A Gather-to-Guide Network for Remote Sensing Semantic Segmentation of RGB and Auxiliary Image

Xianwei Zheng, Xiujie Wu, Linxi Huan, Wei He, *Member, IEEE*, and Hongyan Zhang, *Senior Member, IEEE*

***Abstract*— Convolutional neural network (CNN)-based feature fusion of RGB and auxiliary remote sensing data is known to enable improved semantic segmentation. However, such fusion is challengeable because of the substantial variance in data characteristics and quality (e.g., data uncertainties and misalignment) between two modality data. In this article, we propose a unified gather-to-guide network (G2GNet) for remote sensing semantic segmentation of RGB and auxiliary data. The key aspect of the proposed architecture is a novel gather-to-guide module (G2GM) that consists of a feature gatherer and a feature guider. The feature gatherer generates a set of cross-modal descriptors by absorbing the complementary merits of RGB and auxiliary modality data. The feature guider calibrates the RGB feature response by using the channel-wise guide weights extracted from the cross-modal descriptors. In this way, the G2GM can perform RGB feature calibration with different modality data in a gather-to-guide fashion, thus preserving the informative features while suppressing redundant and noisy information. Extensive experiments conducted on two benchmark datasets show that the proposed G2GNet is robust to data uncertainties while also improving the semantic segmentation performance of RGB and auxiliary remote sensing data.**

***Index Terms*— Deep learning, remote sensing, semantic segmentation.**

I. Introduction

SEMANTIC segmentation of high-resolution aerial/satellite images is a fundamental task in remote sensing, in which the aim is to classify each pixel in a given image with a semantic category. The applications of semantic segmentation range from urban planning, change detection, and landcover classification to urban 3-D semantic modeling [1]–[5]. In recent years, the progress of deep learning in RGB scene image parsing has significantly promoted the development of remote sensing image semantic segmentation [6], [7]. Unfortunately, the performance from single RGB data is limited due to the inefficient feature exploration [8].

Beyond RGB, auxiliary remote sensing data are also widely used for semantic segmentation, such as synthetic aperture radar (SAR) images [9] and digital surface models (DSMs) [10]. As different modalities of RGB, these auxiliary remote sensing data measure the specific properties of the same geospatial object and provide different insights into the comprehensive learning of a semantic object [11], [12]. Therefore, more interests are paid to utilize the complementary information from the RGB and auxiliary remote sensing data to improve the performance of semantic segmentation [10], [13]. In this article, we focus on semantic segmentation of RGB and DSM (can also be extended to IRRG and NDVI), which has been widely investigated as the development of convolutional neural networks (CNNs) [10], [14]–[16]. Although progress has been witnessed in the past years, some issues of semantic segmentation of RGB and auxiliary remote sensing data are still worth studying.

The first issue is the notable variations between RGB and auxiliary data. It is reported that RGB and auxiliary data share different physical and numeric characteristics [10]. Therefore, it is necessary to develop an efficient strategy to derive informative representation from various types of data for semantic segmentation. One simple way is to concatenate the RGB and auxiliary data into a multichannel tensor and feed the tensor into a one-stream vision-based network for training [17]. The different data can thus be directly fused by local convolution operations across channels, but this scheme ignores the specificities of RGB and auxiliary data. Consequently, the two-stream network architectures that process RGB and auxiliary data in two parallel CNNs have become more advisable than the one-stream models [18]–[20]. The fusion based on two-stream CNNs is generally performed either by assembling the predictions of two modalities into a unified final output or by merging hierarchical features from two separate encoders to form enhanced representations for the shared decoder [16]. The former prediction fusion solutions usually encounter difficulties in fully exploiting the complementary information of two modalities since the two CNN branches have no information interaction during the forward and backward propagations. Meanwhile, the feature fusion schemes tend to equally treat RGB and

Manuscript received May 14, 2021; revised June 26, 2021; accepted July 28, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2018Y-FB0505401 and in part by the National Natural Science Foundation of China Project under Grant 42071370 and Grant 41871361. *(Corresponding author: Wei He.)*

Xianwei Zheng, Xiujie Wu, Linxi Huan, and Hongyan Zhang are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan 430079, China (e-mail: zhengxw@whu.edu.cn; xiujiewu@whu.edu.cn; whu_hlx@whu.edu.cn; zhanghongyan@whu.edu.cn).

Wei He is with the RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan (e-mail: wei.he@riken.jp).

Digital Object Identifier 10.1109/TGRS.2021.3103517

auxiliary data, which leads to issues of redundant information learning, as the fused cross-modal features are often directly fed into the subsequent layer during the forward propagation.

Another issue of semantic segmentation in fusing RGB and auxiliary data lies in data uncertainties. Although the combination of RGB and auxiliary data can bring benefits to remote sensing tasks, the quality and diversity of the data can bring troubles for efficient information exploration. The Earth observation data in different modalities are usually acquired by various platforms and sensors. For example, RGB signals are measured using optical sensors, whereas LiDAR-derived DSMs are originally captured by laser scanners. On the one hand, those measurements may contain varying degrees of noises and/or anomalies, due to the physical characteristics of sensors and the complexity of the geographic environment [21], [22]. On the other hand, the image pixels of the data acquired by multiple sensors are not always precisely aligned, even within a careful co-registration process. Low-quality measurements can introduce exceptional features into CNN models and impede the learning of informative features, causing a misunderstanding of geospatial objects. Therefore, when using auxiliary data to complement RGB for improving the performance of semantic segmentation, the fusion scheme should be able to diminish the deep model's sensitivity to the unstable quality of auxiliary data.

Based on the aforementioned two issues, we are motivated to develop a comprehensive deep neural network with a specially designed gather-to-guide module (G2GM) for semantic segmentation of RGB and auxiliary data. We term our network as gather-to-guide network (G2GNet), which gathers cross-modal information to guide the calibration of RGB response, rather than directly fusing two types of features to generate cross-modal inputs for the subsequent layers of the network. The G2GNet provides a mechanism to incorporate the auxiliary information only for RGB representation refinement, thus diminishing the model's sensitivity to the low-quality measurements of additional data. To this end, we insert our G2GM at the end of the two encoders. The G2GM consists of two components: a feature gatherer and a feature guider. The feature gatherer harnesses the two types of information to generate a cross-modal guidance for subsequent RGB feature calibration. First, the global information from each modality is encoded via a self-adaptive semantic region pooling operation, in which the modality-specific features are distilled with less noise inference. The encoded cross-modal global information is then combined into a group of cross-modal descriptors via a gated fusion mechanism. The feature guider further encapsulates the cross-modal representation into a channel-wise weight vector, which serves as the cue to guide the calibration of RGB representations for improved segmentation performance.

The main contributions of this work are summarized as follows.

1) We develop a general G2GNet for remote sensing semantic segmentation, which can selectively absorb the complementary merits of RGB and auxiliary data to improve the semantic segmentation performance.
2) A novel G2GM is proposed to adaptively gather the informative features from the inputs RGB and auxiliary data by a self-adaptive attention mechanism and then fuse the gathered features as a set of global descriptors to guide the calibration of RGB responses for enhanced feature representation.
3) Extensive experiments demonstrate that the G2GNet can effectively fuse RGB and auxiliary data to boost segmentation performance and achieve competitive performance on two benchmark datasets, i.e., ISPRS Vaihigen 2-D and Potsdam 2-D.

II. Related Work

A. RGB-Based Semantic Segmentation

Semantic segmentation has been a long-standing image analysis task across multiple research fields. In the earlier stages, low-level visual cues had been extensively studied for semantic segmentation. Recently, deep learning has shown its superiority in semantic segmentation, as it can learn high-level semantic information that is hard to obtain with the traditional methods.

With the advent of the first end-to-end fully convolutional neural network (FCN) [23], the accuracy of urban natural scene semantic segmentation has been greatly improved. FCN allows images with arbitrary sizes to be fed as input and generates pixel-wise semantic predictions for the whole image. Subsequently, the encoder–decoder architecture was introduced to semantic segmentation for more effective preservation of object spatial details. The decoder structure progressively recovers the spatial dimension of objects, making the FCN upsampling process learnable. Segnet [24] and U-Net [25] also studied the issue of spatial detail recovery by equipping the encoder–decoder structure with skip connections to reuse the detail information from low-level convolutional stages. Following this line, the U-Net++ [26] and Refinenet [27] quickly improved the skip connection with more advanced transfer schemes. To strengthen feature representations for objects with diverse sizes, other researchers also focused on incorporating multiscale processing into FCN models [28], [29]. For example, Zhao *et al.* [30] utilized the image pyramid to parallelly learn multiscale features with different input scales. A more popular way is to deploy an atrous spatial pyramid pooling (ASPP) module as in the case of DeepLab family [31].

The advances in deep network design and natural scene parsing have immediately inspired progress in remote sensing tasks. Vakalopoulou *et al.* [32] built an automated building detector for very high-resolution remote sensing data based on CNN. Längkvist *et al.* [6] developed a CNN-based approach for the pixel-wise classification of multiple objects in satellite images. For full-resolution labeling of aerial images, Sherrah [7] adopted the FCN architecture that did not require a downsampling operation to obviate the deconvolution or interpolation. Zhao and Du [33] presented a multiscale convolutional neural network (MCNN) to learn spatial-related deep features for hyperspectral imagery classification. Maggiori *et al.* [34] designed a multiscale neuron module to alleviate the common tradeoff between recognition

and precise localization for satellite imagery segmentation. Liu *et al.* [35] proposed a self-cascaded network to address the challenges of human-made object confusion and fine-structured object intricacy. Sun *et al.* [36] adopted a residual encoder–decoder architecture to mitigate the problem of insufficient learning. Zheng *et al.* [37] designed a foreground-aware relation network to alleviate the problems of large intraclass variance in background and foreground imbalance in high spatial resolution remote sensing imagery.

B. Semantic Segmentation of RGB and Auxiliary Data

Although the vision-based deep networks have attained promising segmentation results for RGB scenes, the ambiguity of visual cues continues to restrict the network performance. Therefore, numerous researchers have become interested in unifying information from other modality data to improve RGB segmentation accuracy.

In the computer vision community, a revealing experiment in FCN [23] proved that depth information is useful in promoting RGB segmentation performance. With the rapid development of consumer-level depth sensors, e.g., Microsoft Kinect [38], RealSense [39], considerable effort was subsequently devoted to RGB-D semantic segmentation. To facilitate an immediate use of vision-based FCNs, Su and Wang [40] directly treated depth data as an additional channel, which is fed together with RGB data into a single FCN for semantic prediction. Other researchers utilize depth data as a guidance to tailor 2-D CNN into 2.5-D behaviors, to explicitly incorporate geometry information into CNN [41]. To better identify the differences between two modalities, a more popular approach for RGB-D segmentation is to deploy a two-stream network architecture that can parallelize the convolutional feature extraction for each modality with two FCNs. For instance, FuseNet [42] and RedNet [43] equipped their encoders with two FCN branches for each input modality and extended cross-modal fusion (CMF) from the final layer to the multilevel convolutional stages. Residual learning was further used in RDFNet [44], with the aim of fully preserving the modality-specific characteristics during fusion.

As the Earth observation data are rich in different modalities [45], [46], multimodal semantic segmentation of RGB and auxiliary data [18], [47] has also been investigated recently with the help of CNNs. An early work of [14] showed that combining multimodal features with CNNs is essential in labeling some specific categories and can significantly boost the overall performance of remote sensing classification. In [48], the CNN-based analytic outputs of multispectral images and spaceborne remote sensing videos are fused for semantic scene interpretation. To enable an immediate feature-level fusion, Paisitkriangkrai *et al.* [15] trained a CNN with a five-channel input data concatenated by orthophoto, DSM, and normalized DSM (nDSM) images. Similarly, Volpi and Tuia [17] used all spectral channels (near infrared, R, G, and B) and DSM as the input for their network. Alternatively, Liu *et al.* [19] proposed a decision-level fusion scheme that first obtain two probabilistic results from an FCN trained on VHR optical imagery and a linear classifier performed on handcrafted LiDAR features and then combine the two results with high-order CRFs for dense semantic labeling. To counter the blurry effect of boundaries in semantic segmentation, Marmanis *et al.* [20] explicitly incorporated class boundaries detected from DSM data and color images into semantic predictions for the refined segmentation results. Audebert *et al.* [10] applied two SegNets to orthophoto image and DSM/nDSM/NDVI (normalized difference vegetation index) data and investigated the early fusion and the late fusion of convolutional features of two modalities. Similarly, Piramanayagam *et al.* [49] studied the early fusion, the composite fusion, and the late fusion. For remote sensing classification beyond RGB data, a newly proposed X-ModalNet [9] exploited the mechanism of semisupervised transfer learning with cross-modality data in remote sensing.

In summary, most of the existing methods focus on merging features extracted from RGB and auxiliary data into cross-modal representation, which is utilized as the input for the following classifier. However, we consider that overuse of auxiliary data could be a risk factor for segmentation accuracy because low-quality measurements from additional data could disturb network learning. For example, DSM data generated by oblique photogrammetry usually contain missing data and noisy height values. In this work, we address these problems by performing CMF in a gather-to-guide fashion, which is capable of filtering exceptional features while preserving the specificity of different input modalities.

C. Attention Mechanism

Attention mechanism has been widely adopted as an effective learning tool to bias the allocation of computational resources of CNN models to emphasize the most informative features at the regional or global scale. Common deep networks utilize the convolution operation to fuse spatial and channel-wise information for extracting informative features, but they are limited to the local nature of the convolution kernel. To overcome this limitation, a variety of attention modules have been developed for global context information capturing in semantic segmentation. To capture the spatial correlations between features, a nonlocal operation has been devised [50] to calibrate the response at a position by weighted averaging the features at all positions. SENet [51] proposed a channel-attention mechanism for adjusting the importance of different feature maps by explicitly modeling the dependency between channels. To achieve adaptive receptive fields of neurons, SKNet [52] fused multiple kernels with channel-wise attention. DANet [53] inserted two types of attention modules, i.e., a position attention and a channel attention, into a dilated FCN to model the semantic interdependencies in spatial and channel dimensions. Moreover, as remote sensing images usually cover a large spatial extent, the long-range spatial relations between entities are more prominent than natural scene images. Recent works [54], [55] also introduced spatial and channel-attention modules to capture the global relationships between any two positions or channel maps for semantic segmentation of aerial images.

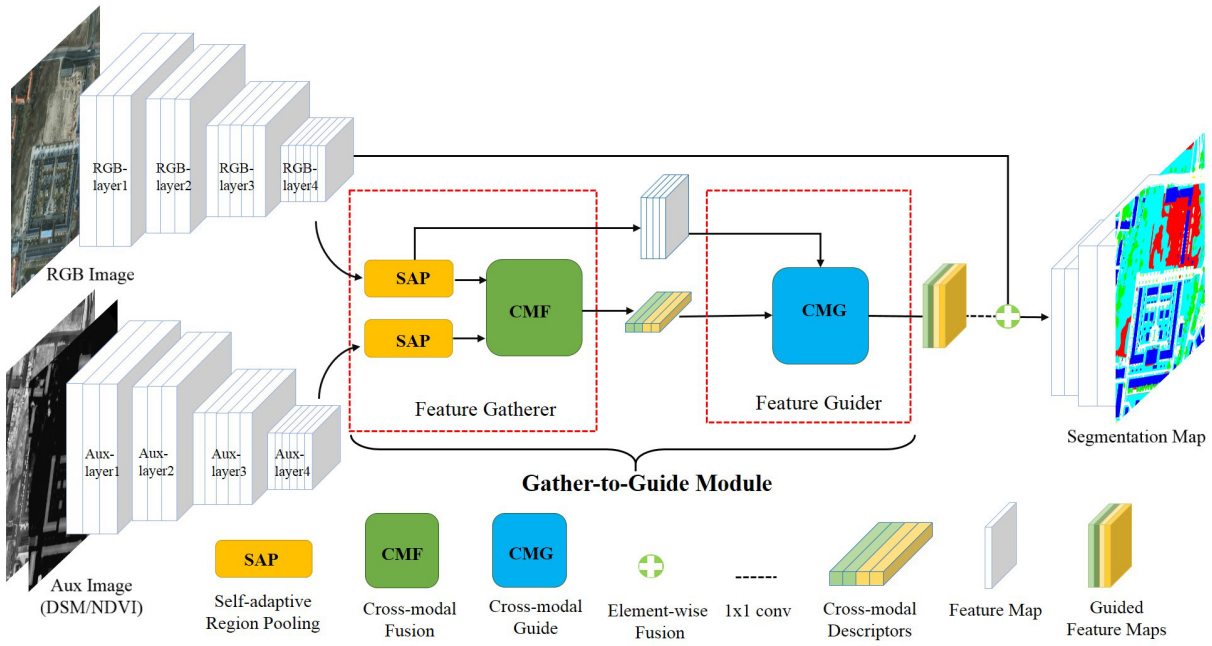


Fig. 1. Overview of the proposed G2GNet architecture.

III. Methodology

In this section, we elaborate on the proposed G2GNet for improved semantic segmentation of RGB and auxiliary data, starting with an overview of the network.

A. Overview

The complete framework of the proposed G2GNet is shown in Fig. 1. The G2GNet is deployed as a two-stream encoder–decoder network architecture for full-resolution segmentation. The encoder of each stream consists of four residual layers. The G2GM, as the center building block of our network architecture, is appended on top of the two encoders to unify information from the two streams and calibrate the RGB feature response. The inputs of the G2GNet are the paired images consist of an RGB and an auxiliary modality data. The output is a pixel-wise semantic segmentation map.

In G2GNet, the two encoders separately extract hierarchical features from each input modality and gradually generate their high-level semantic feature maps. First, the feature maps generated by RGB-layer 4 and Aux-layer 4 are fed into the feature gatherer of the G2GM. In the feature gatherer, the self-adaptive region pooling (SAP) squeezes the feature maps of a given modality into compact global descriptors (also generates dimension reduced feature maps with its internal operation), and the CMF block aggregates the global descriptors of two modalities into a set of cross-modal global descriptors, which absorbs the complementary merits of the two modalities. Second, the feature guider of the G2GM regards the cross-modal global descriptors and the reduced feature maps of RGB-layer 4 as inputs. In detail, the cross-modal guider (CMG) derives per-channel weights from the cross-modal global descriptors and applies them to the reduced feature maps of RGB-layer 4 to guide RGB feature response calibration. Finally, the guided feature maps are fused with the feature maps of RGB-layer 4, and the fused features are propagated to the segmentation decoder for resolution reconstruction and semantic prediction.

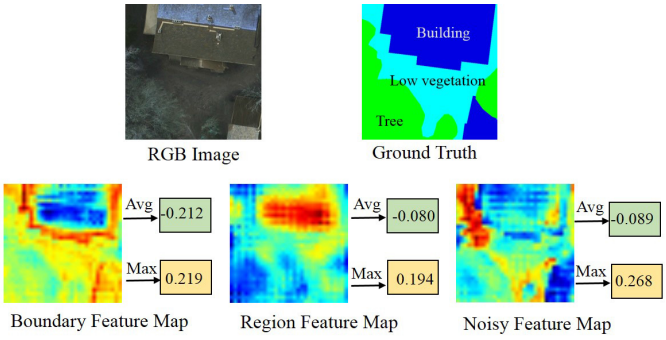


Fig. 2. Different responses encoded by global max-pooling and global average-pooling operations. Values in rectangles are weights derived by different pooling operations on different feature maps.

B. Feature Gatherer

Features learned by deep CNNs contain not only representative cues but also redundant information and noises that may degrade the performance of semantic segmentation. To calibrate the response of different feature maps, previous works generally encoded the importance of feature maps into channel-wise weights via global max and/or average pooling operations. The weights are then applied to feature maps to highlight the informative features while suppressing the redundant ones [51]. However, these global pooling techniques treat all pixels equally and ignore the semantic/geometric content differences between the regions. Therefore, the previous techniques either fail to deal with noisy features or accidentally lose the geometric details. An example is shown in Fig. 2.

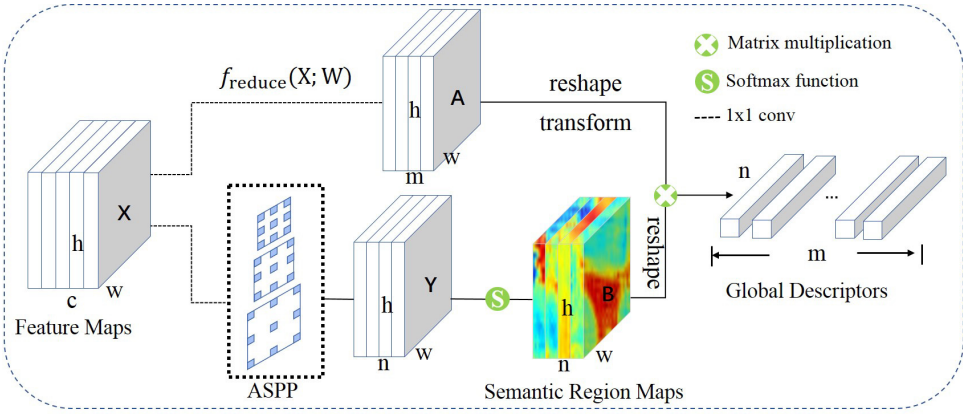


Fig. 3. SAP operation. $f_{\text{reduce}}(\mathbf{X}; \mathbf{W}) = \mathbf{W}(\mathbf{X})$ is a 1×1 convolution layer without bias.

In Fig. 2, we chose three types of representative feature maps for illustration, termed boundary, region, and noisy feature maps. Note that the warmer is the color, the higher is the feature response (pixel value). For instance, the region feature map has a high response on pixels of building region and it is thus important for correctly segmenting a building. In contrast, the noisy feature map has a high response on pixels of uninformative region and it thus contributes less or even harmful for semantic segmentation. Therefore, our goal is to extract the global information of these feature maps and encode them as weights to apply on the original feature maps to preserve the informative features while suppressing the uninformative ones. However, for instance, if adapting a global average pooling, the boundary feature map will yield a small weight value (-0.212), meaning that the importance of geometric (edge) features is underestimated. As edge pixels only occupy an extremely small area, their weights could be smoothed through an average pooling. From Fig. 2, the global max pooling is also problematic. It can yield a large weight value (0.268) for noisy feature map that has strong response on uninformative features, which is even larger than that of boundary feature map (0.219) and region feature map (0.194). This does not help to filter the noisy features and even impedes the learning of informative features.

To gather global representations with awareness to context, we therefore design an SAP method to encode the feature maps of different modalities into representative global descriptors. These descriptors are further fused across modalities for information exchange. With the rich cross-modal global descriptors, we finally infer a channel-wise weight vector for RGB feature refinement.

1) Self-Adaptive Region Pooling: From Fig. 2, we can observe that different feature maps of a given data modality focus on different contexts, i.e., some maps highlight the object regions of specific classes, while some others prefer to highlight boundaries of objects. To encode the importance of a feature map at the region level, we introduce an SAP operation, which encodes the context of each feature map into a global descriptor, as shown in Fig. 3. The SAP first enhances the given features $\mathbf{X} \in \mathbb{R}^{c \times h \times w}$ through a 1×1 convolution layer, followed by a commonly used ASPP module [31] for feature enrichment with multiscale contextual information. The enriched feature maps $\mathbf{Y} \in \mathbb{R}^{n \times h \times w}$ are then processed by a softmax function to generate the weight maps $\mathbf{B} \in \mathbb{R}^{n \times h \times w}$, which is formulated as follows:

$$\mathbf{B}_k[i, j] = \frac{e^{\mathbf{Y}_k[i,j]}}{\sum_{i=1}^{h} \sum_{j=1}^{w} e^{\mathbf{Y}_k[i,j]}} \tag{1}$$

where $\mathbf{B}_k[i, j]$ and $\mathbf{Y}_k[i, j]$ represent the values at position (i, j) in the kth map in $\mathbf{B}$ and $\mathbf{Y}$, respectively. The adoption of a softmax operation helps to preserve the response of the most prominent regions (e.g., semantic regions that have high response) while smoothing the less representative and noisy areas in an obtained weight map. Hence, compared to the input feature maps $\mathbf{Y}$, the weight (feature) maps in $\mathbf{B}$ focus more on the prominent semantic regions and we term $\mathbf{B}$ as semantic region maps.

With the semantic region maps $\mathbf{B}$, it is ready to embed the global context information in $\mathbf{X}$ into a series of representative descriptors. To create more compact representation and alleviate the computation burden, the feature maps $\mathbf{X}$ with dimension c is first reduced to $\mathbf{A} = f_{\text{reduce}}(\mathbf{X}; \mathbf{W}) \in \mathbb{R}^{m \times h \times w}$ with 1×1 convolution layer, in which case m is empirically set as 128. For each feature map in $\mathbf{A}$, we compute an n-dimensional global descriptor encoding its global information by applying $\mathbf{B}$ to it, as shown in Fig. 3. Specifically, we utilize $\mathbf{A}_l \in \mathbb{R}^{h \times w}$ to represent the lth feature map of $\mathbf{A}$, with $\mathbf{a}_l \in \mathbb{R}^{hw}$ as the corresponding reshaped vector. Similarly, the semantic region maps $\mathbf{B}$ are reshaped to the matrix $\mathbf{B}_{\text{re}} \in \mathbb{R}^{n \times hw}$. Then, the lth global descriptor vector $\mathbf{g}_l \in \mathbb{R}^n$ is computed by

$$\mathbf{g}_l = \mathbf{B}_{\text{re}} \times \mathbf{a}_l. \tag{2}$$

The explanation of (2) is as follows. As each semantic region map in $\mathbf{B}$ focuses on different semantic regions, the kth value in the global descriptor $\mathbf{g}_l$ then indicates the response intensity of an lth feature map $\mathbf{A}_l$ to the semantic regions in the kth region map $\mathbf{B}_k$. In other words, the kth value in $\mathbf{g}_l$ also implies whether $\mathbf{A}_l$ contains useful information in semantic region represented by $\mathbf{B}_k$. The SAP can be regarded as performing a region-guided pooling to encode the global information of a feature map into a semantic region-aware

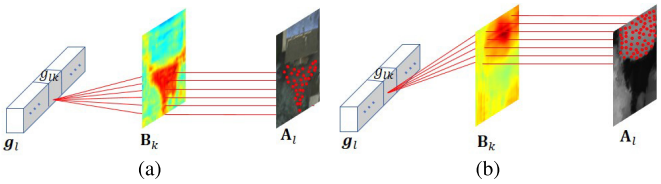


Fig. 4. Visualized mapping relationship between a feature map and its global descriptor. (a) and (b) Mapping relationship for RGB and auxiliary data, respectively. The value g_{lk} indicates how much informative information is contained in $\mathbf{A}_l$, with respect to the semantic region given in $\mathbf{B}_k$. With different semantic regions learned from cross-modal data, different information is encoded in RGB and auxiliary global descriptors. (a) RGB global descriptor. (b) Aux global descriptor.

vector. Compared with global max-/average-pooling operation that compresses the global information of a feature map into a single scalar, our SAP encodes more regional semantic information with a vector. To better understand our SAP, an illustration of the relationship between a feature map and its global descriptor is given in Fig. 4.

Take Fig. 4(a) as an example, where $\mathbf{A}_l$ is the lth feature map in $\mathbf{A}$, $\mathbf{B}_k$ is the kth semantic region map in $\mathbf{B}$, and $\mathbf{g}_l$ is the global descriptor regarding $\mathbf{A}_l$. For intuitive understanding, we use the original RGB image to assist the visualization of feature map $\mathbf{A}_l$ and use the red dots to represent the encoded semantic region of $\mathbf{A}_l$. Then, the value g_{lk} in $\mathbf{g}_l$ indicates the response of $\mathbf{A}_l$ with respect to the semantic region (hot color) in $\mathbf{B}_k$. In this case, if $\mathbf{A}_l$ contains rich information in the region represented by $\mathbf{B}_k$, g_{lk} will receive a high value. Otherwise, if $\mathbf{B}_k$ prefers boundary information, g_{lk} with a large value will indicate that $\mathbf{A}_l$ contain boundary clues. In this way, the semantic information of several regions in an original feature map can be distilled into a context-aware global descriptor, instead of a scalar calculated by global average- or max-pooling operators with only a smoothed or maximum response intensity.

2) Cross-Modal Fusion: RGB data can provide much more favorable features for multiobject segmentation than other data types, such as DSM/nDSM, but also brings ambiguity because of complex textures. It is hence favorable to integrate complementary parsing merits of an auxiliary modality to enhance the features derived from RGB data. For example, the color images derive similar textures of the low and high vegetation objects, while nDSM data can provide clear different features.

To fully exploit the advantages of different modality data, we extract global descriptors for features learned from RGB and auxiliary data with SAP. Then, we aggregate the global descriptors of two modalities into cross-modal features that embed the information from auxiliary data into the descriptors of RGB images. Although multimodal features contain complementary information, they may also have different uninformative noise/controversial information and context gap caused by the discrepancy between data domains. In such a case, aggregating descriptors obtained from different input modalities without information selection can be disadvantageous for CMF. Therefore, we introduce an effective CMF procedure with a gated structure, as shown in Fig. 5.

Let $\mathbf{G}^{\text{rgb}} \in \mathbb{R}^{n\times m}$ and $\mathbf{G}^{\text{aux}} \in \mathbb{R}^{n\times m}$ be the two modality-specific descriptor sets generated from RGB data

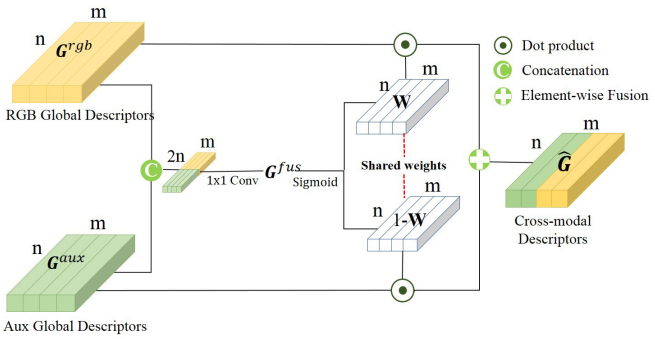


Fig. 5. Pipeline of the proposed CMF operation.

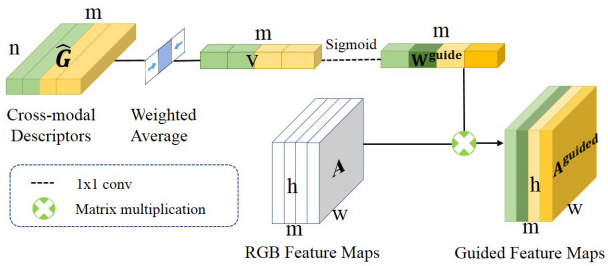


Fig. 6. Calibration of RGB feature response with cross-modal guidance.

and auxiliary data, respectively. Information in $\mathbf{G}^{\text{rgb}}$ and $\mathbf{G}^{\text{aux}}$ is first correlated through a concatenation operation and a 1×1 convolution layer to obtain a more abundant fused descriptor set $\mathbf{G}^{\text{fus}} \in \mathbb{R}^{n\times m}$. A sigmoid function is then applied to $\mathbf{G}^{\text{fus}}$ to yield weights $(\mathbf{W}, \mathbf{1}\text{-}\mathbf{W}) \in \mathbb{R}^{n\times m}$ that measures the contributions of descriptors in $\mathbf{G}^{\text{rgb}}$ and $\mathbf{G}^{\text{aux}}$. With $\mathbf{W}$, the gated fusion can be performed as follows:

$$\widehat{\mathbf{G}}^{\text{rgb}} = \mathbf{G}^{\text{rgb}} \odot \mathbf{W}; \quad \widehat{\mathbf{G}}^{\text{aux}} = \mathbf{G}^{\text{aux}} \odot (1 - \mathbf{W}) \tag{3}$$

where $\odot$ denotes the Hadamard product. In (3), the gated fusion optionally allows information to flow through along with $\mathbf{W}$ and $\mathbf{1}\text{-}\mathbf{W}$, which determines how much the information of a certain modality should be preserved. With the refined descriptors $\widehat{\mathbf{G}}^{\text{rgb}}$ and $\widehat{\mathbf{G}}^{\text{aux}}$, we can acquire a series of fine-grained cross-modal global descriptors by

$$\widehat{\mathbf{G}} = \widehat{\mathbf{G}}^{\text{rgb}} + \widehat{\mathbf{G}}^{\text{aux}}. \tag{4}$$

The cross-modal global descriptors $\widehat{\mathbf{G}}$ are then used to enhance the RGB feature representation, which can be described as a feature guider.

C. Feature Guider

The feature guider is designed to calibrate the compact RGB feature maps $\mathbf{A}$ with the global information from enriched cross-modal descriptor $\widehat{\mathbf{G}}$. The architecture of the feature guider is shown in Fig. 6.

The feature guider first derives a channel-wise guide weight vector $\mathbf{W}^{\text{guide}}$ from the cross-modal global descriptor $\widehat{\mathbf{G}}$ to explore the importance of each feature map in $\mathbf{A}$. The guide weight vector is then used to calibrate the information in $\mathbf{A}$ by

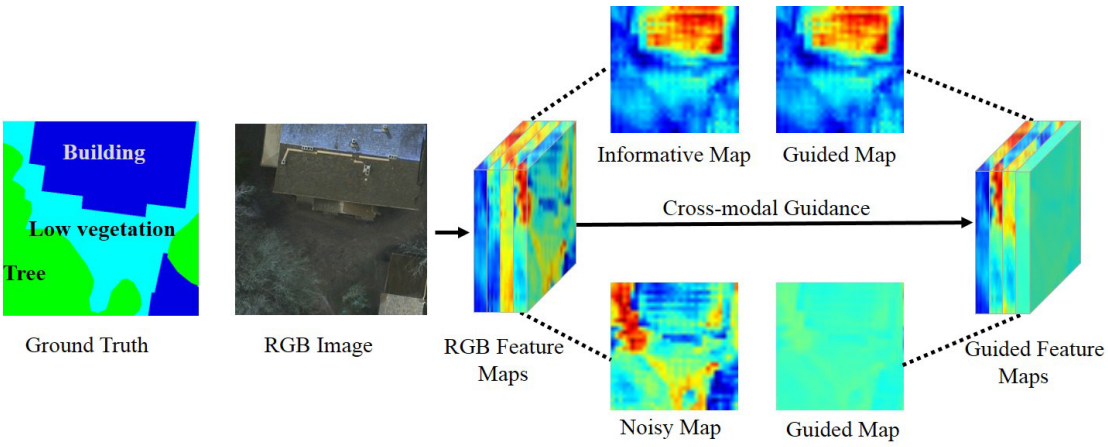


Fig. 7. Example of heat map visualization of unguided feature maps and guided feature maps. The warmer is the color, the higher is the response of features.

adaptively preserving the representative feature maps and suppressing the uninformative/noise features. In order to obtain the channel-wise guide weight vector $\mathbf{W}^{\text{guide}}$, we squeeze the cross-modal global descriptors into a global vector denoted as $\mathbf{v} \in \mathbb{R}^{m\times 1}$ via a weighted average operation and then transform the global vector $\mathbf{v}$ via two 1×1 convolution layers to learn the nonlinear interaction. Subsequently, the sigmoid function is used to generate a channel-attention weight. Detailed workflow for the generation of the guide weight vector is described by the following equation:

$$\mathbf{W}^{\text{guide}}[k] = \frac{1}{1+e^{-\mathbf{v}[k]}};\ \mathbf{v}[k] = \text{Conv}\left(\frac{\sum_{i=1}^{n}\widehat{\mathbf{g}}_k[i]}{n}\right). \quad (5)$$

In (5), $\mathbf{W}^{\text{guide}}[k]$ and $\mathbf{v}[k]$ are the values at the kth position of $\mathbf{W}^{\text{guide}}$ and $\mathbf{v}$, respectively, while $\widehat{\mathbf{g}}_k$ is the kth vector (descriptor) in $\widehat{\mathbf{G}}$, where $\widehat{\mathbf{G}}$ comprises the set of cross-modal global descriptors.

The vector $\mathbf{W}^{\text{guide}} \in \mathbb{R}^{m\times 1}$ can now be readily used to guide the refinement of the RGB feature maps $\mathbf{A} \in \mathbb{R}^{m\times h\times w}$ through a gated mechanism as follows:

$$\mathbf{A}_l^{\text{guided}} = \mathbf{A}_l \cdot \mathbf{W}^{\text{guide}}[l] \quad (6)$$

where $l \in \{1, \ldots, m\}$ is the channel index. The information in $\mathbf{A}_l$ will be largely preserved with $\mathbf{W}^{\text{guide}}[l]$ approximating 1, while it will be erased when $\mathbf{W}^{\text{guide}}[l]$ is close to 0. The guide weight vector $\mathbf{W}^{\text{guide}}$ thus controls the information flow of $\mathbf{A}_l$. For perceptual understanding, the refinement effect of the feature guider is shown in Fig. 7.

Fig. 7 shows the response distributions of different feature maps. The feature map shown in the top right of Fig. 7 focuses on the interior area of building class, whereas the counterpart in the bottom right of Fig. 7 yields noisy responses across the image. With the cross-modal guidance from feature guider, the informative features are efficiently preserved, whereas the noisy features are sufficiently suppressed, in the corresponding guided feature maps. In this way, the redundant and noisy information is largely filtered by the feature guider, enabling the RGB features to be more concentrated on the semantic segmentation task.

IV. Experiment

To validate the effectiveness of the proposed G2GNet, experiments were conducted on two widely used benchmark datasets, ISPRS Vaihingen 2-D and Postdam 2-D [56]. In Sections IV-A–IV-E, we will detailedly describe the two datasets, the experimental settings, and results, and the comprehensive analysis.

A. Dataset

The ISPRS Vaihingen 2-D and Postdam 2-D are two benchmark datasets launched in the ISPRS 2-D semantic labeling contest. These datasets are manually annotated with pixel-wise labels. Each pixel is classified into one of the following six land cover classes: impervious surface (Imp.S.), buildings (Build.), low vegetation (Low.V.), trees, cars, and clutter/background (e.g., containers, tennis courts, or swimming pools).

1) ISPRS Vaihingen: The Vaihingen dataset contains 33 patches with 9-cm/pixel resolution collected over a 1.38 km^2 area of the city. Each of the patches consists of a three-brand IRRG (NIR, R, and G) true orthophoto (TOP) image and two complementary modality data (the corresponding DSM and nDSM). The size of each image is approximately 2500 $\times$ 2000 pixels. The dataset is officially split into two groups, 16 images for training and 17 images for testing. This official setting is also followed in our experiments.

2) ISPRS Potsdam: The Potsdam dataset contains 38 patches with 5-cm/pixel resolution, covering a spatial area of 3.42 km^2. Each patch consists of a four-brand (NIR, R, G, and B) TOP image and the corresponding DSM. In addition to the DSMs, the nDSMs are provided by the official setting with two different normalization methods. The size of all images is 6000 $\times$ 6000 pixels. In our experiment, 24 images are used for training, and the remaining 14 images are used for testing, according to the official setting.

B. Experimental Setup

1) Implementation Details: The proposed G2GNet was implemented with the Pytorch framework, and all our models

were trained with the stochastic gradient descent optimizer on NVIDIA GTX 2080Ti GPUs. The initial learning rate was set to 0.01, and the value is decreased step by step via a poly learning rate strategy. The momentum was set to 0.9, and the weight decay was set to $5e^{-}4$. The ResNet-101 [57] was employed as the backbone model for our experiments, and the batch size was set to 6. Considering that the original remote sensing images are too large to be used as input, we cropped the training images into 640×640 patches. Data augmentation techniques, including random flipping, angle rotation, scaling, and cropping, are adopted on these patches to avoid overfitting. The run time of the G2GNet for an image of 640×640 with one GPU is about 0.2 s.

2) Evaluation Metric: The numeric performance of the proposed G2GNet model on the two datasets is evaluated by the following commonly used metrics: precision, recall, F1-score (F1), mean F1-score (mF1), and overall accuracy (OA). The evaluation is based on an accumulated confusion matrix, from which precision, recall, F1, and OA can be derived

$$\begin{aligned} \text{Precision} &= \frac{1}{N}\sum_{k=1}^{N}\frac{\text{TP}_k}{\text{TP}_k+\text{FP}_k} \\ \text{Recall} &= \frac{1}{N}\sum_{k=1}^{N}\frac{\text{TP}_k}{\text{TP}_k+\text{FN}_k} \\ \text{F1} &= 2\cdot\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} \\ \text{OA} &= \frac{\sum_{k=1}^{N}\text{TP}_k}{\sum_{k=1}^{N}\text{TP}_k+\text{FP}_k+\text{TN}_k+\text{FN}_k} \end{aligned} \tag{7}$$

where TP_k, FP_k, TN_k, and FN_k denote the number of true positive, false positive, true negative, and false negative pixels for object indexed as class k, respectively.

3) Experimental Description: The proposed G2GNet was evaluated on the Vaihingen test set. Following the previous works, the three-band IRRG images on this dataset were adopted as a substitution for the RGB input. Besides, as many methods utilized the NDVI data as the auxiliary image for segmentation, we also extend our G2GNet to the segmentation of RGB(IRRG) and NDVI data. The NDVI of each TOP image can be computed as follows:

$$\text{NDVI} = \frac{\text{NIR}-R}{\text{NIR}+R} \tag{8}$$

where NIR and R denote the value of near-infrared and red bands, respectively. The nDSM and NDVI data were, respectively, fed together with IRRG images into our two-stream G2GNet. The second group of experiments was conducted on the Potsdam test set. The data configurations are similar to those in the experiments on the Vaihingen dataset, without the usage of primary data. The normal RGB images were fed together with nDSM or NDVI into the G2GNet for semantic inference.

As the two datasets both contain color images and several types of auxiliary data (e.g., DSM, nDSM, or NDVI), hence, they are suitable for testing whether G2GNet can be generalized into different remote sensing data. Moreover, measurement uncertainties (e.g., missing data of DSM) and labeling inaccuracy of these datasets are common in real-world conditions. From a practical point of view, those data quality issues are helpful in revealing our model's ability to prevent exceptional feature propagation from low-quality data.

C. Results on Vaihingen Dataset

The comparison results with the state of the art on the Vaihingen test set are listed in Table I. Details on data used by different methods are also presented in Table I. Most of the methods utilize the auxiliary DSM data; some of them leverage both DSM and NDVI data, and a few of them use the single IRRG (TOP) image. This trend suggests that most of the methods manage to exploit the rich complementary information of different remote sensing data for semantic segmentation.

In Table I, it is interesting that in the top-ranking list, the two methods of HUSTW5 and SBANet only use IRRG data for segmentation but achieved higher OA and/or mF1 score than the other comparison methods that fuse the IRRG and auxiliary data. We suspect that those fusion-based semantic segmentation methods do not fully consider the data uncertainties of the additional modalities when fusing them with IRRG data, thus leading to inferior performance. For instance, the DSM data contained in the Vaihingen dataset are generated by dense image matching, resulting in the data missing of textureless areas and inaccurate features in vegetation areas. The low-quality measurements could introduce exceptional features into deep networks and propagate error information for semantic prediction. Compared with those listed fusion-based methods, the proposed G2GNet seems to be more capable of incorporating the additional data to complement the IRRG segmentation. The proposed G2GNet gains over the other fusion-based methods by 1.4%–7.3% in terms of OA. The G2GNet also favorably outperforms HUSTW5 and SBANet in terms of both OA and mF1 score, by incorporating either nDSM or NDVI data with the IRRG images. The G2GNet performs CMF in a gather-to-guide manner, by which the exceptional features are suppressed, while the informative ones are aggregated to guide the IRRG response refinement. The qualitative comparison results are shown in Fig. 8 for visual inspection.

According to the types of data used for segmentation, three methods (with available visual results) are chosen for comparison, as shown in Fig. 8. The samples include HUSTW5 (IRRG-only), DLR_9 (IRRG+nDSM), and RIT_7 (IRRG+nDSM+NDVI). The results of our G2GNet in the first two rows are obtained by fusing IRRG with nDSM, and those in the last two rows are derived by combining IRRG with NDVI. The results in Fig. 8 show that the G2GNet can achieve coherent labeling of various types of urban objects, whereas the other methods somewhat suffer from inconsistent segmentation effects, particularly for the building category that presents a large variation in appearance. The HUSTW5 uses only IRRG data for semantic segmentation, which is prone to be affected by confusing textures. For example, in the first row, part of the building (bottom right of the original IRRG image) is visually similar to the category of low vegetation.

TABLE I

QUANTITATIVE COMPARISON RESULTS ON THE ISPRS VAIHINGEN CHALLENGE 2-D TEST SET, WHERE THE VALUES IN BOLD ARE THE BEST. THE SHORT NAMES OF DIFFERENT METHODS ARE CITED FROM THE CHALLENGE EVALUATION WEBSITE

Method	IRRG	DSM/nDSM	NDVI	F1					OA	mF1
				Imp.S.	Build.	Low.V.	Tree	Car		
UZ_1 [17]	✓	✓		89.2	92.5	81.6	86.9	57.3	87.3	81.50
DSCNN [18]	✓	✓	✓	91.0	94.5	84.4	89.9	77.8	89.8	87.52
DST_2 [7]	✓	✓		90.5	93.7	83.4	89.2	72.6	89.1	85.88
ADL_3 [15]	✓	✓		89.5	93.2	82.3	88.2	63.3	88.0	83.30
GSN [58]	✓			92.2	95.1	83.7	89.9	82.4	90.3	88.70
RSCNN [43]	✓	✓	✓	87.9	92.0	78.4	85.4	72.1	85.9	83.16
RIT_7 [49]	✓	✓	✓	91.7	95.2	83.5	89.2	82.8	89.9	88.48
V-FuseNet [10]	✓	✓	✓	91.0	94.4	84.5	89.9	86.3	90.0	89.22
DLR_9 [20]	✓	✓		92.4	95.2	83.9	89.9	81.2	90.3	88.52
FMD [59]	✓	✓		92.3	95.8	83.8	89.6	86.4	90.6	89.58
CASIA2 [35]	✓			93.2	96.0	84.7	89.9	86.7	91.1	90.10
Kaiqiang [60]	✓	✓		88.7	92.9	80.6	86.4	72.6	87.1	84.24
HUSTW5 [36]	✓			93.3	96.1	**86.4**	**90.8**	74.6	91.6	88.24
SVL_4 [61]	✓	✓		86.1	90.9	77.6	84.9	59.9	84.7	79.88
DDCM [62]	✓			92.7	95.3	83.3	89.4	88.3	90.4	89.80
AFNet [63]	✓			93.4	95.9	86.0	90.7	87.2	91.6	90.64
SBANet [64]	✓			**94.4**	92.9	83.4	89.6	**91.4**	90.6	90.34
CF-Net [65]	✓			92.3	95.6	81.3	89.3	**91.4**	90.0	89.97
CCANet [66]	✓			93.3	94.3	82.0	88.6	86.6	91.1	88.96
G2GNet(ours)	✓		✓	93.6	96.4	86.1	**90.8**	87.2	91.9	90.82
G2GNet(ours)	✓	✓		93.7	**96.5**	86.1	**90.8**	88.2	**92.0**	**91.06**

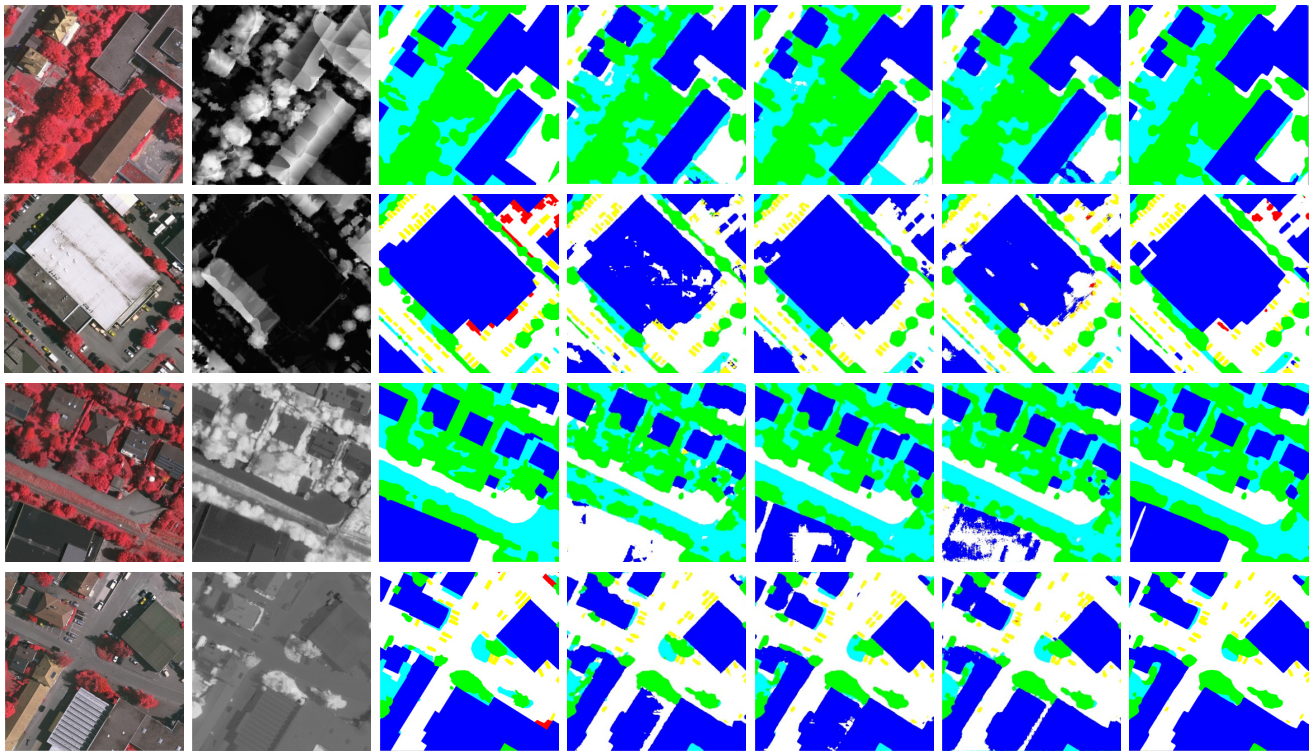


Fig. 8. Qualitative comparison results on the Vaihingen test set. (a) IRRG. (b) nDSM/NDVI. (c) Label. (d) RIT_7. (e) DLR_9. (f) HUSTW5. (g) G2GNet(ours).

Meanwhile, in the last row, the same building at the bottom of the IRRG image has notably different appearances from its roof. However, we are unable to determine the reason for the behavior of DLR_9 and RIT_7. As the nDSM in the first row offers available geometric cues to detect a complete building, the two methods both misclassified a small part of the building as low vegetation. In the second row, the large building with a weak roof texture was poorly reconstructed in nDSM. The anomalies of nDSM seem to have an impact on RIT_7, but not on DLR_9. We speculate that the direct feature-level fusion (via feature summation or concatenation) as in RIT_7 may introduce exceptional features from low-quality measurements

TABLE II

QUANTITATIVE COMPARISON RESULTS ON ISPRS POTSDAM CHALLENGE 2-D TEST SET, WHERE THE VALUES IN BOLD ARE THE BEST. THE SHORT NAMES OF DIFFERENT METHODS ARE CITED FROM THE CHALLENGE EVALUATION WEBSITE

Method	RGB	DSM/nDSM	NDVI	F1					OA	mF1
				Imp.S.	Build.	Low.V.	Tree	Car		
DST5 [7]	✓	✓		92.5	96.4	86.7	88.0	94.7	90.3	91.66
L.Yansong [19]	✓	✓		91.2	94.6	85.1	85.1	92.8	88.4	89.76
RIT_4 [49]	✓	✓	✓	92.6	97.0	86.9	87.4	95.2	90.3	91.82
V-FuseNet [10]	✓	✓	✓	92.7	96.3	87.3	88.5	95.4	90.6	92.04
UFMG_4 [43]	✓	✓		90.8	95.6	84.4	84.3	92.4	87.9	89.50
M-MRS [67]	✓	✓		90.9	97.0	76.3	73.4	88.6	90.7	85.24
CASIA2 [35]	✓			93.3	97.0	87.7	88.4	96.2	91.1	92.52
HUSTW4 [36]	✓			93.6	97.6	88.5	88.8	94.6	91.6	92.62
SVL_1 [61]	✓	✓		83.5	91.7	72.2	63.2	62.2	77.8	74.56
BUCTY5 [61]	✓	✓		93.1	97.3	86.8	87.1	94.1	90.6	91.68
SWJ_2 [61]	✓			**94.4**	97.4	87.8	87.6	94.7	91.7	92.38
AZ3 [61]	✓	✓	✓	93.1	96.3	87.2	88.6	96.0	90.7	92.24
DDCM [62]	✓			92.9	96.9	87.7	89.4	94.9	90.8	92.30
AFNet [63]	✓			94.2	97.2	89.2	89.4	95.1	92.2	93.02
SBANet [64]	✓			93.8	**98.0**	**89.0**	89.5	94.7	**92.8**	93.01
CF-Net [65]	✓			90.9	94.2	86.5	84.7	95.5	88.3	90.37
G2GNet(ours)	✓		✓	**94.4**	97.5	88.8	**89.8**	**96.7**	92.2	**93.44**
G2GNet(ours)	✓	✓		**94.4**	97.6	88.8	89.7	**96.7**	92.2	**93.44**

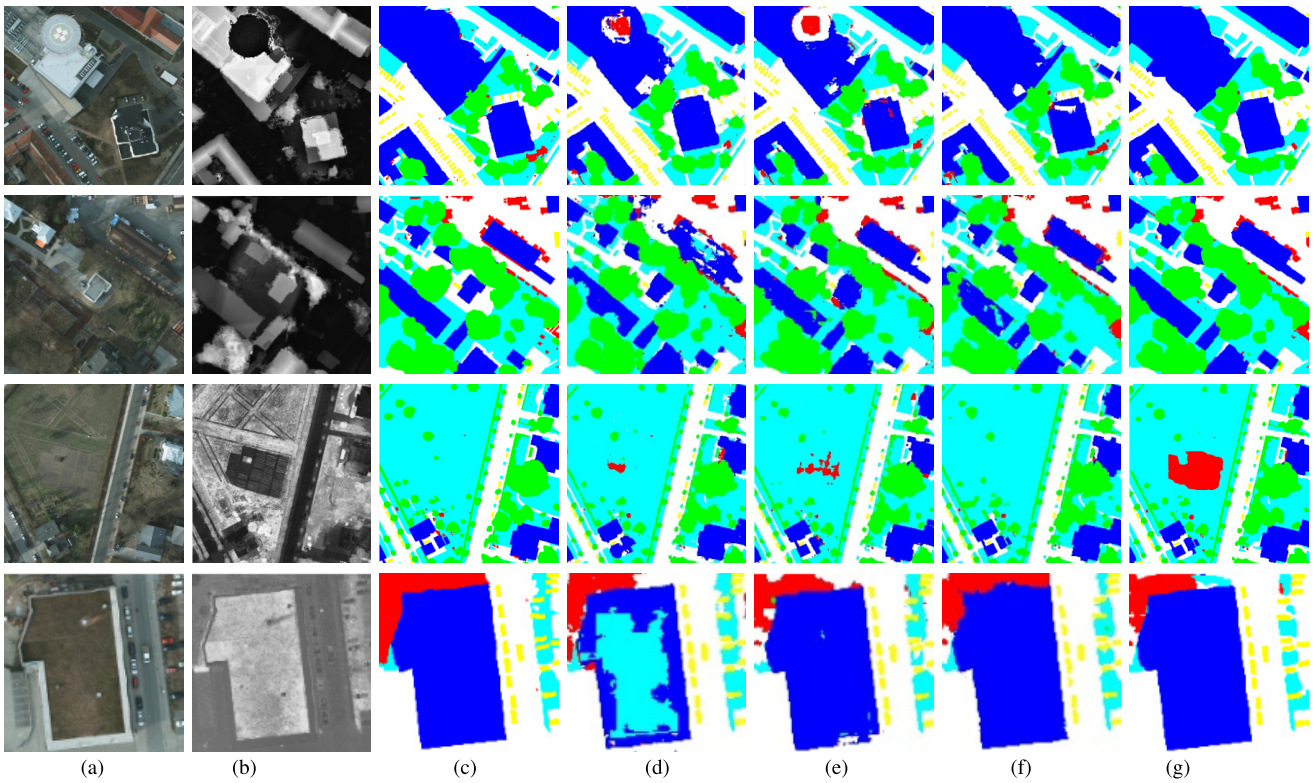


Fig. 9. Qualitative comparison results on the Potsdam test set. (a) RGB. (b) nDSM/NDVI. (c) Label. (d) AZ3. (e) RIT_4. (f) SWJ_2. (g) G2GNet(ours).

that disturb the final semantic inferring, while simply ensembling results of different modalities or models as in DLR_9 can be dominated by IRRG predictions. Both the above fusion strategies can hardly fully exploit the complementary merits of the two input data. By comparison, the proposed G2GNet is better at taking advantage of the two input data. This superiority can also be interpreted from the fourth row, where the building (at the bottom) in the NDVI image shows a more consistent appearance in different parts, and the G2GNet satisfactorily captured such cues to complement IRRG for generating intact segmentation maps.

D. Results on Potsdam Dataset

The quantitative comparison results between G2GNet and other competitors on the Potsdam test set are listed in Table II, where the overall situation is similar to that of the Vaihingen dataset. The two RGB-only methods, AFNet and SBANet,

achieve better performance in terms of OA, and they also gained higher mF1 scores compared with most of the other methods including those configured with various types of auxiliary data. We again suspect that those fusion-based methods may lack the ability to simultaneously suppress the erroneous information from input modalities while preserving their modality-specific features during fusion. Hence, the auxiliary data did not substantially improve their RGB counterpart, and they sometimes may also worsen the final segmentation results. From the table, the G2GNet reaches the value of 92.2% for the OA and 93.44% for the mF1 score when using nDSM as the additional input. G2GNet gains over the other fusion-based methods by 1.5%–14.4% in OA. Besides, by combining RGB with NDVI, the G2GNet still obviously excels all other methods in terms of mF1 score. As our segmentation decoder is similar to many of the comparison methods, the gain in numerical accuracy implies that the proposed G2GM is able to strengthen RGB feature representation for improved semantic prediction. The qualitative comparison results are shown in Fig. 9.

In Fig. 9, the SWJ_2 and AZ3 (with available visual results) are chosen as the representatives of the RGB-only and fusion-based methods, respectively. The RIT_4 presented in the same work of RIT_7 is also added to assist the visual evaluation. The visual results show that despite the use of DSM/nDSM data, AZ3 and RIT_4 still fail to coherently label some buildings with confusing textures. As shown in the first row, the completeness and the geometric structures of some buildings labeled by AZ3 and RIT_4 are both unsatisfactory. Moreover, the two methods both misclassified the helipad on top of a building as a road due to interruptions caused by erroneous height information (missing data) of nDSM. SWJ_2, which uses only the TOP image, is able to avoid the influence of data noises and anomalies from the additional modalities, but it is also easily affected by texture ambiguity and illumination changes, as depicted by the results in the sixth column. Equipped with a two-stream network, our proposed G2GNet model is able to procure highly robust and precise segmentation results for the uneven regions, as shown in the last column of Fig. 9. The G2GM deployed in our model is adept at absorbing the useful information of two input modalities to maximize the RGB segmentation accuracy. More representative examples can be found in the third and fourth rows of Fig. 9. In the third row, the bare land is difficult to decipher in the RGB image, but it is distinguishable in NDVI data. Although this bare land was wrongly annotated as low vegetation in the ground truth labels, it is successfully delimited by our G2GNet. By contrast, AZ3 and RIT_4 are only able to detect extremely small parts of this bare land, as they also utilize the auxiliary NDVI data for segmentation. Besides, it is not surprising that SWJ_2 completely misinterpreted this bare land as low vegetation due to the lack of essential cues from NDVI data. In the fourth row, the large building is weakly distinguishable in the RGB images but is prominent in the NDVI image. Therefore, most of the methods can recognize the building, but the G2GNet draws out more accurate boundaries by effectively highlighting the RGB response with NDVI information.

TABLE III
ABLATION STUDY ON THE VAIHINGEN AND POTSDAM TEST SETS

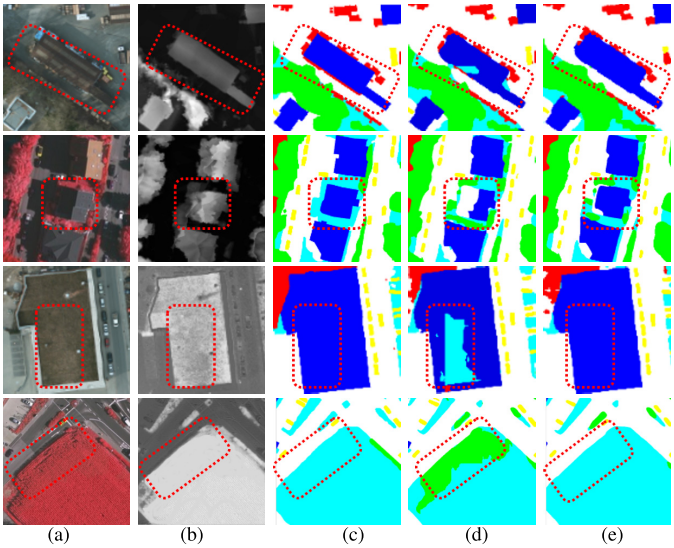
Method	F1					OA	mF1
	Imp.S.	Build.	Low.V.	Tree	Car		
Vaihingen							
Baseline	92.1	94.8	83.8	89.6	85.4	90.2	89.13
G2GNet(nDSM)	**93.7**	**96.5**	**86.1**	**90.8**	**88.2**	**92.0**	**91.06**
G2GNet(NDVI)	93.6	96.4	**86.1**	**90.8**	87.2	91.9	90.82
Potsdam							
Baseline	93.0	96.5	86.3	88.2	95.8	90.3	91.96
G2GNet(nDSM)	**94.4**	**97.6**	**88.8**	89.7	96.7	**92.2**	**93.44**
G2GNet(NDVI)	**94.4**	97.5	**88.8**	**89.8**	96.7	**92.2**	**93.44**

E. Experimental Analysis

1) Ablation Study for Complementarity Exploitation: To further verify whether the G2GNet is truly effective in unifying information of two modalities for extending the performance bounds of RGB/IRRG segmentation, we conducted ablation studies on both Vaihingen and Potsdam test sets. The RGB branch of the G2GNet is used as the baseline network, and the training settings are the same as those in our cross-modal version of G2GNet. The quantitative results are listed in Table III. Compared with the baseline network, the G2GNet yields an obvious increase in F1 score for the different categories on both datasets, by either using nDSM or NDVI as the auxiliary data. For the Vaihingen test set, the G2GNet boosts the baseline network at most by 1.8% and 1.93% in terms of OA and mF1 score. The G2GNet also improves the baseline network by 1.9% and 1.47% in the Potsdam test set. The accuracy gains of G2GNet on both datasets probably contribute to the enhanced segmentation in those uneven regions that are hard to interpret using the RGB/IRRG image. Such evidence can be partially found in the visualized results, as presented in Fig. 10.

In the first two rows of Fig. 10, the textures of the two buildings (highlighted by red rectangle) in the original RGB/IRRG images are visually confusing with the surrounding objects or other categories. The baseline network that uses only the RGB/IRRG images as input can hardly capture the complete context of the building, causing semantic inconsistencies in the predicted results. In the nDSM, the height information can provide clear geometric cues for delimiting the highlighted buildings from their surroundings. Such cues are desirably exploited by the G2GNet to complement the RGB/IRRG segmentation. Similarly, the NDVI images as shown in the last two rows of Fig. 10 offer an essential spectral supplement to the RGB/IRRG signals, and they are absorbed by G2GNet for the coherent labeling of those associated categories.

2) Ablation Study for Different Network Configurations: Ablation studies were also conducted to explore the effects of different network configurations. We first studied the effect of different pooling operations on semantic segmentation performance. In the feature gatherer part of our G2GNet, we replaced the proposed SAP with the global average-pooling (GAP) and the global max-pooling (GMP) operation. The numerical



(a) (b) (c) (d) (e)

Fig. 10. Qualitative results of ablation study on the Vaihingen and Potsdam test sets. (a) RGB/IRRG. (b) DSM/NDVI. (c) Label. (d) Baseline. (e) G2GNet.

TABLE IV
ABLATION STUDY ON THE VAIHINGEN AND POTSDAM TEST SETS WITH DIFFERENT POOLING OPERATIONS

Method	F1					OA	mF1
	Imp.S.	Build.	Low.V.	Tree	Car		
Vaihingen							
G2GNet(GAP)	93.0	95.7	84.6	90.1	**89.4**	91.1	90.56
G2GNet(GMP)	93.2	95.8	84.6	90.0	89.1	91.1	90.54
G2GNet(SAP)	**93.7**	**96.5**	**86.1**	**90.8**	88.2	**92.0**	**91.06**
Potsdam							
G2GNet(GAP)	93.6	97.3	87.8	89.2	96.1	91.3	92.79
G2GNet(GMP)	93.4	96.5	88.1	89.0	96.1	91.1	92.62
G2GNet(SAP)	**94.4**	**97.6**	**88.8**	**89.7**	**96.7**	**92.2**	**93.44**

results are listed in Table IV. For both datasets, the G2GNet with our proposed SAP achieves the best performance in most metrics. Comparing to GAP and GMP that equally treat all pixels in a feature map, the SAP leverages the region context in a feature map for feature refinement is demonstrated to be more effective in the improvement of segmentation performance.

We then studied the number of semantic region maps in feature gatherer denoted as n, which is the main parameter in our G2GM. From Table V, with the number increase of semantic region maps at first, the segmentation accuracy for most of the metrics and categories also increases and reaches an optimal level when $n = 64$. However, as the continuously increasing of the number, the segmentation accuracy decreases for most of the metrics and categories. The experimental results imply that 64 is a good choice for the semantic region map number. We speculate that increasing n from 16 to 64 can distill more informative regional features with the increase in semantic region maps, thus providing rich information for representation enhancement in the subsequent process of G2GM. However, when n is increased from 64 to 128, many

TABLE V
ABLATION STUDY FOR THE NUMBER OF SEMANTIC REGION MAPS

	F1					OA	mF1
	Imp.S.	Build.	Low.V.	Tree	Car		
n=16	93.8	97.3	87.9	89.1	96.2	91.6	92.86
n=32	93.6	97.4	88.1	89.4	96.3	91.7	92.96
n=64	**94.4**	**97.6**	**88.8**	**89.7**	96.7	**92.2**	**93.44**
n=128	94.1	97.5	88.4	89.5	**96.8**	92.0	93.26

TABLE VI
ABLATION STUDY OF DIFFERENT NETWORK CONFIGURATIONS

Stages	F1					OA	mF1
	Imp.S.	Build.	Low.V.	Tree	Car		
Baseline	93.0	96.5	86.3	88.2	95.8	90.3	91.96
+R4	**94.4**	97.6	**88.8**	**89.7**	**96.7**	**92.2**	**93.44**
+R4,R3	94.1	**97.8**	88.6	89.2	96.6	92.0	93.26
+R4,R3,R2	94.0	97.5	88.2	89.5	96.5	91.9	93.14
+R4,R3,R2,R1	94.1	97.7	88.5	89.4	**96.7**	92.1	93.28

newly added semantic region maps become redundant, which not only consume computational resource but also propagate uninformative features to the following process, thus impeding the precise semantic segmentation.

To better understand the G2GNet architecture, the effects of inserting more G2GMs in the different convolutional stages of the G2GNet's encoder are investigated. The quantitative results on the Potsdam test set are listed in Table VI, where R1, R2, R3, and R4 denote the G2GMs inserted in residual layers 1, 2, 3, and 4, respectively. +R4 denotes the current G2GNet architecture. The results indicate that by adding more G2GMs into the model, the segmentation accuracy for the different categories and metrics does not manifest an obvious change. Applying the G2GM in residual layer 4 can obtain a relatively better segmentation performance than applying the other configurations. The results suggest that increasing G2GMs do not capture more valuable information for semantic segmentation. This is probably due to the fact that in lower level convolutional stages, feature maps could contain much noisy information and the semantic region features are not sufficiently prominent. Therefore, we consider inserting G2GM only in residual layer 4 as a best choice, as adding more G2GMs will also add burden on computational efficiency.

3) Robustness to Data Uncertainties: The issue of data uncertainties during cross-modal remote sensing semantic segmentation needs to be further highlighted in this study. Actually, the data quality problem of the two benchmark datasets, i.e., ISPRS Vaihingen 2-D and Potsdam 2-D, has been issued in many previous works [10], [20]. Especially, the DSM/nDSM of the two datasets contains so much error height information. The DSM data used in this study are generated by dense image matching, which could have missing data in textureless areas. Such DSM data are also inaccurate in vegetation areas and noisy at object boundaries. In the above experiments, we have shown examples about the influence

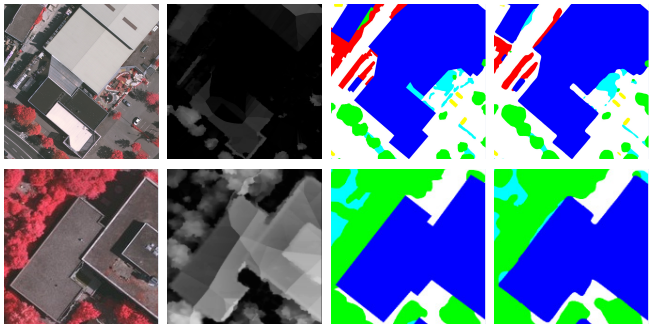


Fig. 11. Examples of semantic segmentation with G2GNet under data uncertainties. (a) IRRG. (b) nDSM. (c) Label. (d) Prediction.

of low-quality measurements on segmentation performance (see Figs. 8 and 9).

Concerns on data uncertainties are necessary when designing CNN-based fusion models, as quality issues are common in various types of Earth observation data [68]. It is not expected to indistinguishably propagate both useful and exceptional features from additional data. In our G2GNet, this problem was circumvented with a two-phase fusion scheme. The feature gatherer encodes and fuses feature maps of different modalities into representative cross-modal global descriptors, in which the less useful information is suppressed to a certain degree. Then, the feature guider infers a channel-wise weight vector for RGB feature refinement with the rich cross-modal global descriptors, which further mitigates the influence of noisy/exceptional features from the additional modality. Therefore, the G2GNet is more robust to data uncertainties than the previous methods. More examples can be seen in Fig. 11. The boundary noise and error height information in nDSM are rejected, while the sharp contours and complete structures of objects in the RGB images are preserved.

V. Conclusion

This study presented a G2GNet for robust semantic segmentation of RGB and auxiliary remote sensing data. The center building block of this architecture is the G2GM, which consists of two parts: a feature gatherer to self-adaptively filter the exceptional and less useful features while aggregating the informative ones from two input modalities, and a feature guider to refine the RGB feature response with the aggregated fused descriptor. Extensive experiments conducted on two challenging benchmark datasets show that the proposed G2GNet can achieve excellent segmentation results on different datasets. Besides the qualitative and quantitative comparison results, the ablation studies also verify that our G2GM is capable of taking advantage of two complementary modalities to boost the RGB segmentation performance. Further experimental analysis also convincingly demonstrates that the proposed G2GNet is extremely robust to data uncertainties. The G2GNet can deliver high-quality labeling results even under severe interruptions caused by data noises and anomalies from auxiliary data.

In the future, we would like to exploit the effectiveness of the proposed G2GNet to more types of Earth observation data and extend our idea of CMF to other remote sensing tasks.

References

[1] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[2] D. Hong *et al.*, "Interpretable hyperspectral artificial intelligence: When nonconvex modeling meets hyperspectral remote sensing," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 2, pp. 52–87, Jun. 2021.

[3] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, "Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 96–107, Nov. 2018.

[4] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2020.

[5] H. Zhang, L. Liu, W. He, and L. Zhang, "Hyperspectral image denoising with total variation regularization and nonlocal low-rank tensor decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3071–3084, May 2020.

[6] M. Längkvist, A. Kiselev, M. Alirezaie, and A. Loutfi, "Classification and segmentation of satellite orthoimagery using convolutional neural networks," *Remote Sens.*, vol. 8, no. 4, p. 329, 2016.

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

[8] Z. Cao, W. Diao, X. Sun, X. Lyu, M. Yan, and K. Fu, "C3Net: Cross-modal feature recalibrated, cross-scale semantic aggregated and compact network for semantic segmentation of multi-modal high-resolution aerial images," *Remote Sens.*, vol. 13, no. 3, p. 528, Feb. 2021.

[9] D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot, and X. X. Zhu, "X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS J. Photogramm. Remote Sens.*, vol. 167, pp. 12–23, Sep. 2020.

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

[11] M. D. Mura, S. Prasad, F. Pacifici, P. Gamba, J. Chanussot, and J. A. Benediktsson, "Challenges and opportunities of multimodality and data fusion in remote sensing," *Proc. IEEE*, vol. 103, no. 9, pp. 1585–1601, Sep. 2015.

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

[13] D. Hong *et al.*, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

[14] A. Lagrange *et al.*, "Benchmarking classification of Earth-observation data: From learning explicit features to convolutional networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 4173–4176.

[15] S. Paisitkriangkrai, J. Sherrah, P. Janney, and V. D. Hengel, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *Proc. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 36–43.

[16] W. Zhang, H. Huang, M. Schmitz, X. Sun, H. Wang, and H. Mayer, "Effective fusion of multi-modal remote sensing data in a fully convolutional network for semantic labeling," *Remote Sens.*, vol. 10, no. 2, p. 52, Dec. 2017.

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

[18] N. Audebert, B. L. Saux, and S. Lefèvre, "Semantic segmentation of Earth observation data using multimodal and multi-scale deep networks," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 180–196.

[19] Y. Liu, S. Piramanayagam, S. T. Monteiro, and E. Saber, "Dense semantic labeling of very-high-resolution aerial imagery and LiDAR with fully-convolutional neural networks and higher-order CRFs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 76–85.

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

[21] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[22] W. He, H. Zhang, and L. Zhang, "Total variation regularized reweighted sparse nonnegative matrix factorization for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3909–3921, Jul. 2017.

[23] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[24] V. Badrinarayanan, A. Handa, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," 2015, *arXiv:1505.07293*. [Online]. Available: http://arxiv.org/abs/1505.07293

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[26] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, vol. 11045. Cham, Switzerland: Springer, 2018, pp. 3–11.

[27] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1925–1934.

[28] X. Zheng, L. Huan, G.-S. Xia, and J. Gong, "Parsing very high resolution urban scene images by learning deep ConvNets with edge-aware loss," *ISPRS J. Photogramm. Remote Sens.*, vol. 170, pp. 15–28, Dec. 2020.

[29] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2881–2890.

[30] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for real-time semantic segmentation on high-resolution images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 405–420.

[31] 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. 2017.

[32] M. Vakalopoulou, K. Karantzalos, N. Komodakis, and N. Paragios, "Building detection in very high resolution multispectral data with deep learning features," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 1873–1876.

[33] W. Zhao and S. Du, "Learning multiscale and deep representations for classifying remotely sensed imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 113, pp. 155–165, Mar. 2016.

[34] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 645–657, Feb. 2017.

[35] Y. Liu, B. Fan, L. Wang, J. Bai, S. Xiang, and C. Pan, "Semantic labeling in very high resolution images via a self-cascaded convolutional neural network," *ISPRS J. Photogram. Remote Sens.*, vol. 145, pp. 78–95, Nov. 2018.

[36] Y. Sun, Y. Tian, and Y. Xu, "Problems of encoder-decoder frameworks for high-resolution remote sensing image segmentation: Structural stereotype and insufficient learning," *Neurocomputing*, vol. 330, pp. 297–304, Feb. 2019.

[37] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 4096–4105.

[38] Z. Zhang, "Microsoft Kinect sensor and its effect," *IEEE MultiMedia*, vol. 19, no. 2, pp. 4–10, Feb. 2012.

[39] L. Keselman, J. I. Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, "Intel realsense stereoscopic depth cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 1–10.

[40] W. Su and Z. Wang, "Regularized fully convolutional networks for RGB-D semantic segmentation," in *Proc. Visual Commun. Image Process. (VCIP)*, Nov. 2017, pp. 1–4.

[41] W. Wang and U. Neumann, "Depth-aware CNN for RGB-D segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 135–150.

[42] 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.* Cham, Switzerland: Springer, 2016, pp. 213–228.

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

[44] S.-J. Park, K.-S. Hong, and S. Lee, "RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4980–4989.

[45] D. Hong, J. Yao, D. Meng, Z. Xu, and J. Chanussot, "Multimodal GANs: Toward crossmodal hyperspectral–multispectral image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5103–5113, Jun. 2021.

[46] R. Fernandez-Beltran, J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and F. Pla, "Multimodal probabilistic latent semantic analysis for Sentinel-1 and Sentinel-2 image fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 9, pp. 1347–1351, Sep. 2018.

[47] M. Y. Yang, B. Rosenhahn, and V. Murino, *Multimodal Scene Understanding: Algorithms, Applications and Deep Learning*. New York, NY, USA: Academic, 2019.

[48] L. Mou and X. X. Zhu, "Spatiotemporal scene interpretation of space videos via deep neural network and tracklet analysis," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 1823–1826.

[49] S. Piramanayagam, E. Saber, W. Schwartzkopf, and F. Koehler, "Supervised classification of multisensor remotely sensed images using a deep learning framework," *Remote Sens.*, vol. 10, no. 9, p. 1429, Sep. 2018.

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

[51] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[52] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jun. 2019, pp. 510–519.

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

[54] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12416–12425.

[55] H. Li, K. Qiu, L. Chen, X. Mei, L. Hong, and C. Tao, "SCAttNet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 905–909, May 2021.

[56] M. Gerke, "Use of the stair vision library within the ISPRS 2D semantic labeling benchmark (Vaihingen)," ITC, Univ. Twente, Enschede, The Netherlands, Tech. Rep., 2015.

[57] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1026–1034.

[58] H. Wang, Y. Wang, Q. Zhang, S. Xiang, and C. Pan, "Gated convolutional neural network for semantic segmentation in high-resolution images," *Remote Sens.*, vol. 9, no. 5, p. 446, May 2017.

[59] W. Sun and R. Wang, "Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM," *IEEE Geosci. & Remote Sens. Lett.*, vol. 15, no. 3, pp. 474–478, Mar. 2018.

[60] K. Chen *et al.*, "Effective fusion of multi-modal data with group convolutions for semantic segmentation of aerial imagery," in *Proc. IGARSS - IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 3911–3914.

[61] I. ISPRS2D. (2020). *ISPRS 2D Semantic Labeling Contest*. Accessed: Jun. 16, 2020. [Online]. Available: https://www2.isprs.org/commissions/comm2/wg4/results/

[62] Q. Liu, M. Kampffmeyer, R. Jenssen, and A.-B. Salberg, "Dense dilated convolutions' merging network for land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6309–6320, Sep. 2020.

[63] R. Liu, L. Mi, and Z. Chen, "AFNet: Adaptive fusion network for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, early access, Nov. 16, 2020, doi: 10.1109/TGRS.2020.3034123.

[64] A. Li, L. Jiao, H. Zhu, L. Li, and F. Liu, "Multitask semantic boundary awareness network for remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, early access, Feb. 2, 2021, doi: 10.1109/TGRS.2021.3050885.

[65] C. Peng, K. Zhang, Y. Ma, and J. Ma, "Cross fusion net: A fast semantic segmentation network for small-scale semantic information capturing in aerial scenes," *IEEE Trans. Geosci. Remote Sens.*, early access, Jan. 29, 2021, doi: 10.1109/TGRS.2021.3053062.

[66] G. Deng, Z. Wu, C. Wang, M. Xu, and Y. Zhong, "CCANet: Class-constraint coarse-to-fine attentional deep network for subdecimeter aerial image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, early access, Feb. 12, 2021, doi: 10.1109/TGRS.2021.3055950.

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

[68] W. He *et al.*, "Non-local meets global: An integrated paradigm for hyperspectral image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 29, 2020, doi: 10.1109/TPAMI.2020.3027563.

Xianwei Zheng received the M.S. and Ph.D. degrees from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China, in 2010 and 2015, respectively.

He is currently working as an Associate Professor in computer vision and 3-D geographical information science (GIS) with Wuhan University. His research interests include indoor and outdoor scene parsing, 3-D computer vision and reconstruction, and geovisualization.

Xiujie Wu received the B.S. degree from the School of Mathematics and Statistics, Wuhan University, Wuhan, China, in 2019. She is currently pursuing the M.S. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University.

Her research interests include indoor and outdoor scene parsing, and multimodal learning.

Linxi Huan received the B.S. degree from the School of Mathematics and Statistics, Wuhan University, Wuhan, China, in 2018. She is currently pursuing the Ph.D. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University.

Her research interests include machine learning, scene parsing, and 3-D reconstruction.

Wei He (Member, IEEE) received the B.S. degree from the School of Mathematics and Statistics, Wuhan University, Wuhan, China, in 2012, and the Ph.D. degree from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, in 2017.

From 2018 to 2020, he was a Researcher with the Geoinformatics Unit, RIKEN Center for Advanced Intelligence Project, Tokyo, Japan, where he is currently the Research Scientist. His research interests include image quality improvement, remote sensing image processing and low-rank representation, and deep learning.

Hongyan Zhang (Senior Member, IEEE) received the B.S. degree in geographic information system and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2005 and 2010, respectively.

He has been a Full Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University, since 2016. He has authored/coauthored more than 90 research articles and eight talents. His research interests include image reconstruction for quality improvement, hyperspectral information processing, and agricultural remote sensing.

Dr. Zhang is a Reviewer of more than 30 international academic journals, including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. He scored first in the Pairwise Semantic Stereo Challenge of the 2019 Data Fusion Contest organized by the IEEE Image Analysis and Data Fusion Technical Committee. He is a Young Chang-Jiang Scholar appointed by the Ministry of Education of China. He serves as the Session Chair for the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) Conference and the 2015 IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS) Conference. He serves as an Associate Editor for *Photogrammetric Engineering and Remote Sensing* and *Computers and Geosciences*.