SCIENCE CHINA
Information Sciences

April 2023, Vol. 66 140305:1–140305:14
https://doi.org/10.1007/s11432-022-3599-y

• RESEARCH PAPER •
Special Topic: Artificial Intelligence Innovation in Remote Sensing

MFVNet: a deep adaptive fusion network with multiple field-of-views for remote sensing image semantic segmentation

Yansheng LI, Wei CHEN*, Xin HUANG, Zhi GAO, Siwei LI, Tao HE & Yongjun ZHANG*

School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079*, China*

Received 19 June 2022/Revised 25 August 2022/Accepted 14 October 2022/Published online 27 March 2023

Abstract In recent years, the remote sensing image (RSI) semantic segmentation attracts increasing research interest due to its wide application. RSIs are difficult to be processed holistically on current GPU cards on account of their large field-of-views (FOVs). However, the prevailing practices such as downsampling and cropping will inevitably decrease the quality of semantic segmentation. To address this conflict, this paper proposes a new deep adaptive fusion network with multiple FOVs (MFVNet), which is specially designed for RSI semantic segmentation. Different from existing methods, MFVNet takes into consideration the differences among multiple FOVs. By pyramid sampling the RSI, we first obtain images on different scales with multiple FOVs. Images on the high scale with a large FOV can capture larger spatial contexts and complete object contours, while images on the low scale with a small FOV can keep the higher spatial resolution and more detailed information. Then scale-specific models are chosen to make the best predictions for all scales. Next, the output feature maps and score maps are aligned through the scale alignment module to overcome spatial misregistration among scales. Finally, the aligned score maps are fused with the help of adaptive weight maps generated by the adaptive fusion module, producing the fused prediction. The performance of MFVNet surpasses the previous state-of-the-art semantic segmentation models on three typical RSI datasets, demonstrating the effectiveness of the proposed MFVNet.

Keywords semantic segmentation, remote sensing image (RSI), field-of-view (FOV), adaptive fusion, convolutional neural network

Citation Li Y S, Chen W, Huang X, et al. MFVNet: a deep adaptive fusion network with multiple field-of-views for remote sensing image semantic segmentation. Sci China Inf Sci, 2023, 66(4): 140305, https://doi.org/10.1007/s11432-022-3599-y

1 Introduction

With the advance in remote sensing technologies [1–5], the quantity of remote sensing images (RSIs) has been thoroughly increased. Due to the wide applications such road extraction [6], wildfire detection [7], building extraction [8], cloud detection [9], land-cover classification [10–12], and change detection [13,14]. RSI semantic segmentation has been attracting growing research interest in recent years.

One RSI generally has a large field-of-view (FOV), consisting of extensive geospatial objects (e.g., buildings, forests, and water) and abundant geospatial information [15]. When conducting RSI semantic segmentation based on visual interpretation [16], humans tend to observe the whole RSI on different scales and classify the pixels of geospatial objects by combining the local details and global structures, to make good use of the geospatial information inside the RSI [15–18].

On the contrary, the large FOV of RSI brings big trouble to machines. Although the rapid development of deep neural networks has made extraordinary progress in semantic segmentation [19–31], these machine-vision-based methods fail to process one RSI holistically, limited by the GPU memories. The prevailing practice to segment one RSI is either to downsample it to a smaller one or to separately segment the

* Corresponding author (email: weichenrs@whu.edu.cn, zhangyj@whu.edu.cn)

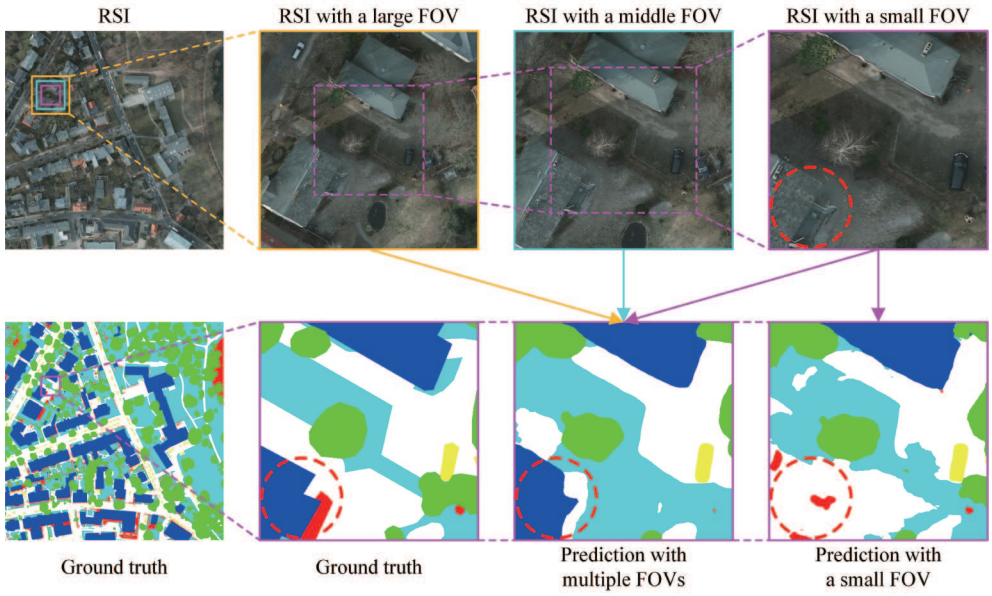

Figure 1 (Color online) The necessity of multiple FOVs. The upper row shows one RSI and three parts of it with different FOVs which are getting smaller from left to right. The lower row is the ground truths and predictions. They are the ground truth of the whole RSI, the ground truth of the part with a small FOV, the fused prediction with multiple FOVs, and the prediction with only a small FOV. As can be seen, small FOV leads to incomplete and erroneous predictions (i.e., the last column). However, with the help of multiple FOVs, the quality of segmentation is remarkably improved (i.e., the third column).

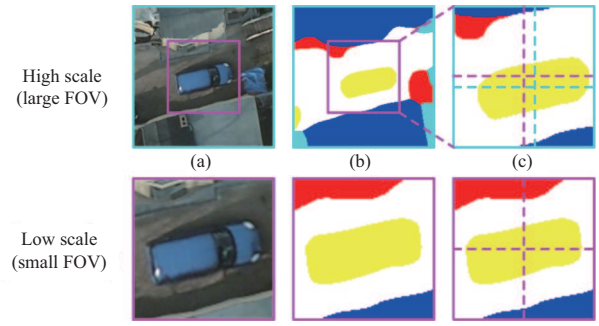

Figure 2 (Color online) The problem of spatial misregistration. (a) Images; (b) predictions; (c) spatial misregistration between scales. The lower part displays the image on the low scale, while the higher part displays the image on the high scale. The pink dashes and blue dashes denote the center points of the predicted car on different scales. As can be seen, they are not aligned well since resampling in the fusion process results in spatial misregistration.

partitioned patches and merge their results into a whole one. Downsampling keeps the FOV of one RSI but definitely ruins the fine details and geospatial information inside one RSI. The cropped patches maintain the image quality but obtain a pretty smaller FOV, which tends to deliver incomplete information and may easily lead to incorrect semantic segmentation. These two trivial practices both have their own imperfections.

As depicted in the fourth column of Figure 1, due to the lack of complete structure of geospatial objects, it is hard for both humans and machines to recognize the building in the lower left corner when only the RSI with a small FOV is given. The pixels of that building look similar to the ground (i.e., impervious surfaces) and may lead to confusion without the context (i.e., structural information) [32]. The naive fusion of multiple FOVs is one straight-forward solution to this. However, the use of multiple FOVs results in new problems. As can be seen in Figure 2, the predictions of multiple FOVs are not aligned well because of the spatial misregistration caused by resampling in the fusion process, which is needed to be considered before the fusion of multiple FOVs.

Therefore, we propose a deep adaptive fusion network with multiple FOVs (MFVNet) in this paper, to make full use of the large FOV of the RSI. By cropping the RSI with different pyramid sampling rates, images with multiple FOVs are obtained and resampled to the same size. Consequently, the image with a small FOV (i.e., low scale) preserves precise locations and contours while the image with a large

FOV (i.e., high scale) provides complete contexts and structural information. As embodied in the third column of Figure 1, the combination of them can enhance the robustness of semantic segmentation. Since images with multiple FOVs are different in their actual spatial resolutions, the geospatial information is expressed in different scales and a single model is unable to process all of them well. Thus, we perform the scale-specific model (SSM) searching on each scale to choose the best model for a specific scale. The images on different scales are separately fed to the corresponding SSMs, which produce scale-specific feature maps and score maps. After that, the feature maps and score maps from high scales are aligned to the low scale by the scale alignment modules (SAMs) to tackle the spatial misregistration between scales as illustrated in Figure 2. Moreover, it is easy to find that images from high scales specialize in capturing continuous geospatial objects and complete outlines, whereas images from low scales are good at keeping precise boundaries and detailed information. Intuitively, it is preferable to perform adaptive fusion instead of naive average or maximum fusion among these scales to combine their advantages. Therefore, the feature maps from all scales are concatenated and used to compute the scale-specific weight maps. Finally, the score maps on different scales are fused adaptively with the help of weight maps in the adaptive fusion modules (AFMs), which generate the final prediction result. Then we carry out detailed studies and analyses on the proposed MFVNet in the experiment part. The effectiveness of MFVNet is verified on three typical RSI datasets including GID [33], GF1WHU [34], and Potsdam. MFVNet achieves considerable performance gains compared with all previous state-of-the-art semantic segmentation models [21,22,24,26,35,36] and methods designed for RSI semantic segmentation [32,37–39].

Recently, some methods attempt to jointly and complementarily utilize local and global contexts to handle the semantic regions with large variations in RSIs, the ideas of which are similar to our proposed MFVNet. GLNet [39] takes the downsampled entire image and its cropped local patches as respective inputs and fuses feature maps from two branches, capturing both the high-resolution fine structures from zoomed-in local patches and the contextual dependency from the downsampled input. FCTL [32] introduces a locality-aware contextual correlation-based segmentation model to process local image patches and presents a contextual semantics refinement network that is enabling to reduce the boundary artifacts and refine mask contours during the process of creating the final mask. Here is the main difference between our MFVNet and these methods: they attempt to utilize multiple FOVs but do not take into consideration the differences and spatial misregistration among multiple FOVs, so the advantages of multiple FOVs are not well utilized. While MFVNet uses SSM for model choosing, SAM for spatial alignment, and AFM for adaptive fusion, which better exploits the multiple FOVs in RSI semantic segmentation. Codes and pre-trained models will be made publicly available online along with this paper (https://github.com/weichenrs/MFVNet). To summarize, the main contributions of this paper are three-fold.

(1) This paper proposes a novel deep adaptive fusion network with multiple FOVs called MFVNet to make full use of the large FOV of the RSI. MFVNet can take all existing and future semantic segmentation models as its backbones (i.e., SSMs). Besides, it can be trained in an end-to-end way even with its complicated architecture.

(2) This paper proposes SAM to solve the spatial misalignment among different scales in RSI for the first time. And AFM is proposed to fuse the predictions on different scales adaptively, which combines the strengths of different scales.

(3) Extensive experiments and analyses on three typical RSI datasets indicate the efficacy of MFVNet. The proposed MFVNet is compared with previous state-of-the-art semantic segmentation models and methods specially designed for RSI semantic segmentation. Our proposed MFVNet achieves state-of-the-art results on those datasets.

The rest of this paper is organized as follows. Section 2 reviews the related work of this study. Section 3 systematically introduces the proposed MFVNet. Section 4 reports the datasets used in this study and the experimental results. The conclusion and discussion are summarized in Section 5.

2 Related work

Since the concept of multiple FOVs has not been discussed before, here we relate some relevant studies with similar ideas about it in the following.

2.1 Long-range contextual dependency in one single field-of-view

Common deep neural networks designed for semantic segmentation explicitly exploit multi-scale fusion of features. To surpass the limit of the local receptive field by convolution layer and to capture the contextual information at multiple scales, multi-scale feature fusion is mainly explored in natural image semantic segmentation [22,24,26,40]. Dilated convolutions are used to aggregate multi-scale contextual information without losing resolution or analyzing rescaled images [40]. ASPP [22] applies several parallel atrous convolutions with different rates, while PSPNet [24] performs pooling operations at different grid scales. HRNet [26] maintains high-resolution representations by connecting high-to-low resolution convolutions in parallel and repeatedly conducting multi-scale fusions across parallel convolutions, which produces not only strong but also spatially precise high-resolution representations. Although these methods exploit the multi-scale feature fusion, only limited context information inside the image with a small FOV can be used.

Along with the tremendous success of self-attention-based Transformers [41,42] in the natural language processing domain, self-attention is adapted to the computer vision domain [35,36,43–46] to capture the global spatial context (i.e., long-range context). DANet [35] introduces a self-attention mechanism to capture feature dependencies in the spatial and channel dimensions, respectively. OCNet [43] presents an object context that aims at only gathering the pixels that belong to the same category as a given pixel as its context. CCNet [36] harvests the contextual information of all the pixels on its crisscross path. By taking a further recurrent operation, each pixel can finally capture the full-image dependencies from all pixels. Recently, with the proposal of Vision Transformer [47], the computer vision domain has been witnessing the breakthrough on Transformers [27, 48, 49], which shows huge potential in semantic segmentation for capturing the long-range contextual dependency. Similarly, the contexts explored by these methods are also limited to the input image with a small FOV, which restricts the performance from further improvements.

2.2 Global-local contextual dependency among multiple field-of-views

Recently, preliminary explorations about global-local contextual dependency on images have been conducted, which is similar to multiple FOVs [32, 37–39, 50]. A two-stage multi-scale training strategy is utilized in a semantic embedding network in [37], which is designed to fuse complementary information learned from multiple levels to make predictions. WiCNet [38] uses an extra context branch to explicitly model the context information in a larger image area, where the information communication between context branches is built through a context Transformer. Nevertheless, these methods have not addressed the spatial registration problem in the fusion process, and the fusion strategies of some are inflexible, which need further consideration.

Recent studies [32, 39, 50] take into consideration the outside contexts for adaptive fusion. GLNet [39] proposes to incorporate the local and global information via a two-stream network that separately processes the downsampled global image and cropped local patches, as well as a feature-sharing module that shares the concatenated local and global features in both streams. CascadePSP [50] uses a global step to refine the entire image and provides sufficient image contexts for the subsequent local step to perform full-resolution high-quality refinement. FCTL [32] leverages the locality-aware contextual correlation and the adaptive feature fusion scheme, which associates and combines local-context information to strengthen local segmentation. And a contextual semantics refinement network is used to leverage the relevance of local segmentation and context mask to avoid boundary-vanishing artifacts and refine the local semantic mask. Generally, these methods do not take into consideration the differences among multiple field-of-views. Furthermore, they have not addressed the spatial registration problem in the fusion process, and the fusion strategies of some are inflexible, which is required to be further explored.

3 Method

In this section, we will systematically introduce the proposed MFVNet. An overview of the architecture of MFVNet is presented in Figure 3. First, the original RSIs are cropped with different pyramid sampling rates and resampled to the same size. In detail, the original RSIs are down-sampled according to different sampling rates, which will be discussed in the implementation detail in Subsection 4.2. The sampling rates indicate the scales (i.e., FOVs). Then a fixed-size sliding window is used to crop the sampled RSIs

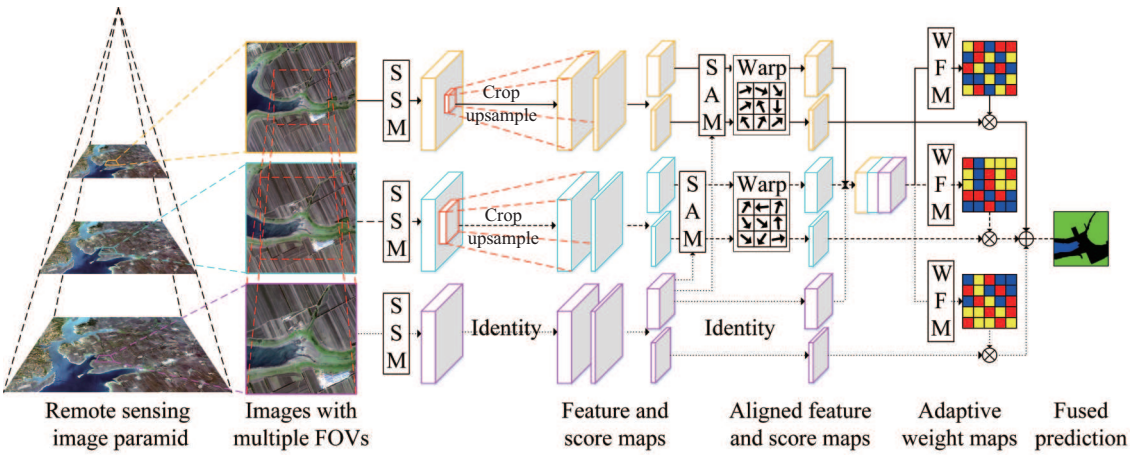

Figure 3 (Color online) The overall architecture of the proposed MFVNet. The accessible spatial contexts are enlarged by pyramid sampling the original RSIs, forming images on different scales with multiple FOVs. SSM searching is performed on each scale to choose the best model for the specific scale. The images are then fed to the corresponding SSMs and scale-specific feature maps and score maps are produced. They are aligned via the SAMs and fused adaptively with AFMs, which generates the final prediction result.

from all scales. It is noted that when dealing with the image patches on the fringe of the original whole image, since there is no more information outside the original image, we will expand the original image by flipping it along the fringe. After that, the image patches are obtained by cropping the expanded image. As the down-sampling rates are different, the windows from all scales have the same size but contain different spatial context information. Where the scale is higher, the FOV is larger. In this way, images with multiple FOVs can be obtained. Afterward, these images will go through several special-designed modules in the MFVNet to generate the final fused prediction.

3.1 Scale-specific model searching

Since the architecture designs of deep neural network models vary a lot, their focuses and abilities for extracting information are quite different. Theoretically and practically, a model usually fails to perform well on all scales and all datasets. Thus, we conduct the scale-specific model searching. For images on scale i, they are used to train several semantic segmentation models contained in the model pool. After training, these models are evaluated on the corresponding scale i. Then the SSM will be selected according to the performances of these models. The SSM searching process can be approximately depicted as

$$\theta_{ss}^{i} = \mathrm{argmin}(L_{\mathrm{ce}}(\sigma_{ss}^{i}(I^{i}), M^{i})), \quad \sigma_{ss}^{i} \in \text{model pool}, \tag{1}$$

where I^i and M^i denote the image and label on scale i, respectively; σ_{ss}^{i} denotes the candidate semantic segmentation model; L_{ce} denotes the cross-entropy loss; and θ_{ss}^{i} denotes the selected SSM of scale i. For middle scale and high scale, the original RSIs are down-sampled with the rate of 1.5 and 2. Then they are cropped into 512×512 tiles with automatically computed strides according to the image sizes. Three scales are used considering the trade-off between efficacy and efficiency. The necessity of SSM will be discussed in Subsection 4.2. Then the SSMs of all scales are used to compute the feature maps and score maps as

$$F^{i}, P^{i} = \theta_{ss}^{i}(I^{i}), \tag{2}$$

where F^i and P^i denote the feature maps and score maps of image I^i on scale i, respectively. Note that F^i and P^i from larger scales are center-cropped and up-sampled according to the valid region of I^i, as shown in Figure 3.

3.2 Scale alignment module

SSM searching in this subsection tends to select different models for each scale. Due to the resampling inside the network models and the original spatial resolution difference among scales, we are inevitably faced with spatial misregistration as depicted in Figure 2, which has not been considered well by most

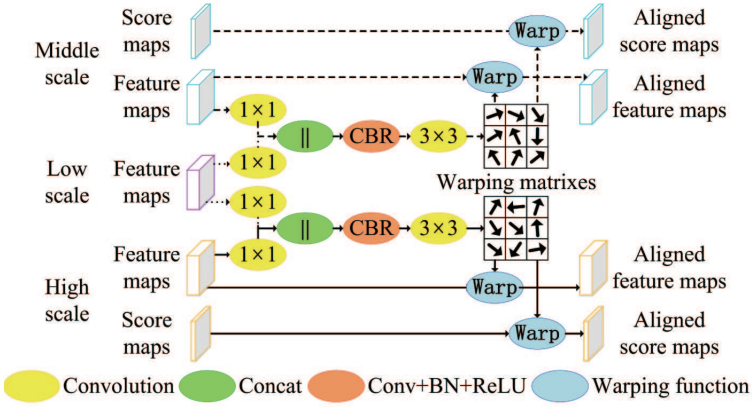

Figure 4 (Color online) The architecture of SAM. It is used to align both middle scale and high scale to low scale. Feature maps are first fed to a convolution function to align the channel dimension. Then they are concatenated and fed through a Conv-BN-ReLU block and a convolution function to obtain the warping matrixes, which helps to align the feature maps and score maps.

existing methods. Inspired by [51], we propose the SAM to align the feature maps and score maps on different scales. It first uses the feature maps on different scales to compute the warping matrixes as

$$X_{\mathrm{sa}}^{i}, Y_{\mathrm{sa}}^{i} = \delta_{\mathrm{sa}}^{i}(F^{1}, F^{i}), \tag{3}$$

where F^1 and F^i denote the feature maps on scale 1 and scale i, respectively; δ_{sa}^{i} is the scale alignment function, X_{sa}^{i},Y_{sa}^{i} are the warping matrixes with regard to the coordinates X and Y.

The detailed architecture of the scale alignment function is illustrated in Figure 4. The alignment is conducted between scale 1 and each larger scale independently. For the scale i, F^1 and F^i are independently fed to a 1×1 convolution to align the channel dimension. Then they are concatenated together and fed through a Conv-BN-ReLU block and a 3×3 convolution to obtain the warping matrixes.

After that, the warping matrixes are fed to the warping function for aligning the feature maps and score maps from larger scales to scale 1 as

$$F_{\mathrm{a}}^{i}, P_{\mathrm{a}}^{i} = \eta((X_{\mathrm{sa}}^{i}, Y_{\mathrm{sa}}^{i}), F^{i}, P^{i}), \tag{4}$$

where F^i and P^i are the feature maps and score maps on scale i, respectively; η denotes the warping function, F_{a}^{i} and P_{a}^{i} are the aligned feature maps and score maps on scale i, respectively.

It is noted that low-level features are used to align high-level features in [51], where all used features are from the input image. But in our proposed SAM, feature maps from higher scales are used to compute the warping matrixes together with feature maps from low scales, and the warping matrixes are used to align the feature maps and score maps from larger scales to the scale 1, which is different from [51].

3.3 Adaptive fusion module

When SSM searching has chosen the best model for each scale in our MFVNet, it is recommended to conduct decision fusion rather than feature fusion in order to avoid repeated computation. Benefiting from larger FOV, the high scale specializes in predicting large geospatial objects with complete outlines maintained, yet the low scale is good at capturing detailed information due to its high spatial resolution. Thus, the predicted score maps need to be fused adaptively. After the scale registration by the SAM, the aligned feature maps are concatenated and then used to compute the scale-specific weight maps as

$$F_{\mathrm{a}}^{\mathrm{all}} = \mathrm{concat}(F_{\mathrm{a}}^{1}, F_{\mathrm{a}}^{2}, \ldots, F_{\mathrm{a}}^{i}), \tag{5}$$

$$w^{i} = \phi_{\mathrm{af}}^{i}(F_{\mathrm{a}}^{\mathrm{all}}), \tag{6}$$

where F_{a}^{1}, F_{a}^{2}, F_{a}^{i} denote the aligned feature maps on scale 1, 2, i, respectively; ϕ_{af}^{i} is the adaptive fusion function for scale i; w^i is the scale-specific weight map for scale i.

The detailed architecture of the adaptive fusion function is illustrated in Figure 5. The concatenated feature maps are fed to two Conv-BN-ReLU blocks. Afterward, three 1×1 convolutions are used to compute the scale-specific weight maps for each scale separately.

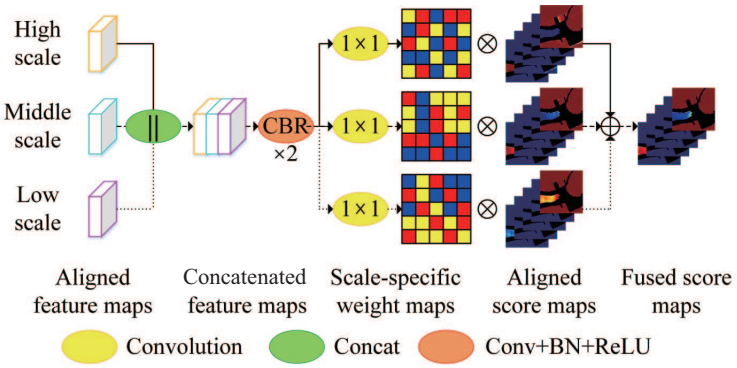

Figure 5 (Color online) The architecture of AFM. The feature maps are concatenated and fed to two Conv-BN-ReLU blocks. Then, three convolution functions are used to compute the scale-specific weight maps for each scale separately. Finally, the score maps on different scales are fused adaptively with the scale-specific weight maps to generate the final prediction result.

Finally, the score maps on different scales are fused adaptively with the help of scale-specific weight maps, which generate the final prediction result as

$$P^{\text{fuse}} = \sum_{i=1}^{S} (w^i \times P_{\text{a}}^i), \tag{7}$$

where P_{a}^i denotes the aligned score maps on scale i, P^{fuse} denotes the adaptively fused score maps, and S denotes the index of scales.

3.4 Loss function

For adequately utilizing the multiple FOVs in the RSIs, we build the framework of the adaptive fusion network with SSM, SAM, and AFM, which formulates the whole loss function as

$$\begin{aligned} L &= L_{\text{ce}}(P^{\text{fuse}}, M^1) \\ &= L_{\text{ce}}\left(\sum_{i=1}^{S} (\phi_{\text{af}}^i(F_{\text{a}}^{\text{all}}) \times \eta(\delta_{\text{sa}}^i(F^1, F^i), P^i)), M^1\right), \end{aligned} \tag{8}$$

where P^{fuse} denotes the adaptively fused score maps produced by (7), M^1 is the ground truth of image I^1, L_{ce} stands for the cross-entropy loss function.

It is noted that the proposed MFVNet can be end-to-end trained after the SSM searching process. Moreover, MFVNet can take all existing and future networks for semantic segmentation as its backbones for all scales with no regard to their architectures, which shows good generalization ability. The overall performance of MFVNet is based on the basic performances of SSMs from all scales and the cooperation of them. A more powerful SSM will definitely enhance the performance of MFVNet.

4 Experiments

In this section, we conduct experiments on three typical RSI semantic segmentation datasets, including GID [33], GF1WHU [34], and Potsdam, which are specially selected to demonstrate the effectiveness of the proposed MFVNet. The datasets and metrics used in the experiments are introduced first. Then we illustrate the necessity and effectiveness of SSM searching. After that, we ablate the important design modules (i.e., SSM, SAM, and AFM) of the proposed MFVNet. Finally, MFVNet is compared with previous state-of-the-art semantic segmentation methods and specially designed methods on three RSI semantic segmentation datasets.

4.1 Data descriptions

Three typical RSI datasets, including GID [33], GF1WHU [34], and Potsdam are utilized in this study.

The GID dataset [33] contains 150 RSIs with the size of 6800×7200. Five major categories are annotated: built-up, farmland, forest, meadow, and water. Areas not belonging to the above five categories

and clutter regions are labeled as background. We randomly choose 90 images for training, 30 images for validation, and 30 images for testing.

The GF1WHU dataset [34] is composed of 108 RSIs with an average size of about 16000×17000. It annotates three categories: cloud, cloud shadow, and clear-sky. We choose 84 images for training, 12 images for validation, and 12 images for testing randomly.

The Potsdam dataset[1)] is composed of 38 RSIs with the size of 6000×6000. Six classes are labeled, including impervious surfaces (i.e., Imp. sur.), car, tree, Low vegetation (i.e., Low veg.), building, and clutter. Twenty-four images are randomly chosen for training, 7 images for validation, and 7 images for testing.

4.2 Implementation details

We re-implement previous state-of-the-art semantic segmentation methods [21,22,24,26,35,36] and methods specially designed for RSI semantic segmentation [32,37–39]. It is noted that all models are trained following the hyper-parameters listed in their studies. Empirically, all the models are trained for 200 epochs on the GID and the Potsdam datasets, 50 epochs on the GF1WHU dataset, except for our MFVNet which is trained for 20 epochs, with their largest batch size on a single RTX3090 GPU card. We use the original structure in [21] for UNet, HRNet V2-W48 for HRNet and PSPNet with ResNet-101 backbone [52] for PSPNet. It is noted that the numbers of channels of their first and last convolutional layers are changed according to the input RS images (i.e., 4 channels) and the number of object categories. The data are augmented by random horizontal flipping, random vertical flipping, and random rotation of 90°. For some methods [37,38] that do not release the source codes, our re-implementations may be a little different from the original versions because of the missing detail in the studies. We also conduct limited modifications on the models without violating the original designs to pursue the best results possible. Some methods [32,39] are not designed for the datasets we used, so that their performances may not be as promising as those listed in their studies although we endeavor to achieve the best results possible.

We use the mean intersection over union (mIoU) as the main metric. Considering the class imbalance problems in datasets, we also display the frequency-weighted intersection over union (FWIoU) and the mean F_1 scores of all classes (mF_1).

4.3 Ablation study

As the images with multiple FOVs are generated via pyramid sampling, these images on different scales have disparate spatial contexts and resolutions. Empirically, we find that the best semantic segmentation model varies on different scales. And that is why we conduct the scale-specific model searching as in Subsection 3.2 on all scales. The model pool contains three prevalent semantic segmentation models (i.e., UNet [21], HRNet [26], and PSPNet [24]) considering their complementary abilities in capturing information.

Table 1 gives the results of SSM on the Potsdam dataset. It can be seen from the table that models vary a lot on the performances of different scales. It also verifies that the fixed architecture will definitely hurt the performance. The best models for low scale, middle scale, and high scale are PSPNet, UNet, and HRNet on the GID dataset; HRNet, HRNet, and UNet on the GF1WHU dataset; and PSPNet, PSPNet, and UNet on the Potsdam dataset. So it is data-driven.

To evaluate how each module in the proposed MFVNet influences the semantic segmentation performance, Table 2 demonstrates the quantitative ablation study results on the Potsdam dataset. For the baseline method, we use HRNet [26] trained on the low scale. When SSM is not used, we use HRNet for the training on all three scales since it performs better than the other two in the model pool. The use of multiple FOVs brings a performance gain of 0.6%. SSM leads to another performance gain of 0.4%. Moreover, AFM and SAM can both bring a performance gain of 1.1% and the best results can be acquired by the combination of the two. In general, with the help of SSM, AFM, and SAM, the mIoU score is improved by more than 2%, achieving state-of-the-art performance. The visualization of the ablation study is depicted in Figure 6.

1) https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-potsdam.

Table 1 Results of SSM on the Potsdam dataset

Scale	Method	Imp. sur.	Car	Tree	Low veg.	Building	Clutter	mIoU	FWIoU	mF_1
Low	UNet [21]	82.2	**82.9**	**73.9**	72.1	88.6	31.7	71.9	78.6	81.9
	HRNet [26]	83.0	81.3	72.7	72.5	90.0	36.2	72.6	79.2	82.7
	PSPNet [24]	**84.0**	80.5	74.7	**73.4**	**90.5**	**36.9**	**73.3**	**80.2**	**83.2**
Middle	UNet [21]	82.3	**81.5**	72.6	71.2	88.6	33.1	71.6	78.3	81.8
	HRNet [26]	81.4	81.0	68.6	69.6	88.6	35.1	70.7	77.5	81.0
	PSPNet [24]	**83.6**	79.4	**73.6**	**73.0**	**90.1**	**37.1**	**72.8**	**79.7**	**82.9**
High	UNet [21]	**80.9**	**80.5**	**71.5**	**69.5**	88.3	**31.4**	**70.4**	**77.2**	**80.9**
	HRNet [26]	80.4	79.7	67.6	67.8	**88.5**	28.3	68.7	75.9	79.5
	PSPNet [24]	79.6	72.4	68.1	68.1	88.2	30.1	67.7	75.6	79.1

Table 2 Performance contribution of each module in MFVNet on the Potsdam dataset

Method	Multiple FOV	SSM	SAM	AFM	mIoU	FWIoU	mF_1
Baseline					72.6	79.2	82.7
Baseline+MFOV	✓				73.2	80.0	83.1
Baseline+MFOV+SSM	✓	✓			73.6	80.8	83.6
Baseline+MFOV+SSM+SAM	✓	✓	✓		74.7	81.3	84.3
Baseline+MFOV+SSM+SAM+AFM (our MFVNet)	✓	✓	✓	✓	**74.8**	**81.5**	**84.3**

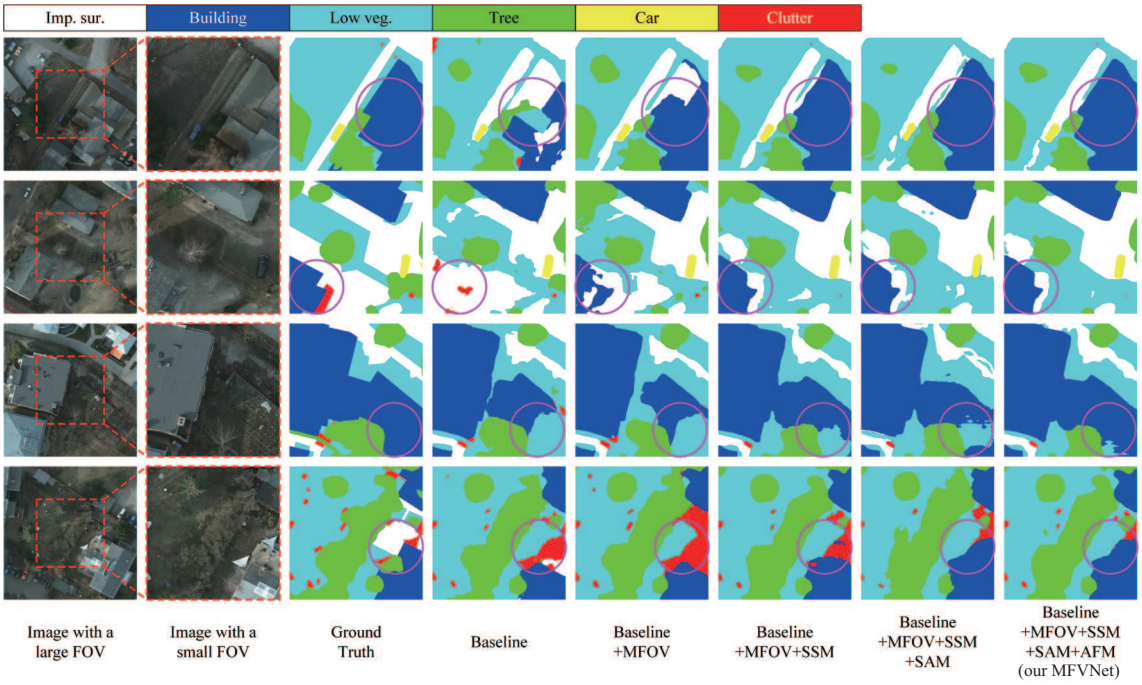

Figure 6 (Color online) Visualization of ablation study.

4.4 Comparison with existing methods

In this subsection, we evaluate the performance of the proposed framework against recent state-of-the-art methods. All methods reported here are re-implemented with the hyper-parameters listed in the studies to ensure fair comparisons, including previous state-of-the-art semantic segmentation models [21, 22, 24, 26, 35, 36] and methods designed for RSI semantic segmentation [32, 37–39].

The quantitative results are presented in Tables 3–5, our proposed MFVNet leads to a significant gain over existing methods. It can be observed that the mIoU scores of the proposed MFVNet are about 0.6%, 1.8%, and 1.3% higher than those of the state-of-the-art methods for the GID, GF1WFV, and Potsdam datasets, respectively.

Furthermore, most of the existing methods (e.g., PSPNet [24], DANet [35]) fail to work well on all three RSI datasets, while our MFVNet performs well on these three datasets since MFVNet can exploit

Table 3 Comparison with existing methods on the GID dataset

Method	Build-up	Farmland	Forest	Meadow	Water	mIoU	FWIoU	mF_1
UNet [21]	94.4	97.1	90.6	90.1	97.3	93.9	95.9	96.8
PSPNet [24]	94.2	96.6	**95.1**	92.3	97.1	95.0	96.0	97.4
Deeplabv3+ [22]	92.4	96.3	94.3	92.7	97.2	94.6	95.6	97.2
HRNet [26]	93.7	96.7	93.2	93.5	97.4	94.9	96.0	97.4
DANet [35]	93.3	96.6	88.8	85.9	97.3	92.4	95.1	96.0
CCNet [36]	92.2	95.7	90.8	87.9	94.3	92.2	94.2	95.9
TS-MTA [37]	94.1	96.4	92.7	88.3	96.3	93.8	95.1	96.8
WiCNet [38]	93.3	96.5	94.2	93.3	97.0	94.9	95.8	97.3
GLNet [39]	83.9	92.2	69.1	88.0	93.3	85.3	90.0	91.8
FCTL [32]	91.5	96.1	89.0	87.4	96.6	92.1	94.6	95.9
MFVNet (ours)	**95.1**	**97.4**	93.8	**93.7**	**97.9**	**95.6**	**96.7**	**97.7**

Table 4 Comparison with existing methods on the GF1WHU dataset

Method	Cloud	Shadow	Clear	mIoU	FWIoU	mF_1
UNet [21]	91.2	56.8	86.2	78.1	87.2	87.0
PSPNet [24]	91.2	53.8	86.0	77.0	87.2	87.0
Deeplabv3+ [22]	86.9	52.9	92.3	77.4	87.8	86.0
HRNet [26]	92.5	56.8	87.3	78.9	88.3	87.2
DANet [35]	85.7	53.0	91.3	76.7	86.8	85.7
CCNet [36]	85.2	53.1	90.5	76.3	86.2	85.5
TS-MTA [37]	90.8	51.3	**92.6**	78.3	89.8	86.4
WiCNet [38]	90.6	52.5	92.0	78.4	89.5	86.6
GLNet [39]	**94.1**	39.9	86.1	73.4	**90.9**	82.2
FCTL [32]	91.7	27.6	89.6	69.6	88.7	77.8
MFVNet (ours)	92.5	**58.5**	91.1	**80.7**	90.2	**88.4**

Table 5 Comparison with existing methods on the Potsdam dataset

Method	Imp. sur.	Car	Tree	Low veg.	Building	Clutter	mIoU	FWIoU	mF_1
UNet [21]	82.2	**82.9**	73.9	72.1	88.6	31.7	71.9	78.6	82.3
PSPNet [24]	84.0	80.5	74.7	73.4	90.5	36.9	73.3	80.2	83.2
Deeplabv3+ [22]	83.4	80.1	74.2	73.1	90.6	37.4	73.1	79.9	83.2
HRNet [26]	83.0	81.3	72.7	72.5	90.0	36.2	72.6	79.2	82.7
DANet [35]	84.1	80.5	75.1	73.7	90.7	38.2	73.7	80.5	83.6
CCNet [36]	83.6	81.2	74.4	72.5	90.2	32.3	72.4	79.6	82.3
TS-MTA [37]	83.5	82.0	73.8	72.3	91.3	38.2	73.5	80.0	83.4
WiCNet [38]	82.4	80.6	74.1	72.3	89.9	36.8	72.7	79.3	82.8
GLNet [39]	68.7	54.5	67.3	60.5	81.3	21.9	59.0	66.6	72.2
FCTL [32]	72.0	52.2	66.3	61.8	81.0	28.5	60.3	69.3	73.7
MFVNet (ours)	**85.2**	82.2	**76.0**	**74.9**	**91.4**	**39.2**	**74.8**	**81.5**	**84.3**

the multiple FOVs contained in the RSIs and combine the strengths of different scales by adaptive fusion, which owns good generalization ability. It is noted that GLNet [39] and FCTL [32] are supposed to handle the RSI semantic segmentation with a large FOV, but they perform not very well on these three datasets with a much longer time consuming than other methods. The reasons may require further exploration.

Some qualitative results are presented in Figures 7–9, which also demonstrate that our proposed MFVNet is superior to the baseline method (i.e., HRNet [26]) in a large margin. The baseline method fails to predict precisely on some locations such as the cloudy area in Figure 8 and the building area in Figure 9. However, our proposed MFVNet can overcome the limited information with the help of multiple FOVs. MFVNet combines the advantage of all scales and fuses them adaptively for better results.

With the aid of SSM, SAM, and AFM, our proposed MFVNet makes good use of multiple FOVs in remote sensing images. As shown in the given results in the paper, it can obviously improve the overall performance. More specifically, the model on the lower scale can only see limited and incomplete information, whereas the model on the higher scale tends to enlarge the accessible spatial contexts but

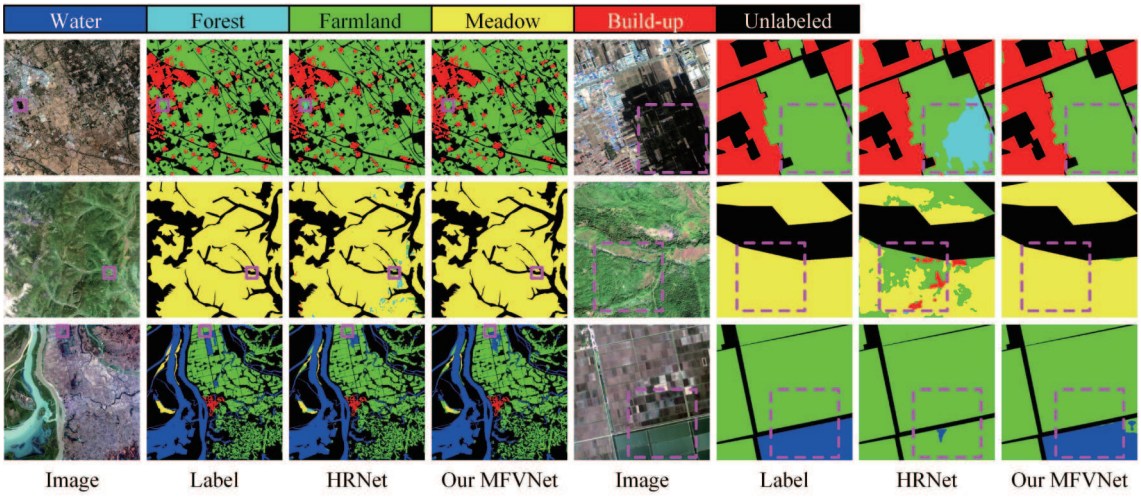

Figure 7 (Color online) Visualization results on the GID dataset. Pink boxes demonstrate the effectiveness of MFVNet.

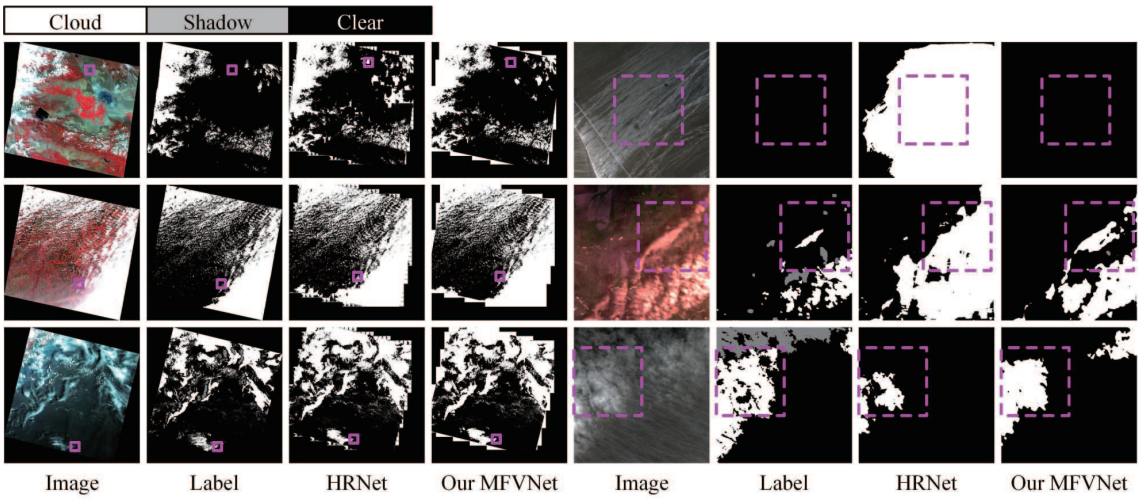

Figure 8 (Color online) Visualization results on the GF1WHU dataset. Pink boxes demonstrate the effectiveness of MFVNet.

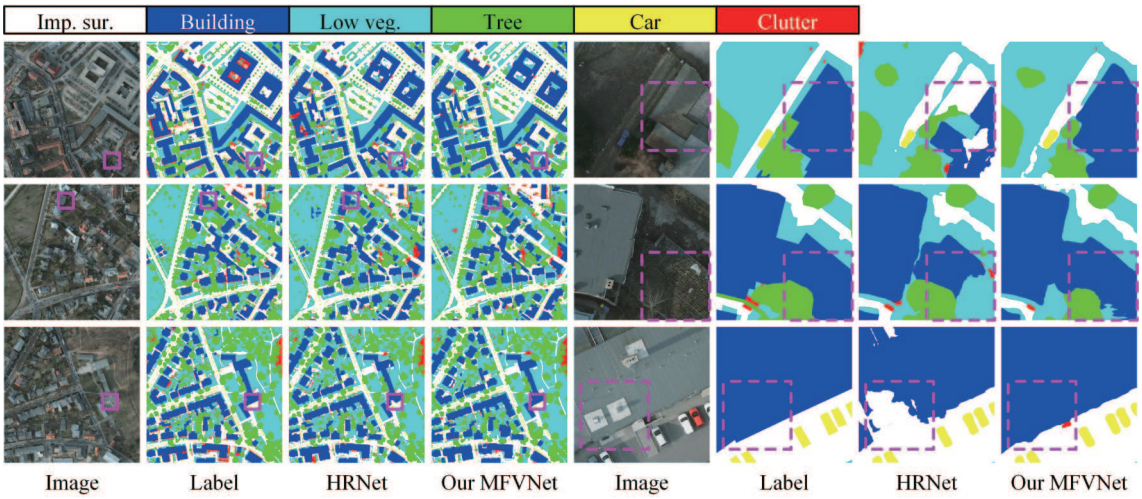

Figure 9 (Color online) Visualization results on the Potsdam dataset. Pink boxes demonstrate the effectiveness of MFVNet.

inevitably brings the potential noises (e.g., the ignored small objects) to the whole MFVNet. Even though our proposed MFVNet aims to adaptively fuse the models with multiple FOVs and pursue the best overall performances, it is quite challenging for a unified model to achieve the top in all categories due to the scale variation of geospatial objects in the mission of remote sensing image semantic segmentation. For instance, the special designs of the U-shape and the skip-connection in UNet are beneficial to the segmentation of small objects (e.g., the car category in the Potsdam dataset), while the performance on large objects is extremely degenerated. In general, our proposed MFVNet inherits the advantage of UNet on small objects and handles large objects well meanwhile. It is for a similar reason that MFVNet achieves suboptimal performance on some large objects (e.g., the forest category in the GID dataset). After all, given the best overall performance, the performance decline of our MFVNet in some categories (e.g., the car category in the Potsdam dataset or the forest category in the GID dataset) compared with the different optimal networks is acceptable.

5 Conclusion and discussion

In this paper, we propose a deep adaptive fusion network to make full use of multiple FOVs in the RSIs. We enlarge the accessible spatial contexts by pyramid sampling the original RSIs, which form images with multiple FOVs due to the different pyramid sampling rates. Based on the differences among scales, we perform SSM searching for each scale, to choose the best model for a specific scale. The images on different scales are independently fed to the corresponding SSMs and produce scale-specific feature maps and score maps, which need to be aligned first because of spatial misregistration caused by resampling. Hence, the feature maps and score maps from high scales are aligned to low scales by the SAMs. Then these feature maps are concatenated and used to compute the scale-specific weight maps. Finally, the score maps on different scales are fused adaptively with the help of weight maps in the AFMs, which generates the final prediction result.

Although the proposed MFVNet can obtain promising results, the SSM searching which is used to select the best model for each scale costs much time currently, since we have to train each of the candidate models for all scales thoroughly to compare their performances fairly. Furthermore, the model pool for choosing only contains three models (i.e., UNet [21], HRNet [26], and PSPNet [24]) because of the time costs. We hope the development of network architecture searching [53, 54] will help to overcome this problem. And, only simple cross-entropy loss is studied in this paper, we believe that additional constraints on the SAM and AFM will further enhance the performances of spatial registration and the adaptive fusion, which can help to produce a better-fused prediction. It is noted that we do not conduct time-consuming hyper-parameter (e.g., learning rate) tuning in this work, yet we focus on the architecture design and fusion problems. We believe that hyper-parameter tuning can be solved with automatic hyper-parameter searching [55] for further performance improvements, which will also be studied in our future work.

Acknowledgements This work was supported in part by State Key Program of the National Natural Science Foundation of China (Grant No. 42030102), Foundation for Innovative Research Groups of the Natural Science Foundation of Hubei Province (Grant No. 2020CFA003), National Natural Science Foundation of China (Grant No. 41971284), Fundamental Research Funds for the Central Universities (Grant No. 2042022kf1201), and Special Fund of Hubei Luojia Laboratory.

References

1 He Q, Sun X, Yan Z, et al. Multi-object tracking in satellite videos with graph-based multitask modeling. IEEE Trans Geosci Remote Sens, 2022, 60: 1–13
2 He Q, Sun X, Diao W, et al. Transformer-induced graph reasoning for multimodal semantic segmentation in remote sensing. ISPRS J Photogrammetry Remote Sens, 2022, 193: 90–103
3 Sun X, Wang P, Yan Z, et al. FAIR1M: a benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. ISPRS J Photogrammetry Remote Sens, 2022, 184: 116–130
4 Fu S L, Xu F, Jin Y-Q. Reciprocal translation between SAR and optical remote sensing images with cascaded-residual adversarial networks. Sci China Inf Sci, 2021, 64: 122301
5 Gu Y F, Liu T Z, Gao G M, et al. Multimodal hyperspectral remote sensing: an overview and perspective. Sci China Inf Sci, 2021, 64: 121301
6 Mei J, Li R J, Gao W, et al. CoANet: connectivity attention network for road extraction from satellite imagery. IEEE Trans Image Process, 2021, 30: 8540–8552
7 Rashkovetsky D, Mauracher F, Langer M, et al. Wildfire detection from multisensor satellite imagery using deep semantic segmentation. IEEE J Sel Top Appl Earth Observations Remote Sens, 2021, 14: 7001–7016
8 Ding L, Tang H, Liu Y, et al. Adversarial shape learning for building extraction in VHR remote sensing images. IEEE Trans Image Process, 2022, 31: 678–690
9 Li Y, Chen W, Zhang Y, et al. Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning. Remote Sens Environ, 2020, 250: 112045

10 Li Y, Shi T, Zhang Y, et al. Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation. ISPRS J Photogrammetry Remote Sens, 2021, 175: 20–33
11 Li Y, Zhou Y, Zhang Y, et al. DKDFN: domain knowledge-guided deep collaborative fusion network for multimodal unitemporal remote sensing land cover classification. ISPRS J Photogrammetry Remote Sens, 2022, 186: 170–189
12 Workman S, Rafique M U, Blanton H, et al. Revisiting near/remote sensing with geospatial attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022
13 Peng D, Bruzzone L, Zhang Y, et al. SemiCDNet: a semisupervised convolutional neural network for change detection in high resolution remote-sensing images. IEEE Trans Geosci Remote Sens, 2021, 59: 5891–5906
14 Zhu Q, Guo X, Deng W, et al. Land-Use/Land-Cover change detection based on a Siamese global learning framework for high spatial resolution remote sensing imagery. ISPRS J Photogrammetry Remote Sens, 2022, 184: 63–78
15 Datcu M, Seidel K. Human-centered concepts for exploration and understanding of Earth observation images. IEEE Trans Geosci Remote Sens, 2005, 43: 601–609
16 Lillesand T, Kiefer R W, Chipman J. Remote Sensing and Image Interpretation. Hoboken: John Wiley & Sons, 2015
17 Haar R, Bart M T, Florack L. A multiscale geometric model of human vision. In: The Perception of Visual Information. New York: Springer, 1993. 73–114
18 Romeny B M H. Front-End Vision and Multi-Scale Image Analysis: Multi-Scale Computer Vision Theory and Applications, Written in Mathematica. Berlin: Springer Science & Business Media, 2008
19 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 3431–3440
20 Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 2481–2495
21 Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention, 2015. 234–241
22 Chen L, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of European Conference on Computer Vision, 2018. 801–818
23 Lin G, Milan A, Shen C, et al. RefineNet: multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 1925–1934
24 Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 2881–2890
25 Xiao T, Liu Y, Zhou B, et al. Unified perceptual parsing for scene understanding. In: Proceedings of European Conference on Computer Vision, 2018. 418–434
26 Wang J, Sun K, Cheng T, et al. Deep high-resolution representation learning for visual recognition. IEEE Trans Pattern Anal Mach Intell, 2021, 43: 3349–3364
27 Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE International Conference on Computer Vision, 2021
28 Zhang Z J, Pang Y W. CGNet: cross-guidance network for semantic segmentation. Sci China Inf Sci, 2020, 63: 120104
29 Ma S, Pang Y W, Pan J, et al. Preserving details in semantics-aware context for scene parsing. Sci China Inf Sci, 2020, 63: 120106
30 Feng J P, Wang X G, Liu W Y. Deep graph cut network for weakly-supervised semantic segmentation. Sci China Inf Sci, 2021, 64: 130105
31 He N J, Fang L Y, Plaza A. Hybrid first and second order attention Unet for building segmentation in remote sensing images. Sci China Inf Sci, 2020, 63: 140305
32 Li Q, Yang W, Liu W, et al. From contexts to locality: ultra-high resolution image segmentation via locality-aware contextual correlation. In: Proceedings of the IEEE International Conference on Computer Vision, 2021. 7252–7261
33 Tong X Y, Xia G S, Lu Q, et al. Land-cover classification with high-resolution remote sensing images using transferable deep models. Remote Sens Environ, 2020, 237: 111322
34 Li Z, Shen H, Li H, et al. Multi-feature combined cloud and cloud shadow detection in GaoFen-1 wide field of view imagery. Remote Sens Environ, 2017, 191: 342–358
35 Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 3146–3154
36 Huang Z, Wang X, Huang L, et al. CCNet: criss-cross attention for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, 2019. 603–612
37 Ding L, Zhang J, Bruzzone L. Semantic segmentation of large-size VHR remote sensing images using a two-stage multiscale training architecture. IEEE Trans Geosci Remote Sens, 2020, 58: 5367–5376
38 Ding L, Lin D, Lin S, et al. Looking outside the window: wide-context transformer for the semantic segmentation of high-resolution remote sensing images. IEEE Trans Geosci Remote Sens, 2022, 60: 1–13
39 Chen W, Jiang Z, Wang Z, et al. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 8924–8933
40 Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. In: Proceedings of International Conference on Learning Representations, 2016
41 Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks. In: Proceedings of Advances in Neural Information Processing Systems, 2014. 3104–3112
42 Devlin J, Chang M, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. 2018. ArXiv:1810.04805
43 Yuan Y, Huang L, Guo J, et al. OCNet: object context network for scene parsing. 2021. ArXiv:1809.00916
44 Li D, Hu J, Wang C, et al. Involution: inverting the inherence of convolution for visual recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021. 12321–12330
45 Woo S, Park J, Lee J, et al. CBAM: convolutional block attention module. In: Proceedings of European Conference on Computer Vision, 2018. 3–19
46 Zhao H, Zhang Y, Liu S, et al. PSANet: point-wise spatial attention network for scene parsing. In: Proceedings of European Conference on Computer Vision, 2018. 267–283
47 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale. In: Proceedings of International Conference on Learning Representations, 2021
48 Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention. In: Pro-

ceedings of International Conference on Machine Learning, 2021. 10347–10357
49 Zheng S, Lu J, Zhao H, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021. 6881–6890
50 Cheng H K, Chung J, Tai Y, et al. CascadePSP: toward class-agnostic and very high-resolution segmentation via global and local refinement. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 8890–8899
51 Li X, You A, Zhu Z, et al. Semantic flow for fast and accurate scene parsing. In: Proceedings of European Conference on Computer Vision, 2020. 775–793
52 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770–778
53 Liu C, Chen L, Schroff F, et al. Auto-DeepLab: hierarchical neural architecture search for semantic image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 82–92
54 Zhang X, Xu H, Mo H, et al. DCNAs: densely connected neural architecture search for semantic image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021. 13956–13967
55 He X, Zhao K, Chu X. AutoML: a survey of the state-of-the-art. Knowledge-Based Syst, 2021, 212: 106622