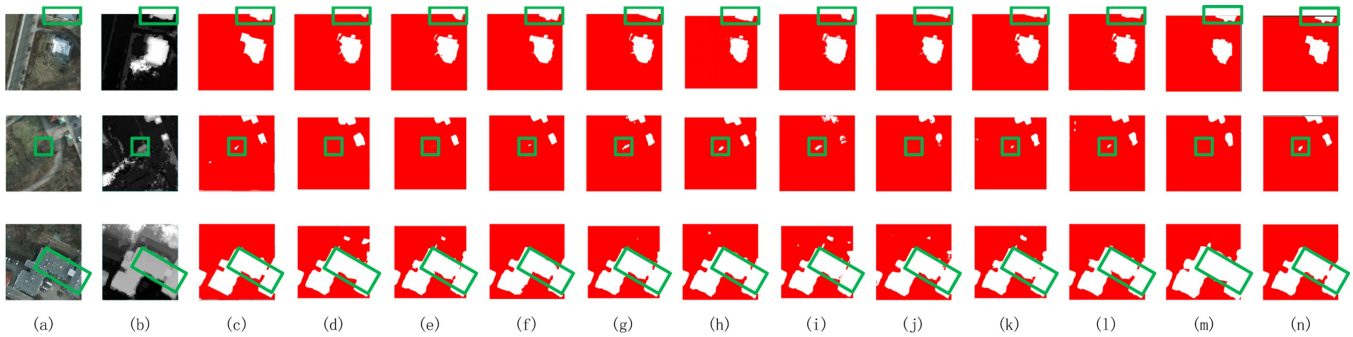


Fig. 6. Qualitative comparison of binary-classes semantic segmentation for different methods on Potsdam dataset. (a) RGB. (b) DSM. (c) Ground-truth. (d) ResUNet++. (e) PSPNet. (f) Deeplab V3+. (g) CGSANet. (h) ACFNet-Single. (i) FuseNet. (j) RedNet. (k) HAFNet. (l) $(MS)^2$-Net. (m) MCANet. (n)ACFNet. Red area represents background and white area represents buildings.

TABLE II
BINARY-CLASSES SEMANTIC SEGMENTATION RESULTS OF DIFFERENT R ON DFC2018 DATASET

First branch R	Second branch R	Third branch R	OA(%)	FWIoU(%)	IoU(%)
9	11	13	96.3	93.2	88.7
9	11	15	98.9	97.4	92.6
9	13	15	98.2	96.3	91.7
11	13	15	98.4	96.9	92.2

results of the proposed method for binary-classes semantic segmentation with different R. From the Table II, different R of different branches are set to 9, 11, and 15, respectively.

IV. RESULTS AND ANALYSIS

A. Comparison to State-of-the-Art Methods

We selected a range of state-of-the-art deep semantic segmentation models as competitors to fully validate the performance of the proposed ACFNet. To compare the performance of single-branch networks that only oriented to semantic segmentation of HR images, discarding LiDAR-branch. We simplify the structure of encoder in the proposed ACFNet to five layers of downsampling operation and then feed the resulting features into the IFG module. In the decoder stage, the cascade operation is omitted and the output of the IFG module is combined with four layers of DFCL with downsampling operation. We call this proposed simplified network ACFNet-Single. As shown in Tables III and IV, ACFNet-Single are compared with three classical deep models for semantic segmentation, including ResUNet++ [33], PSPNet [34], Deeplabv3+ [35], and CGSANet [36]. To evaluate the performance of dual-branch networks that joint using HR and LiDAR images, we compare the proposed ACFNet with several state-of-the-art cross-modal methods, including FuseNet [19], RedNet [37], HAFNet [21], $(MS)^2$-Net [38], and MCANet [39].

1) Binary-Classes Semantic Segmentation: The experimental results of the mentioned methods for building extraction are shown in Table III. Building extraction is a binary-classes semantic segmentation that treats the building class as the target and all other classes as background information. It can be observed that the proposed ACFNet produces the best OA, FWIoU, and MIoU compared to the other competing methods. For the Potsdam dataset, the OA/FWIoU/MIoU of ACFNet results in gains of 1.3%/2.3%/1.8% compared to the best dual-branch competitor $(MS)^2$-Net [38]. For the Vaihinge dataset, the OA/FWIoU/MIoU of ACFNet yields 1.0%/1.8%/2.8% performance as compared to the best dual-branch competitor RedNet [37]. For the DFC2018 dataset, the OA/FWIoU/MIoU of ACFNet notably higher than MCANet [39] by 2.1%/3.5%/3.1%. While, these metrics of the ACFNet-Single also have highest scores in single-branch methods.

Figs. 6– 8 visualize the semantic segmentation results for binary classification for the optimally trained network. As can be seen, ACFNet provides the most accurate and smoothest maps for all three datasets, compared to the other methods. First, although all algorithms generally perform well for large-area building extraction, the proposed method can obtain smooth maps, especially for retaining local information. Shown as the first-row maps in Fig. 6, the bottom-right corner of the first-row maps in Fig. 7, and the centre of the second-row maps in Fig. 8, ACFNet has a strong ability to extract local details of buildings. Second, ACFNet yields optimal extraction effect for small buildings. Show as the second-row maps in Fig. 6, the left of the first-row maps and the bottom-left of second-row maps in Fig. 7, and the top-left corner of the first-row maps in Fig. 8, ACFNet can get binary-classes semantic segmentation results consistent with the actual situation. Third, ACFNet has a satisfactory extraction effect on irregular buildings. Taking the first-row maps in Fig. 6 and the second-row maps in Fig. 8 as examples, although it is difficult for all algorithms to identify the concrete shape of complex buildings, whose contour visualized by ACFNet is the closest to their ground-truth.

In order to further evaluate the performance of the proposed ACFNet, we conducted precision–recall (P-R) experiments. *Precision* represents the accuracy of positive predictions and is used to measure the accuracy of binary-classes semantic segmentation in this work. *Recall* represents the proportion of positive predictions out of all accurate predictions and is conducive to compensate for missed positive predictions. In general, if the P-R curve of *model 1* is surrounded by that of *model 2* or the area enclosed by the curve of *model 2* covers a larger

TABLE III
EVALUATION OF METRIC SCORES (%) OF BINARY-CLASSES SEMANTIC SEGMENTATION FOR THREE DATASETS

		Single-branch networks					Dual-branch networks					
Dataset	Metrics	ResUNet++ [33]	PSPNet [34]	Deeplab V3+ [35]	CGSANet [36]	ACFNet-Single	FuseNet [19]	RedNet [37]	HAFNet [21]	$(MS)^2$-Net [38]	MCANet [39]	ACFNet
Potsdam	OA	95.6 ± 1.927	96.8 ± 0.557	97.5 ± 0.469	97.7 ± 0.345	98.3 ± 0.587	95.8 ± 1.592	96.3 ± 0.271	97.0 ± 0.350	97.6 ± 0.124	96.0 ± 0.498	**98.9 ± 0.153**
	FWIOU	91.7 ± 3.315	94.3 ± 0.682	95.2 ± 0.894	96.3 ± 0.779	96.7 ± 1.254	91.9 ± 3.030	93.7 ± 0.354	94.5 ± 0.652	95.5 ± 0.235	92.4 ± 0.564	**97.8 ± 0.290**
	MIOU	87.2 ± 7.330	91.4 ± 1.072	94.1 ± 1.524	94.7 ± 1.069	95.5 ± 1.102	87.6 ± 4.824	90.1 ± 1.130	91.5 ± 1.072	94.9 ± 0.264	91.7 ± 1.471	**96.7 ± 0.431**
Vaihingen	OA	94.1 ± 1.053	94.0 ± 1.261	96.2 ± 1.398	96.3 ± 1.279	96.9 ± 0.092	96.1 ± 0.251	96.8 ± 0.311	96.4 ± 0.145	95.3 ± 0.990	96.5 ± 0.764	**97.8 ± 0.149**
	FWIOU	90.7 ± 1.988	89.6 ± 2.292	93.0 ± 2.386	93.1 ± 2.378	94.2 ± 0.138	92.5 ± 0.445	94.5 ± 0.125	93.5 ± 0.272	92.1 ± 1.810	93.3 ± 0.439	**96.3 ± 0.237**
	MIOU	89.0 ± 2.448	88.8 ± 4.445	92.0 ± 4.465	92.1 ± 4.398	92.9 ± 0.358	91.1 ± 0.502	92.4 ± 0.502	92.1 ± 0.323	90.8 ± 1.906	92.5 ± 1.136	**95.2 ± 0.332**
DFC2018	OA	91.6 ± 1.023	88.3 ± 1.759	86.0 ± 0.602	88.1 ± 0.485	94.1 ± 0.139	83.2 ± 1.207	95.6 ± 0.114	93.7 ± 0.218	86.4 ± 0.300	96.8 ± 0.098	**98.9 ± 0.310**
	FWIOU	85.4 ± 1.985	82.1 ± 2.095	79.6 ± 1.167	82.4 ± 1.002	88.3 ± 0.233	77.8 ± 1.749	92.7 ± 0.275	89.7 ± 0.370	80.2 ± 0.595	93.9 ± 0.239	**97.4 ± 0.654**
	MIOU	75.0 ± 2.324	73.1 ± 2.823	65.6 ± 2.015	82.4 ± 2.412	81.2 ± 0.422	55.3 ± 1.940	85.3 ± 0.720	81.5 ± 0.467	66.7 ± 1.598	89.5 ± 0.579	**92.6 ± 0.802**

TABLE IV
EVALUATION OF METRIC SCORES (%) OF MULTI-CLASSES SEMANTIC SEGMENTATION FOR THREE DATASET

		Single-branch networks					Dual-branch networks					
Dataset	Metrics	ResUNet++ [33]	PSPNet [34]	Deeplab V3+ [35]	CGSANet [36]	CJSNet-single	FuseNet [19]	RedNet [37]	HAFNet [21]	$(MS)^2$-Net [38]	MCANet [39]	ACFNet
Potsdam	OA	77.2 ± 0.927	82.0 ± 0.522	84.5 ± 1.425	83.2 ± 1.210	85.2 ± 1.145	80.7 ± 0.502	87.1 ± 1.095	84.5 ± 1.254	87.6 ± 0.563	89.1 ± 0.414	**91.4 ± 0.357**
	FWIOU	64.4 ± 1.315	70.6 ± 1.006	74.9 ± 2.827	73.1 ± 2.223	75.6 ± 2.246	70.0 ± 0.983	78.3 ± 1.844	74.7 ± 2.104	78.9 ± 0.957	81.1 ± 0.863	**83.6 ± 0.903**
	MIOU	54.7 ± 2.330	58.7 ± 1.553	59.6 ± 3.960	59.1 ± 2.323	61.3 ± 2.437	57.0 ± 1.520	65.7 ± 3.373	59.3 ± 3.735	69.0 ± 1.648	70.3 ± 1.203	**73.6 ± 1.378**
Vaihingen	OA	88.3 ± 0.294	89.8 ± 1.029	85.2 ± 1.057	87.1 ± 0.835	91.4 ± 0.165	76.9 ± 0.663	92.2 ± 0.584	78.9 ± 0.212	90.6 ± 0.753	92.4 ± 0.388	**96.1 ± 0.274**
	FWIOU	80.0 ± 0.549	82.0 ± 2.660	75.4 ± 2.029	77.9 ± 1.242	83.9 ± 0.596	63.9 ± 1.175	86.1 ± 1.003	66.3 ± 0.377	83.4 ± 1.133	86.9 ± 0.555	**91.1 ± 0.531**
	MIOU	62.3 ± 0.848	70.9 ± 3.308	56.8 ± 1.861	60.9 ± 1.226	72.4 ± 0.906	49.1 ± 1.722	73.4 ± 1.665	51.4 ± 0.424	72.6 ± 1.320	79.3 ± 0.774	**81.6 ± 0.673**
DFC2018	OA	79.8 ± 0.741	77.5 ± 1.315	76.7 ± 1.900	78.1 ± 1.302	82.3 ± 0.376	69.6 ± 1.583	81.6 ± 1.320	72.3 ± 1.397	80.5 ± 0.516	84.5 ± 1.532	**87.8 ± 0.436**
	FWIOU	65.7 ± 1.442	63.2 ± 2.546	62.5 ± 3.509	67.4 ± 2.006	72.5 ± 0.586	54.2 ± 2.615	70.3 ± 1.322	56.5 ± 2.356	68.5 ± 0.731	74.5 ± 1.564	**77.2 ± 1.302**
	MIOU	49.2 ± 3.412	47.1 ± 3.148	46.0 ± 4.178	52.3 ± 3.074	58.7 ± 2.288	38.0 ± 3.204	57.6 ± 2.009	40.1 ± 3.068	56.3 ± 1.320	60.5 ± 2.266	**64.7 ± 2.128**

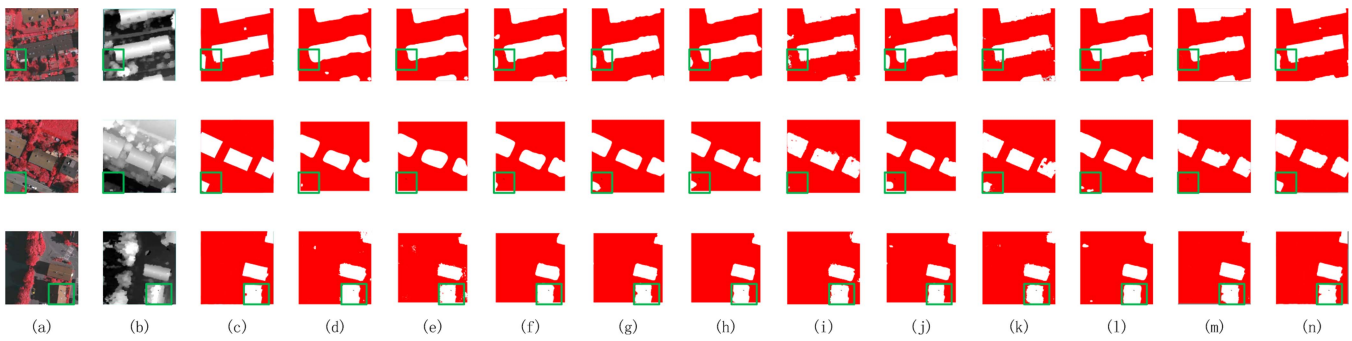


Fig. 7. Qualitative comparison of binary-classes semantic segmentation for different methods on Vaihingen dataset. (a) RGB. (b) DSM. (c) Ground-truth. (d) ResUNet++. (e) PSPNet. (f) Deeplab V3+. (g) CGSANet. (h) ACFNet-Single. (i) FuseNet. (j) RedNet. (k) HAFNet. (l) $(MS)^2$-Net. (m) MCANet. (n)ACFNet. Red area represents background and white area represents buildings.

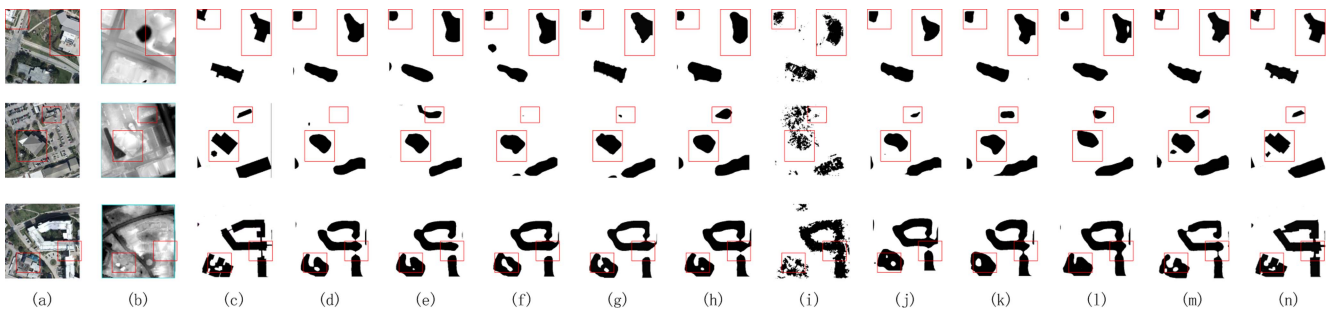


Fig. 8. Qualitative comparison of binary-classes semantic segmentation for different methods on DFC2018 dataset. (a) RGB. (b) DSM. (c) Ground-truth. (d) ResUNet++. (e) PSPNet. (f) Deeplab V3+. (g) CGSANet. (h) ACFNet-Single. (i) FuseNet. (j) RedNet. (k) HAFNet. (l) $(MS)^2$-Net. (m) MCANet. (n)ACFNet. White area represents background and black area represents buildings.

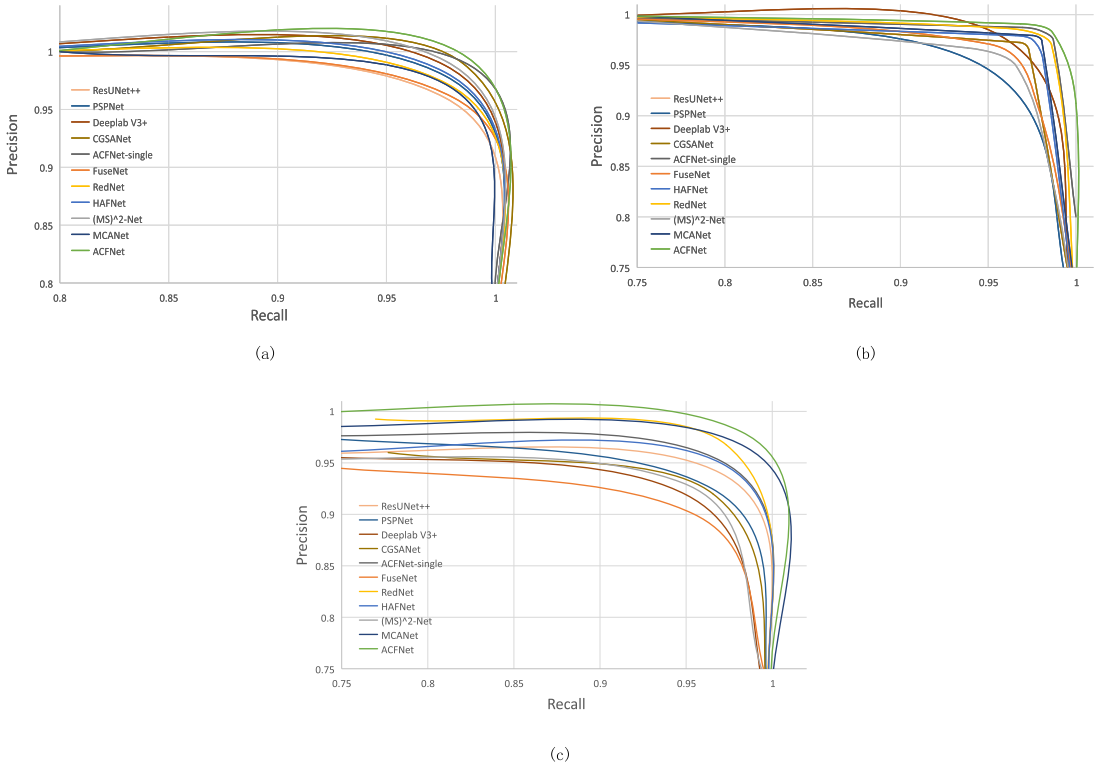


Fig. 9. P-R curves comparison of ResUNet++, PSPNet, Deeplab V3+, CGSANet, ACFNet-Single, FuseNet, RedNet, HAFNet, $(MS)^2$-Net, MCANet, and ACFNet. (a) Potsdam dataset. (b) Vaihingen dataset. (c) DFC2018 dataset.

area, the performance of *model 2* is considered to be better than that of *model 1*. Accordingly, the proposed ACFNet has optimal performance of binary-classes semantic segmentation for abovementioned three datasets, shown as Fig. 9.

To reflect the stability of the different algorithms, the performance of binary-classes semantic segmentation for three datasets as the training ratio varies is depicted in Fig. 10. It can be seen that the performance of the various algorithms decreases with decreasing number of training samples, which is as expected. However, ACFNet always produces higher accuracy and better stability as compared to the other algorithms, even with a small training ratio.

2) Multiclasses Semantic Segmentation: To verify the generalization ability of the proposed algorithm, experiments of multiclasses semantic segmentation also were conducted. The experimental results of different algorithms for multiclasses semantic segmentation on the three datasets are shown in Table IV. It can be observed that the proposed ACFNet produces the best OA, FWIoU, and MIoU compared to the other competing methods. For the Potsdam dataset, the OA/FWIoU/MIoU of ACFNet results in gains of 2.3%/2.5%/3.3% compared to the best dual-branch competitor MCANet [39]. For the Vaihingen dataset, the OA/FWIoU/MIoU of ACFNet yields 3.7%/4.2%/2.3% performance as compared to the best dual-branch competitor MCANet. For the DFC2018 dataset, the OA/FWIoU/MIoU of ACFNet notably higher than MCANet by 3.3%/2.7%/4.2%. While, these metrics of the ACFNet-Single also have highest scores in single-branch methods for multiclasses semantic segmentation.

Figs. 11–13 visualize the multiclasses semantic segmentation results of the best-trained networks. As can be seen, the proposed ACFNet provides the most accurate and smoothest maps for all three datasets, compared to the other methods. Fig. 11 shows visualization results of multiclasses semantic segmentation for the Potsdam dataset. The abovementioned methods have achieved good segmentation results for large areas, especially for *buildings* and *trees*. However, in terms of local details, the comparative models are not robust and effective enough to accurately identify *clutter*, *cars*, and *low vegetation*. For example, as shown in the box of the first-row maps, the *trees* distributed in the large areas of *low vegetation* are segmented by ACFNet accurately. This is also reflected in the following two cases. In the second-row maps, *trees* are segmented from the large areas of *low vegetation*; as shown in the red boxes in the upper-right corner of the third-row maps, *low vegetation* are segmented from the large areas of *trees*.

Fig. 12 shows visualization results of multiclasses semantic segmentation for the Vaihingen dataset. Similar to the Potsdam dataset, comparative models cannot segment local small objects, while ACFNet can effectively segment *clutter* and provide more accurate boundaries for *buildings* and *cars*. As shown in the first-row maps, the classes of *low vegetation*, *buildings,* and *impervious surfaces* can be segmented accurately by ACFNet. As can be seen from the second-row and third-row maps, ACFNet clearly outperforms the other networks for the classes of *trees*, *low vegetation*, and *cars*. Fig. 13 visualize the results of multiclasses semantic segmentation for the the GRSS DFC

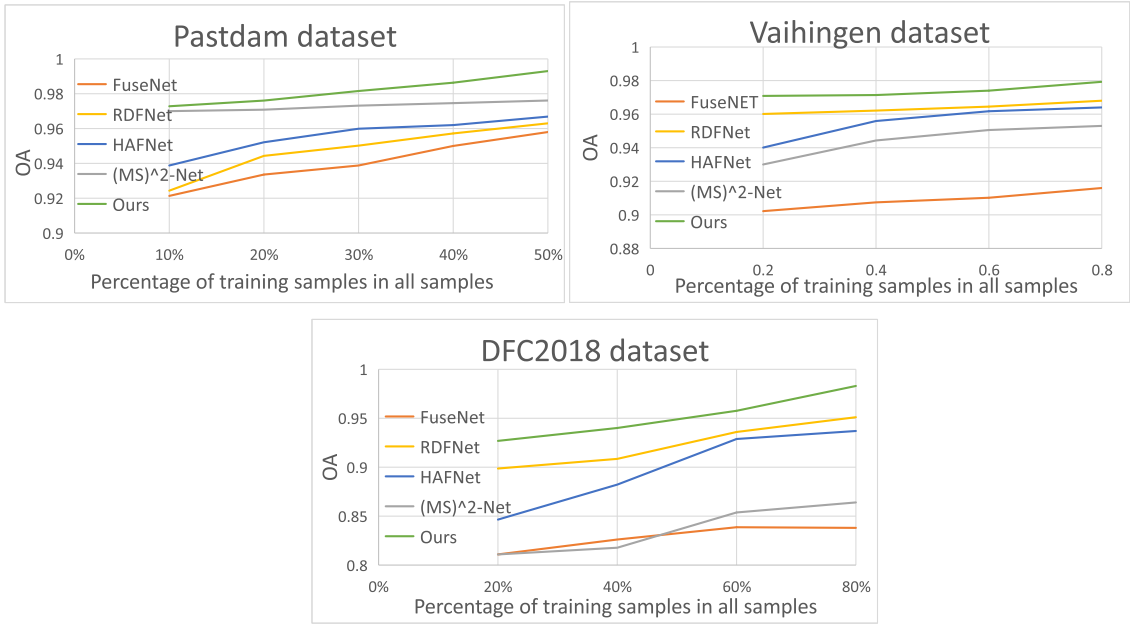


Fig. 10. OA of binary-classes semantic segmentation for varying training ratio. (a) Pastdam dataset. (b) Vaihingen dataset. (c) DFC2018 dataset.

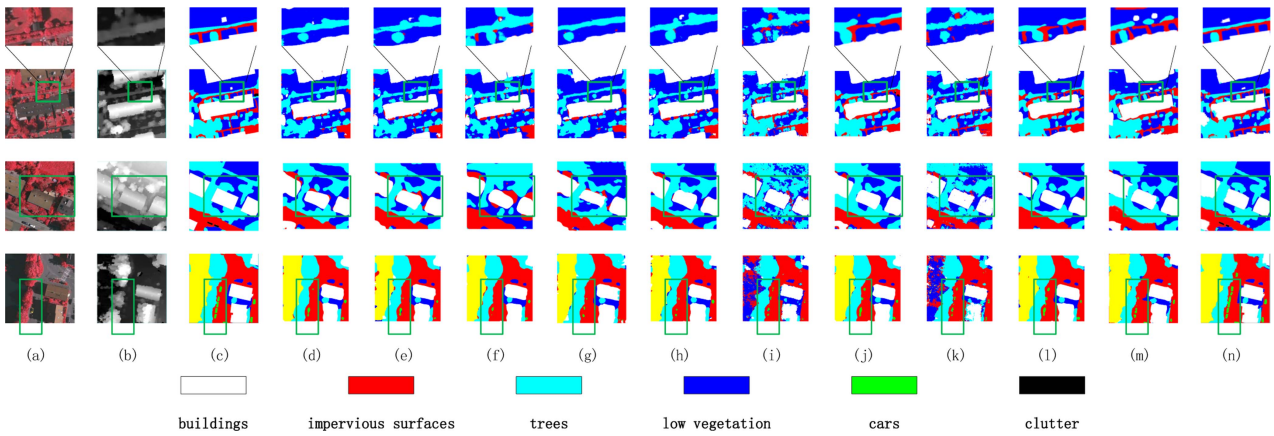


Fig. 11. Qualitative comparison of multiclasses semantic segmentaion for different methods on Potsdam dataset. (a) RGB. (b) DSM. (c) Ground-truth. (d) ResUNet++. (e) PSPNet. (f) Deeplab V3+. (g) CGSANet. (h) ACFNet-Single. (i) FuseNet. (j) RedNet. (k) HAFNet. (l) $(MS)^2$-Net. (m) MCANet. (n) ACFNet.

2018 dataset. As the RGB and DSM images were cropped, three sets of them were selected. As shown in the first-row, ACFNet is the best at identifying *main arterials* and *residential buildings*. Shown as in the second-row and third-row maps, *bare soil* and *main arterials*, respectively, incur fewer misclassifications from ACFNet than the other techniques.

In short, since the fusion network of HR and LiDAR data complements the elevation features compared to the single HR network, it is more accurate in processing semantic categories with elevation information, and this conclusion is reflected in Table III and Figs. 6–8. In addition, the visual results from Figs. 11–13 for multiclasses semantic segmentation show that the fusion networks can handle most of the semantic segmentation categories (e.g., trees, buildings, cars, impervious surfaces, low vegetation, etc.,) but they are not effective in handling a few semantic categories (e.g., stressed grasses) with no obvious elevation structure.

The OAs of multiclasses semantic segmentation for various networks are studied under varying training ratios for three datasets. Fig. 14 shows that the proposed ACFNet always produce higher OA compared to competitors. While the performance of all methods decreases as the number of training samples decreases, which is to be expected. However, ACFNet always achieves higher segmentation accuracy and better stability than other techniques, even at very small training rates.

From all the above experiments, we can conclude that the optimal algorithms in the comparison experiments are not consistent for different tasks and datasets. For example, for binary-classes

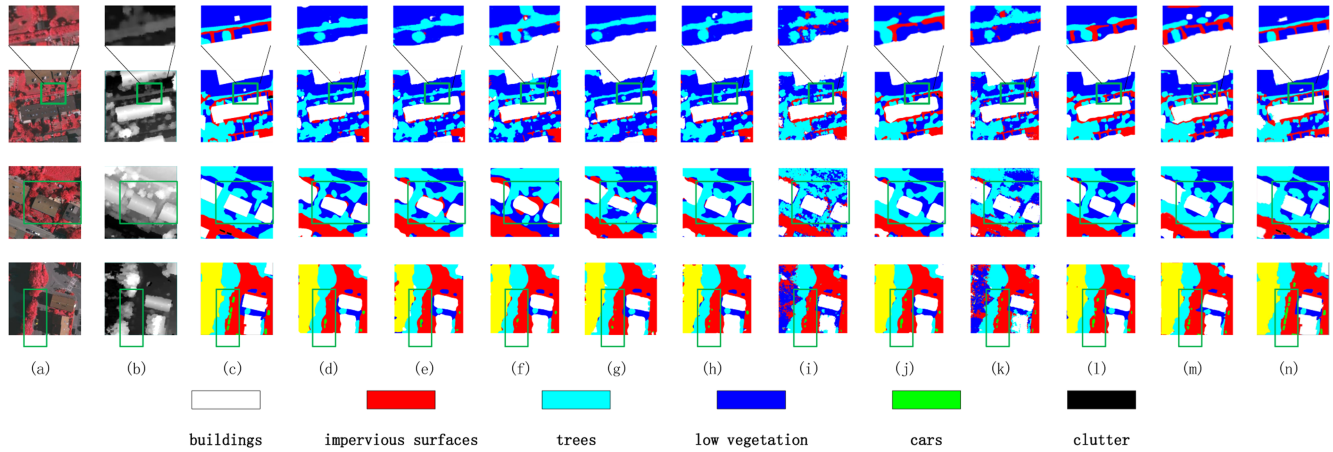


Fig. 12. Qualitative comparison of multiclasses semantic segmentaion for different methods on Vaihingen dataset. (a) RGB. (b) DSM. (c) Ground-truth. (d) ResUNet++. (e) PSPNet. (f) Deeplab V3+. (g) CGSANet. (h) ACFNet-Single. (i) FuseNet. (j) RedNet. (k) HAFNet. (l) $(MS)^2$-Net. (m) MCANet. (n) ACFNet.

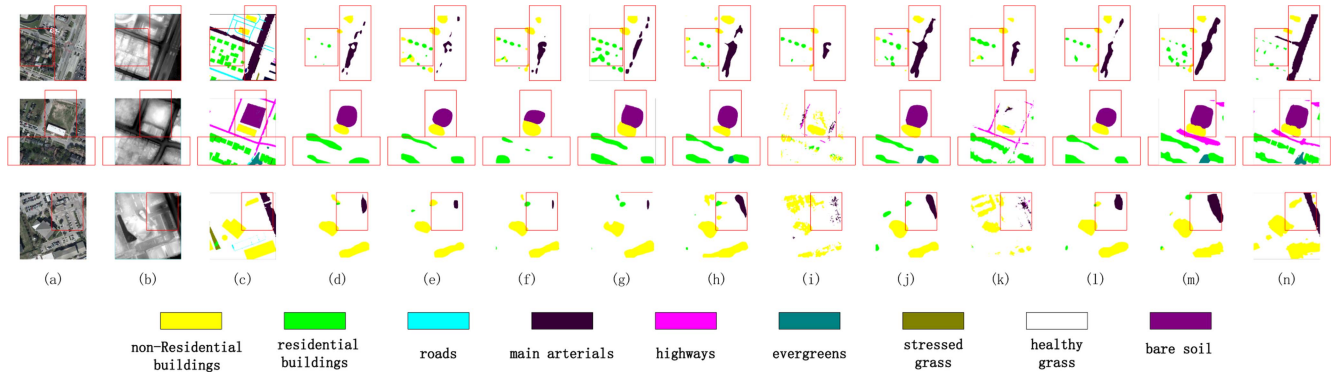


Fig. 13. Qualitative comparison of multiclasses semantic segmentaion for different methods on DFC2018 dataset. (a) RGB. (b) DSM. (c) Ground-truth. (d) ResUNet++. (e) PSPNet. (f) Deeplab V3+. (g) CGSANet. (h) ACFNet-Single. (i) FuseNet. (j) RedNet. (k) HAFNet. (l) $(MS)^2$-Net. (m) MCANet. (n) ACFNet.

semantic segmentation, the optimal algorithm for the Potsdam dataset is $(MS)^2$-Net, the optimal algorithm for the Vaihingen and the optimal algorithm for the DFC2018 datasets is MCANet. However, the proposed ACFNet is the strongest performer for all datasets and tasks compared to other methods, which reflects the higher stability of our network.

B. Ablation Experiments

We conducted a series of ablation experiments to validate the performance of the components in the ACFNet network for semantic segmentation in Vaihingen dataset. Specifically, we investigated the impact of using the ASF module, the important feature guided (IFG) module and the DFCL module within ACFNet. As shown in Table V, using binary-classes semantic segmentation in Vaihingen dataset as an example, seven different networks were designed to measure the impact of these three main components in the proposed ACFNet. For Network A, B, or C, only one module is employed to construct models, the OAs of Network A and Network B are higher than that of Network C, and the OA of Network B is higher than that of Network A. FWIoU and MIoU of these three networks have the same trend. To evaluate the effect of ASF module in ACFNet, we compare the experimental results of Network E and ACFNet. In Network E, ASF is replaced by traditional summation and data fusion strategy. The OA/FWIoU/MIoU of ACFNet is improved by 0.7%/2.4%/2.3% by using the ASF module. To evaluate the

TABLE V
EVALUATION OF METRIC SCORES (%) FOR ABLATION EXPERIMENTS

		ASF	IFG	DFCL	OA	FWIoU	MIoU
semantic segmentation for binary classification	Network A	✓			96.4	93.4	92.3
	Network B		✓		96.5	93.7	92.8
	Network C			✓	96.0	92.4	91.1
	Network D	✓	✓		97.3	94.2	93.2
	Network E		✓	✓	97.1	93.9	92.9
	Network F	✓		✓	97.0	93.5	92.7
	ACFNet	✓	✓	✓	**97.8**	**96.3**	**95.2**
Semantic segmentation for multi -categorisation	Network A	✓			92.7	83.5	70.9
	Network B		✓		93.7	87.4	75.3
	Network C			✓	92.5	82.9	69.1
	Network D	✓	✓		94.8	89.7	77.8
	Network E		✓	✓	94.5	89.3	77.2
	Network F	✓		✓	94.1	88.5	76.7
	ACFNet	✓	✓	✓	**96.1**	**91.1**	**81.6**

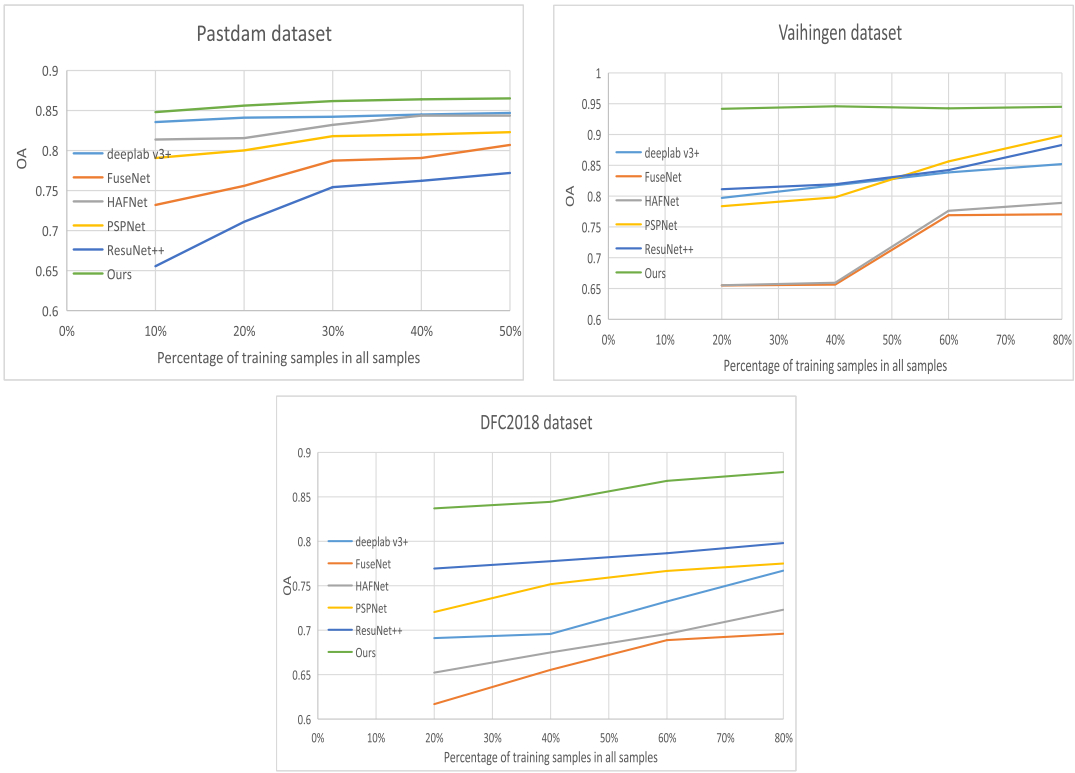


Fig. 14. OA of multiclasses semantic segmentation for varying training ratio. (a) Pastdam dataset. (b) Vaihingen dataset. (c) DFC2018 dataset.

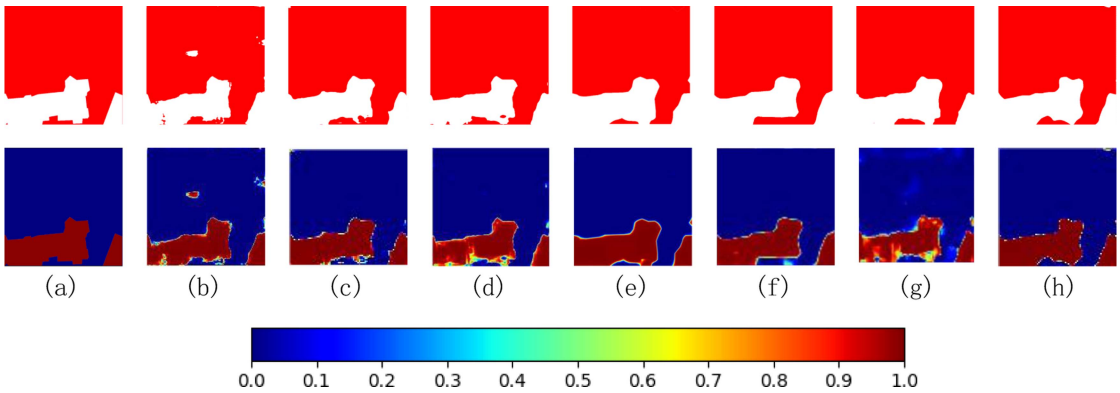


Fig. 15. Contrast result chart of ablation experiment. (a) Label. (b) Network A. (c) Network B. (d) Network C. (e) Network D. (f) Network E. (g) Network F. (h) ACFNet.

impact of the IFG module on the performance of ACFNet, we set up comparison experiments for Network F. It can be seen that OA/FWIoU/MIoU of ACFNet gains 0.8%/2.8%/2.5% comparing to that of Network F. To evaluate the impact of the DFCL module on the performance of ACFNet, we design Network D as comparative method. In Network D, traditional 2-D convolution is employed to instead of the DFCL module, and OA/FWIoU/MIoU of ACFNet is higher than that of Network D by 0.5%/2.0%/2.0%. Through the analysis of three factors, it can be seen that ASF and IFG modules have greater contribution for the performance of the proposed network. The separate use of the DFCL module is unsatisfactory, but its collaboration works effectively for binary-classes semantic segmentation. Semantic segmentation for multiclasses tasks proposes rules that are fully consistent with the binary-classes described above. Therefore, we fully consider the characteristics of ASF, DFCL, and IFG modules and combine them to improve the performance of the proposed ACFNet.

To improve the readability of this work, we also show the visualization maps of the fusion mixtures from different modules or module combinations. As shown in Fig. 15, the first row shows the corresponding binary-class semantic segmentation result for each strategy, and the second row shows the probability density maps. As can be seen from the Fig. 15, the visual effect of the

TABLE VI
COMPUTATION OF TRAINING TIME, TESTING TIME, FLOPS, AND PARAMETERS FOR THREE DATASETS

	Methods	ResUNet++	PSPNet	Deeplab V3+	CGSANet	FuseNet	RedNet	HAFNet	$(MS)^2$-Net	MCANet	ACFNet
Potsdam	Training time	90.76s	105.09s	97.52s	250.32S	140.89s	221.12s	214.44s	196.89s	230.51S	237.24s
	Testing time	33.20s	40.29s	37.78s	83.82S	59.30s	36.96s	77.93s	36.63s	71.77S	73.45s
	FLOPs	98.87007848G	249.210496G	139.31204096G	1097.71043661G	376.5403648G	132.2696704G	737.194673542G	136.77638064G	432.48688998G	670.07579648G
	Params	4.063078M	27.500678M	59.536291M	43.068675M	44.166723M	81.942403M	88.978706M	13.726894M	53.349427M	168.814131M
Vaihingen	Training time	28.71s	41.56s	31.13s	89.32S	44.24s	70.08s	68.52s	65.89s	75.83S	76.32s
	Testing time	3.62s	4.24s	3.78s	7.93S	5.52s	3.67s	6.82s	3.71s	6.05S	6.09s
	FLOPs	98.87007848G	249.210496G	139.31204096G	1097.71043661G	376.5403648G	132.2696704G	737.194673542G	136.77638064G	432.48688998G	670.07579648G
	Params	4.063078M	27.500678M	59.536291M	43.068675M	44.166723M	81.942403M	88.978706M	13.726894M	53.349427M	168.814131M
DFC2018	Training time	2.14s	2.78s	2.64s	6.71S	2.97s	5.07s	3.98s	4.23s	5.32S	5.65s
	Testing time	0.19s	0.26s	0.22s	1.32S	0.41s	0.21s	0.80s	0.22s	0.81S	0.89s
	FLOPs	98.87007848G	249.210496G	139.31204096G	1097.71043661G	376.5403648G	132.2696704G	737.194673542G	136.77638064G	432.48688998G	670.07579648G
	Params	4.063078M	27.500678M	59.536291M	43.068675M	44.166723M	81.942403M	88.978706M	13.726894M	53.349427M	168.814131M

fusion mixtures produced by ACFNet with all three designed modules matches the actual situation best, both in terms of high accuracy of the segmentation map and few noise of the probability map.

C. Number of Parameters and Calculation Time

To demonstrate the effectiveness of different semantic segmentation methods, training time, testing time, floating-point operations (FLOPs) and the number of parameters (Params) are shown as Table VI. The computational complexity of sing-branch methods is usually lower than that of dual-branch methods, due to their model design without considering LiDAR feature learning. In this work, ACFNet is designed as a dual-branch method yielding outstanding performance of semantic segmentation for binary classification and multicategorization. Compared with other state-of-the-art methods, the computational complexity of ACFNet is acceptable.

V. CONCLUSION

In this article, we present a deep ACFNet that can segment objects using HR and LiDAR images. The network performs segmentation excellently by the following three key aspects.

1) The ASF module is introduced, which can adaptively choose the optimal spatial scale for fusion according to the feature quality of multimodal data, making full use of the representation properties of ground object details in different resolution modalities.
2) The designed IFG module is oriented to extract fine-grained spatial-semantic information, and evaluate the influence weights of deep semantic features and shallow spatial detailed features on the segmentation results, achieving adaptive deep and shallow feature fusion, and reducing the semantic-spatial information dilution caused by layer-by-layer up-down sampling.
3) DFCL module is proposed to transform the feature map from spatial domain to frequency domain. Based on the traditional Fourier transform, two odd–even 2-D FFT are used and the amplitude and phase in the frequency domain are adjusted to ensure the computational volume and at the same time better model the dependence of the contextual information, so as to improve the segmentation accuracy of the complex urban features, especially the occluded features.

Comparative experiments and ablation studies of the semantic segmentation for binary classification and multicategorization tasks show that ACFNet has a strong competitive and generalization performance.

REFERENCES

[1] K. Makantasis, A. D. Doulamis, N. D. Doulamis, and A. Nikitakis, "Tensor-based classification models for hyperspectral data analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 6884–6898, Dec. 2018.
[2] K. Makantasis, A. Voulodimos, A. Doulamis, N. Doulamis, and I. Georgoulas, "Hyperspectral image classification with tensor-based rank-R learning models," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 3148–3152.
[3] X. Zhang, P. Xiao, X. Feng, J. Wang, and Z. Wang, "Hybrid region merging method for segmentation of high-resolution remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 98, pp. 19–28, Dec. 2014.
[4] R. Jiang, S. Mei, M. Ma, and S. Zhang, "Rotation-invariant feature learning in VHR optical remote sensing images via nested siamese structure with double center loss," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3326–3337, Apr. 2021.
[5] Y. Liu, J. Liu, X. Ning, and J. Li, "MS-CNN: Multiscale recognition of building rooftops from high spatial resolution remote sensing imagery," *Int. J. Remote Sens.*, vol. 43, no. 1, pp. 270–298, Jan. 2022.
[6] C. Wu, B. Du, and L. Zhang, "Fully convolutional change detection framework with generative adversarial network for unsupervised, weakly supervised and regional supervised change detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9774–9788, Aug. 2023.
[7] X. Chen et al., "Adaptive effective receptive field convolution for semantic segmentation of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3532–3546, Apr. 2021.
[8] 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.
[9] S. Freire et al., "Introducing mapping standards in the quality assessment of buildings extracted from very high resolution satellite imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 90, pp. 1–9, Apr. 2014.
[10] W. Nurkarim and A. W. Wijayanto, "Building footprint extraction and counting on very high-resolution satellite imagery using object detection deep learning framework," *Earth Sci. Informat.*, vol. 16, pp. 515–532, 2022.
[11] 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.

[12] Y. Yu et al., "Building extraction from remote sensing imagery with a high-resolution capsule network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jul. 2022, Art no. 8015905.
[13] H. Jung, H.-S. Choi, and M. Kang, "Boundary enhancement semantic segmentation for building extraction from remote sensed image," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2021, Art no. 5215512.
[14] Y. Sun, Z. Fu, C. Sun, Y. Hu, and S. Zhang, "Deep multimodal fusion network for semantic segmentation using remote sensing image and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2018, Art no. 5404418.
[15] N. Audebert, B. L. Saux, and S. Lefevre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 140, pp. 20–32, Jun. 2018.
[16] Q. Yuan, H. Z. M. Shafri, A. H. Alias, and S. J. b. Hashim, "Multiscale semantic feature optimization and fusion network for building extraction using high-resolution aerial images and LiDAR data," *Remote Sens.*, vol. 13, no. 13, Jul. 2021, Art. no. 2473.
[17] S. A. N. Gilani, M. Awrangjeb, and G. Lu, "An automatic building extraction and regularisation technique using LiDAR point cloud data and orthoimage," *Remote Sens.*, vol. 8, no. 3, Mar. 2016, Art. no. 258.
[18] S. Piramanayagam, E. Saber, W. Schwartzkopf, and F. W. Koehler, "Supervised classification of multisensor remotely sensed images using a deep learning framework," *Remote Sens.*, vol. 10, no. 9, Sep. 2018, Art. no. 1429.
[19] 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.*, 2016, pp. 213–228.
[20] V. Badrinarayanan, A. Handa, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," *Comput. Sci.*, vol. 39, no. 12, pp. 2481–2495, 2015.
[21] P. Zhang et al., "A hybrid attention-aware fusion network (HAFNet) for building extraction from high-resolution imagery and LiDAR data," *Remote Sens.*, vol. 12, no. 22, Nov. 2020, Art. no. 3764.
[22] X. Rui, Z. Li, Y. Cao, Z. Li, and W. Song, "DILRS: Domain-incremental learning for semantic segmentation in multi-source remote sensing data," *Remote Sens.*, vol. 15, no. 10, May 2023, Art. no. 2541.
[23] X. Zhang, H. Jiang, N. Xu, L. Ni, C. Huo, and C. Pan, "MsIFT: Multi-source image fusion transformer," *Remote Sens.*, vol. 14, no. 16, Aug. 2022, Art. no. 4062.
[24] J. Zhao et al., "Multi-source collaborative enhanced for remote sensing images semantic segmentation," *Neurocomputing*, vol. 493, pp. 76–90, Jul. 2022.
[25] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, (Series Lecture Notes in Computer Science), B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., 2016, pp. 630–645.
[26] L. Kaiser, A. N. Gomez, and F. Chollet, "Depthwise separable convolutions for neural machine translation," 2017, *arXiv:1706.03059*.
[27] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.
[28] H. Zhang, K. Zu, J. Lu, Y. Zou, and D. Meng, "EPSANet: An efficient pyramid squeeze attention block on convolutional neural network," in *Proc. Asian Conf. Comput. Vis.*, 2023, pp. 541–557.
[29] J. Huang, X. Zhang, Q. Xin, Y. Sun, and P. Zhang, "Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network," *ISPRS J. Photogrammetry Remote Sens.*, vol. 151, pp. 91–105, May 2019.
[30] Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou, "Global filter networks for image classification," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 980–993.
[31] D. Kingma and J. B. Adam, "A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, vol. 5, no. 6, 2015.
[32] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[33] D. Jha et al., "ResUNet++: An advanced architecture for medical image segmentation," in *Proc. IEEE Int. Symp. Multimedia*, 2019, pp. 225–230.
[34] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. 30TH IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.
[35] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.
[36] S. Chen, W. Shi, M. Zhou, M. Zhang, and Z. Xuan, "CGSANet: A contour-guided and local structure-aware encoder-decoder network for accurate building extraction from very high-resolution remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1526–1542, 2022.
[37] J. Jiang, L. Zheng, F. Luo, and Z. Zhang, "RedNet: Residual encoder-decoder network for indoor RGB-D semantic segmentation," 2018, *arXiv:1806.01054*.
[38] J. Zhao, Y. Zhou, B. Shi, J. Yang, D. Zhang, and R. Yao, "Multi-stage fusion and multi-source attention network for multi-modal remote sensing image segmentation," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 6, pp. 1–20, Nov. 2021.
[39] X. Li et al., "MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification," *Int. J. Appl. Earth Observation Geoinformation*, vol. 106, 2022, Art. no. 102638.

Zhen Ye received the B.S., M.S., and Ph.D. degrees in information and communication engineering from Northwestern Polytechnical University, Xi'an, China, in 2007, 2010, and 2015, respectively.

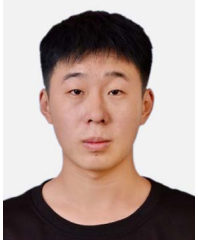
She spent one year as an exchange student of Mississippi State University, Mississippi State, MS, USA. She is currently an Associate Professor with the School of Electronics and Control Engineering, Chang'an University, Xi'an. Her research interests include hyperspectral image analysis, pattern recognition, and machine learning.

Zhen Li received the B.E. degree in mechatronics engineering from the Xi'an University of Science and Technology, Xi'an, China, in 2021. She is currently working toward the M.S. degree in control science and engineering with the School of Electronics and Control Engineering, Chang'an University, Xi'an.

Her current research focuses on multisource remote sensing image fusion and classification.

Nan Wang received the B.E. degree in remote sensing and photogrammetry from Information Engineering University, Zhengzhou, China, in 2009, and the M.S. and Ph.D. degrees in information science and technology from Wuhan University, Wuhan, China, in 2011 and 2014.

She spent two years as a Postdoctoral Researcher with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. She is currently an Associate Professor with the School of Information and Electronics, Beijing Institute of Technology, Beijing. Her research interests include multispectral/hyperspectral image analysis, pattern recognition, and machine learning.



Yuan Li received the B.E. degree in automation from the Jinling Institute of Technology, Nanjing, China, in 2022. He is currently working toward the M.S.degree in control science and engineering with the School of Electronics and Control Engineering, Chang'an University, Xi'an, China.

His current research interests are multistage fusion and multisource network for remote sensing image segmentation.

Wei Li (Senior Member, IEEE) received the B.E. degree in telecommunications engineering from Xidian University, Xi'an, China, in 2007, the M.S. degree in information science and technology from Sun Yat-Sen University, Guangzhou, China, in 2009, and the Ph.D. degree in electrical and computer engineering from Mississippi State University, Starkville, MS, USA, in 2012.

He spent one year as a Postdoctoral Researcher with the University of California, Davis, CA, USA. He is currently a Professor with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China. He has authored or coauthored more than 160 peer-reviewed articles and 100 conference papers. His research interests include hyperspectral image analysis, pattern recognition, and target detection.

Dr. Li is currently working as an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS). He was an Associate Editor for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS) and the IEEE SIGNAL PROCESSING LETTERS. He was the recipient of the JSTARS Best Reviewer in 2016 and TGRS Best Reviewer Award in 2020 from the IEEE Geoscience and Remote Sensing Society (GRSS) and the Outstanding Paper Award at IEEE International Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (Whispers) in 2019.