MISNet: Multiscale Cross-Layer Interactive and Similarity Refinement Network for Scene Parsing of Aerial Images

Wujie Zhou, *Member, IEEE*, Xiaomin Fan, Lu Yu, *Senior Member, IEEE*, and Jingsheng Lei

***Abstract*—Although progress has been made in multisource data scene parsing of natural scene images, extracting complex backgrounds from aerial images of various types and presenting the image at different scales remain challenging. Various factors in high-resolution aerial images (HRAIs), such as imaging blur, background clutter, object shadow, and high resolution, substantially reduce the integrity and accuracy of object segmentation. By applying multisource data fusion, as in scene parsing of natural scene images, we can solve the aforementioned problems through the integration of auxiliary data into HRAIs. To this end, we propose a multiscale cross-layer interactive and similarity refinement network (MISNet) for scene parsing of HRAIs. First, in a feature fusion optimization module, we extract, filter, and optimize multisource features and further guide and optimize the features using a feature guidance module. Second, a multiscale context aggregation module increases the receptive field, captures semantic information, and extracts rich multiscale background features. Third, a dense decoding module fuses the global guidance information and high-level fused features. We also propose a joint learning method based on feature similarity and a joint learning module to obtain deep multilevel information, enhance feature generation, and fuse multiscale and global features to enhance network representation for accurate scene parsing of HRAIs. Comprehensive experiments on two benchmark HRAIs datasets indicate that our proposed MISNet is qualitatively and quantitatively superior to similar state-of-the-art models.**

***Index Terms*—Cross-layer interaction, feature similarity, high-resolution aerial images (HRAIs), multiscale fusion, scene parsing.**

I. Introduction

THE scene parsing of high-resolution aerial images (HRAIs) is a basic processing task for assigning category labels to each image pixel [1]. It plays an important role in urban planning, change detection, 3-D semantic modeling of cities, and other applications [2], [3], [4], [5]. Recently, deep convolutional neural networks (DCNNs) have proven to be effective in many computer vision tasks, such as detection, segmentation, and classification [6], [7], achieving state-of-the-art (SOTA) performance. A DCNN automatically extracts hierarchical feature maps of various objects in an image. It can extract details in shallow layers and complex semantic cues in deep layers of the network. However, existing scene parsing methods for HRAIs often identify only a few categories and process single-source data, thereby limiting their applicability.

In addition to single-source data, scene parsing has benefited from auxiliary aerial image data, such as digital surface models (DSMs) [8] and synthetic aperture radar images [9]. The introduction of multisource data can effectively improve the robustness of the segmentation method [10]. As different forms of spectral data, these types of auxiliary aerial image data capture specific attributes of the same geospatial object and provide different insights for the overall learning of semantic objects [11], [12]. Therefore, complementary information and auxiliary aerial image data in a red–green–blue (RGB) representation can be applied to optimize the performance of scene parsing [9], [13]. We focus on the scene parsing of infrared–red–green (IRRG) images and DSMs, which have been extensively studied using DCNNs [9], [14], [15], [16].

Although progress has been made in recent years, various problems related to scene parsing and auxiliary aerial image data persist. For instance, the high data diversity leads to low interclass variance and high intraclass heterogeneity. Hence, confusion between trees and low vegetation as well as misidentification of human-made objects in urban areas can occur [17]. Moreover, existing methods, particularly those based on deep learning, suffer from two major problems: 1) insufficient spatial information for inference and 2) lack of contextual information. These problems result in poor segmentation around object boundaries and in other difficult areas such as shadowed regions [18].

Over the past few years, numerous studies on DCNNs have been conducted to improve the results of scene parsing in HRAIs. These images typically display complex scenes and cover large areas, posing challenges for scene parsing [10]. Similarly, very-high-resolution (VHR) images, multisource data images, and point clouds increase the complexity of scene parsing. For instance, building roofs can appear to be complex and diverse in urban areas captured in a VHR image. This is a typical issue, in which similar roofs have different spectra, and their imaging

Manuscript received 21 July 2022; revised 2 December 2022; accepted 4 February 2023. Date of publication 8 February 2023; date of current version 20 February 2023. This work was supported by the National Natural Science Foundation of China under Grant 61502429. *(Corresponding author: Wujie Zhou.)*

Wujie Zhou is with the School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China, and also with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: wujiezhou@163.com).

Xiaomin Fan and Jingsheng Lei are with the School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China (e-mail: fanxiaomin@zust.edu.cn; leijingsheng@zust.edu.cn).

Lu Yu is with the Institute of Information and Communication Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: yulu2@zju.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3243247

is affected by occlusions and shadows. Consequently, accurate labeling remains challenging in the segmentation of VHR aerial images [19].

For accurate segmentation, we introduce DSMs as auxiliary data, whose elevation information allows us to reduce the segmentation problems caused by high objects. Accordingly, we introduce a multiscale cross-layer interactive and similarity refinement network (MISNet) consisting of four modules: 1) feature fusion optimization module (FFOM), 2) multiscale context aggregation module (MCAM), 3) dense decoding module (DDM), and 4) joint learning module (JLM). Considering the limitations of single-source data, we use the FFOM to extract and filter complementary information from multisource fused features. To handle changes in object scales and complex scenes in HRAIs, the novel MCAM captures multiscale background features and guides decoding processing. In the decoder network, the DDM fuses global information and cross-source high-level features layer by layer to filter redundant cues. Finally, the JLM calculates the cosine similarity of the auxiliary output and low-level fused features to refine segmentation weights and improve low-level information.

Our primary contributions can be listed and summarized as follows.

1) We propose the MISNet to learn interactions and continuity between image data and selectively merge complementary information extracted from the IRRG and normalized DSMs (nDSMs) of multispectral images to improve scene parsing.
2) The MISNet adopts an encoder–decoder architecture and contains the FFOM, MCAM, DDM, and JLM. The FFOM relates to cross-source hierarchies and cross-hierarchical continuity. The MCAM fully uses dilated convolutions to guide global context information. The DDM uses dense connections to fuse multiscale features and global guidance information. The JLM refines the decoding features and adds auxiliary supervision for optimization.
3) Comprehensive experimental evaluations show that the MISNet achieves higher scene parsing performance than 10 SOTA methods on the Potsdam and Vaihingen benchmark datasets.

II. Related Work

A. Single-Source Scene Parsing

Long et al. [19] presented a novel framework for the scene parsing task. Ronneberger et al. [20] used skip connections to combine shallow features with deeper features and reused low-level features to recover more data. Badrinarayanan et al. [21] introduced an encoder–decoder architecture and applied up-pooling with the recorded pooling method to recover distinct data, such as edges and complex shapes. Chen et al. [22] presented the atrous spatial pyramid pooling, where parallel dilated convolution operation with different rates extracts multiscale cues. Xu et al. [23] proposed a context extraction architecture based on a high-resolution module to solve the class scale imbalance and uncertain boundary information problems. Mou et al. [24] introduced two simple and effective architecture units, spatial correlation block and channel correlation block, to learn and analyze the global correlation between any two spatial positions or feature graphs and then generate the relation-augmented feature representation. Aerial images contain a considerable amount of detailed information about ground objects that result in the images showing large inclass and small interclass variances. This makes it difficult for the images to be recognized. Therefore, an attention mechanism was proposed to solve the convolution locality limitation problem. Li et al. [25] proposed an ABCNet based on a bilateral framework, using a context path to capture global contextual cues and a spatial path to retain spatial details, and designed new modules to integrate and enhance features. Zhao et al. [26] presented a scene parsing architecture based on end-to-end attention, in which a pyramid attention pool module introduced the attention principle into the multiscale block for feature refinement. Based on the characteristics of aerial images, Zhao et al. [27] presented a model based on the regional self-attention mechanism, which can mine the relationships between pixels in the surrounding area. This attention module can effectively decrease the noise of feature mapping and the interference of redundant features.

B. RGB-D Scene Parsing

Unlike single-source scene parsing, RGB-D scene parsing incorporates depth features into the RGB features to improve scene parsing accuracy.

In a previous study on RGB-D scene parsing, Couprie et al. [28] first used the depth cues in the feature learning method to mark the whole scene, laying the foundation for the field of RGB-D indoor scene parsing. Gupta et al. [29] presented a height-above-ground, angle-with-gravity image-learning, and horizontal-disparity DCNN, which differs from the depth image and found that the feature learning effect was better than that of the depth image.

In existing research, the application of depth images in RGB image scene parsing is relatively mature, but there are many aspects that can be improved. Lin et al. [30] presented a multibranch DCNN, which segmented available depth images into feature layers with a common resolution, enriched context information with feature cascade, and improved scene parsing performance. Jiang et al. [31] presented the residual encoder and decoder architecture (RedNet) for the indoor scene analysis task. The effective combination of the long skip connection between the decoder and encoder and the short skip connection in the residual unit enables RedNet to achieve efficient performance. Hu et al. [32] proposed the complementary attention network (ACNet), which selectively collects the features of two different RGB and depth modes to extract weighted features. Chen et al. [33] presented a separation and aggregation gate (SA-Gate) operation to calibrate RGB features and multistage depth information extraction and aggregated the two alternately. Zhou et al. [34] presented a three-branch self-attention architecture (TSNet), which obtained RGB and depth inputs from the two backbone networks. Seichter et al. [35] presented the efficient scene analysis network (ESANet), which has high robustness and achieves fast inference. Qian et al. [36] presented a gated

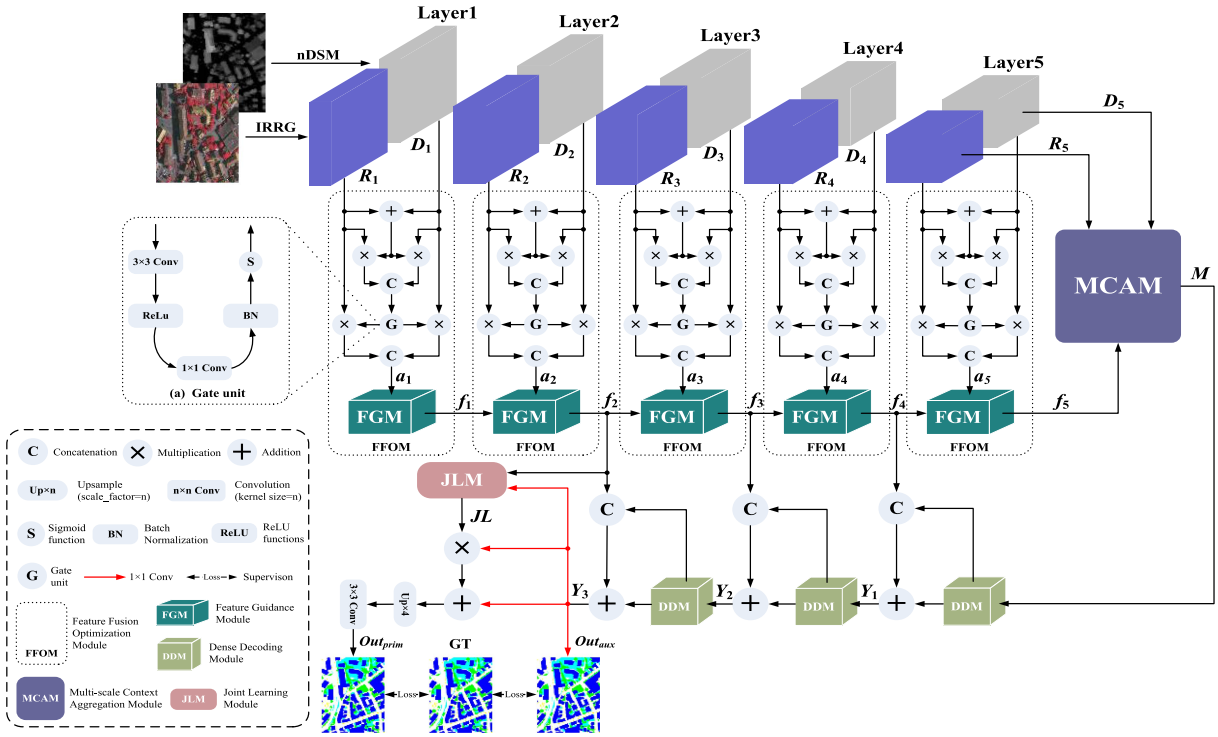

Fig. 1. Overall framework of the proposed MISNet.

residual block to effectively fuse RGB and depth signals and achieved excellent segmentation performance through complementary features calculated by the gate mechanism and specific features aggregated by residual blocks. Zhou et al. [37] designed a coattention fusion module that fused RGB and depth information into channel and spatial dimensions. Zhou et al. [38] presented a kind of collaboration of low-level and high-level cues optimized by depth enhancement and progressive guided fusion networks for indoor scene analysis.

C. Multisource Scene Parsing of HRAIs

Unlike RGB-D scene parsing, multisource scene parsing of HRAIs requires further development for fusing multisource features. However, only few studies are available on this topic.

Considering the automatic extraction of representative features and intensive image classification, Bittner et al. [39] presented a fully convolutional network (FCN), which includes three parallel modules combined at a late stage. The FCN applies three upper inputs, namely, a panchromatic image, RGB image, and nDSM, to combine height and spectral information from different image data. In addition, full-resolution binary building masks are automatically generated, helping propagate fine details from lower to higher layers to obtain accurate building profiles. Peng et al. [40] proposed a dual-branch DP-DCN based on dense connections and FCNs to automatically obtain fine-grained scene parsing maps. The network prevents both gradient explosion as the network deepens and overfitting with scarce labeled aerial image data. It handles the differences between aerial and natural images efficiently. Zheng et al. [41] proposed a gather-to-guide network to improve the fusion of RGB and auxiliary aerial image data. The key part of this model is a gather-to-guide block, including a guider and a gatherer. The gatherer captures complementary cues from the RGB and auxiliary data and generates cross-source descriptors. The guider uses guide weights extracted from the descriptor to calibrate RGB by reducing redundancy and noise while retaining feature data. However, during final feature mapping fusion, the final semantic prediction output is directly obtained by simple addition and not combined with the differences of the two kinds of image data, thereby introducing too much noise and affecting the segmentation effect. In contrast, the proposed MISNet uses feature extraction and optimized filtering to ensure proper multisource interactions and cross-layer continuity; it establishes dependencies between global information and cross-source features for refining multisource complementary features.

III. Proposed MISNet

A. Overview

The encoder–decoder framework of MISNet is shown in Fig. 1. We use an IRRG image and the corresponding nDSM image as inputs and Res2Net-50 instead of the conventional ResNet as the backbone. In Res2Net, the hierarchical residual connection in a single residual block enables a change in the receptive field at a finer granularity to capture both details and global characteristics [42]. We remove the pooling operation and all the fully connected layers of the original Res2Net, thereby increasing the network performance. We use Res2Net pretrained on the ImageNet dataset [43].

We extract five multilevel features, R_i ($i = 1, 2, 3, 4, 5$) and D_i ($i = 1, 2, 3, 4, 5$), in source-specific encoders for the IRRG and nDSM images, respectively. The input resolution of the source-specific encoder is $W \times H$. Thus, $H/4 \times W/4$ is the resolution for the first and second layers, and $H/2^m \times W/2^m$ is the resolution for layer $m > 2$. In addition, the number of channels of the features in the ith layer is given by C_i ($i = 1, 2, 3, 4, 5$) and $C = [64, 256, 512, 1024, 2048]$.

For the encoder, we use the FFOM to fuse the two kinds of image data (i.e., IRRG and nDSM images) features of each layer. Then, the MCAM extracts multiscale context features. For the decoder, the MCAM extracts global information for guidance, and the DDM gradually integrates high-level features added to the global cues. Moreover, we add the output of the last DDM as an auxiliary output, and the JLM further refines the initial fused feature to obtain the final scene parsing map. For prediction, we discard the first-layer fused features because they are noisy and may undermine segmentation.

B. Feature Fusion Optimization Module

There are two main problems related to the fusion of the IRRG spectral and nDSM elevation features. One is the inherent morphological differences that cause feature incompatibility, and the other is the noise and redundancy in low-quality elevation data. Inspired by the method in [44], we introduce the FFOM to optimize the compatibility of multisource features and mine spatial information from elevation features. As shown in Fig. 1, the FFOM includes two parts that are detailed in the following.

Feature Extraction and Optimized Filtering: In MISNet, we apply feature extraction and optimized filtering to the multisource data at each layer, as shown in Fig. 1. Let features R_i and D_i represent feature maps of the ith layer ($i = 1, 2, 3, 4, 5$) from the IRRG and nDSM branches, respectively. First, R_i and D_i are simply fused by elementwise addition. Then, we extract shared information using elementwise multiplication to highlight the similarity between the two kinds of image data, focusing on the common elements between the original features R_i and D_i and the fused feature [45], and thus, obtaining R_D^i ($i = 1, 2, 3, 4, 5$) and D_R^i ($i = 1, 2, 3, 4, 5$). The corresponding formulation is represented as follows:

$$R_D^i = R_i \times (R_i + D_i) \tag{1}$$

$$D_R^i = D_i \times (R_i + D_i) \tag{2}$$

where $\times$ and $+$ represent elementwise multiplication and addition, respectively.

The optimized features R_D^i and D_R^i are concatenated along the channel dimension, but simple concatenation causes redundancy. Thus, we use two gating units for feature filtering and purification. The gate unit is depicted in Fig. 1(a) and includes the following parts: one 3×3 convolution and one 1×1 convolution connected in series with a rectified linear unit (ReLU) operation activated by a sigmoid function after passing through batch normalization (BN). G_{RD}^i ($i = 1, 2, 3, 4, 5$) denotes the outputs of the gate unit that are multiplied by the original features R_i and D_i. The gate unit is described as follows:

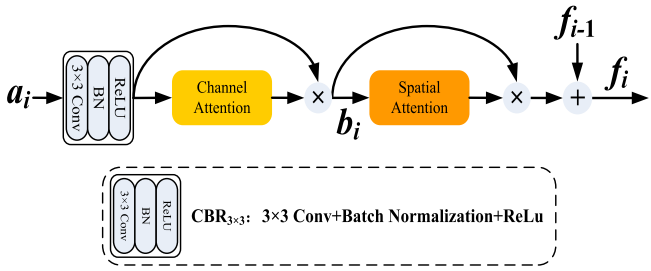

Fig. 2. Architecture of FGM.

$$Gate(x) = \sigma(BN(Conv_{1\times1}(RELU(Conv_{3\times3}(x))))) \tag{3}$$

where $Conv_{n\times n}$ represents a convolution operation with an $n \times n$ kernel, *RELU* denotes ReLU activation, *BN* represents batch normalization, and σ represents the sigmoid function. Hence, G_{RD}^i can be obtained as follows:

$$G_{RD}^i = Gate(Cat(R_D^i, D_R^i)) \tag{4}$$

where *Cat* represents the channelwise concatenation and *Gate* represents the gate unit.

Finally, we purify and filter the original information through pixel-level multiplication and concatenate the two purified features to obtain a_i ($i = 1, 2, 3, 4, 5$) as follows:

$$a_i = Cat((G_{RD}^i \times R_i), (G_{RD}^i \times D_i)). \tag{5}$$

Feature Guidance Module (FGM): Inspired by the method in [46] and [47], we introduce the FGM, as shown in Fig. 2, to optimize the compatibility of multisource features. The FGM contains channel and space attention mechanisms. Channel attention uses the relation between feature channels, whereas spatial attention aims to find locations with informative cues.

We apply the FGM to extract the features, optimize filtering, and optimize the fused feature a_i along the channel and spatial dimensions. The FGM is primarily divided into two parts. The first part uses channel attention to perform weighted optimization along the channel dimension of a_i. The second part uses spatial attention to perform weighting along the spatial dimension of the features after channel optimization. Let *CA* and *SA* denote the channel and spatial attention mechanisms, respectively. In addition, we define the convolution block $CBR_{n\times n}$ that includes a convolutional layer with $n \times n$ kernel, BN, and ReLU activation, as follows:

$$CBR_{n\times n}(x) = RELU(BN(Conv_{n\times n}(x))). \tag{6}$$

Hence, the feature b_i ($i = 1, 2, 3, 4, 5$) after channel attention adjustment can be formulated as follows:

$$b_i = CA(CBR_{3\times3}(a_i)) \times CBR_{3\times3}(a_i). \tag{7}$$

The fused feature output is denoted as f_i and given by

$$f_i = \begin{cases} (b_i \times SA(b_i)), i = 1 \\ (b_i \times SA(b_i)) + f_{i-1}, i = 2 \\ (b_i \times SA(b_i)) + CBR_{3\times3}(Avg(f_{i-1})), i = 3, 4, 5 \end{cases} \tag{8}$$

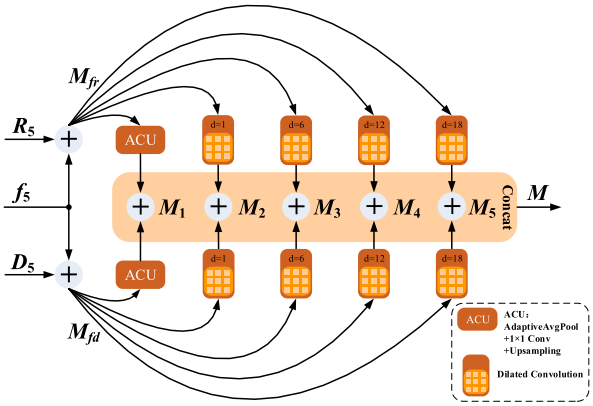

Fig. 3. Architecture of MCAM.

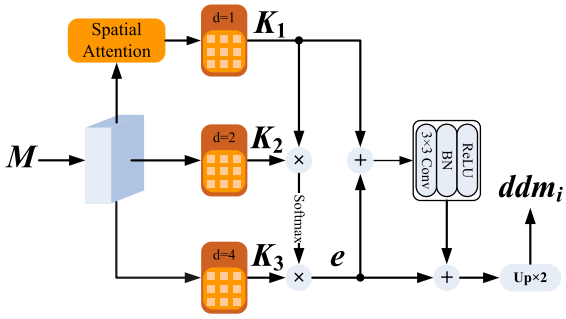

Fig. 4. Architecture of DDM.

where *Avg* represents adaptive average pooling.

C. Multiscale Context Aggregation Module

Considering cross-source fused features, we capture the cross-source complementary information by selectively fusing single-source features using the MCAM (see Fig. 3). Inspired by the method in [48], we obtain multiscale single-source features by adding the fused features of the fifth layer and the two kinds of single-source features. Then, multiscale single-source features with different scales are fused to extract multiscale cross-source features. First, we add the single-source features and fused features to obtain multiscale single-source features M_{fr} *and* M_{fd} as follows:

$$M_{fr} = f_5 + R_5 \tag{9}$$

$$M_{fd} = f_5 + D_5. \tag{10}$$

Then, features M_{fr} and M_{fd} are operated with dilated convolutions at different scales and added together. The multiscale fusion results are concatenated, and the MCAM is formulated as follows:

$$M_1 = ACU(M_{fr}) + ACU(M_{fd}) \tag{11}$$

$$M_2 = Dconv_1(M_{fr}) + Dconv_1(M_{fd}) \tag{12}$$

$$M_3 = Dconv_6(M_{fr}) + Dconv_6(M_{fd}) \tag{13}$$

$$M_4 = Dconv_{12}(M_{fr}) + Dconv_{12}(M_{fd}) \tag{14}$$

$$M_5 = Dconv_{18}(M_{fr}) + Dconv_{18}(M_{fd}) \tag{15}$$

$$M = Cat(M_1, M_2, M_3, M_4, M_5) \tag{16}$$

where M_i (i = 1, 2, 3, 4, 5) represents summation at scale i, *ACU* represents the adaptive average pooling followed by 1×1 convolution and upsampling, $Dconv_i$ represents dilated convolution with dilation rate i, and M represents the MCAM output.

D. Dense Decoding Module

Recent scene parsing methods use encoder–decoder architecture to generate pixel-level predictions. For instance, in the decoder, DeconvNet [49] uses stacked deconvolution operations to gradually recover a full-resolution prediction. SegNet [21] used indices in the encoder pooling block to guide the recovery of image resolution in the decoder, and DeepLabV3+ [22] implements a cascaded decoder. Although decoder improvement is being pursued, the limitations on the resolution of fused features and difficulties of feature aggregation are challenging to address. In addition, fusing low and high-level features and resolving resolution limitations between them are difficult problems. High-level features containing rich semantic cues allow object location and background noise elimination, but they lack details such as object contour and texture. Conversely, low-level features can capture spatial details, but are noisy and unsuitable for accurate segmentation.

We propose the DDM with the architecture shown in Fig. 4. It aims to provide a multiscale receptive field with cascaded dilated convolution to handle changes in the object scale during decoding. The DDM only fuses the features of adjacent layers, avoiding interference caused by large resolution differences. Furthermore, it extracts complementary information to enhance multiscale multilevel features.

The output of the MCAM is the input to the first DDM, which proceeds as follows. First, as concatenation causes redundancy in the feature space, we use the spatial attention module to optimize the hierarchical features along the spatial dimension. Then, we introduce three cascaded dilated convolutions to expand the receptive field for K_1, K_2, and K_3, and multiply the results layer by layer to extract common elements, obtaining e. Subsequently, a residual connection and upsampling provide the DDM output, ddm_i (i = 1, 2, 3). The input of the first DDM is M (i.e., the MCAM output), and the input of the remaining DDMs is the output of the previous decoder block, $Y_{i\text{-}1}$ (i = 2, 3). For convenience, let x represent the DDM input. The DDM is formulated as follows:

$$K_1 = Dconv_1(SA(x)) \tag{17}$$

$$K_2 = Dconv_2(x) \tag{18}$$

$$K_3 = Dconv_4(x) \tag{19}$$

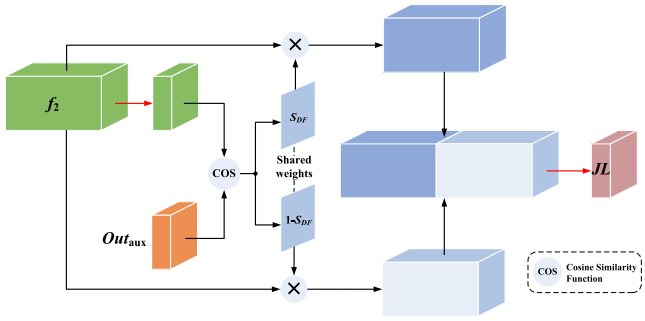

Fig. 5. Architecture of JLM.

$$e = \theta(K_1 \times K_2) \times K_3 \tag{20}$$

$$ddm_i = Up_2(CBR_{3\times3}(K_1 + e) + e) \tag{21}$$

where θ represents the softmax function and Up_n represents n times upsampling.

Output Y_1 of the first decoder block is given by

$$Y_1 = Cat(ddm_1(x), f_4) + ddm_1(x) \tag{22}$$

where ddm_1 represents the first DDM output. The remaining decoder blocks are formulated as follows:

$$Y_i = Cat(ddm_i(Y_{i-1}), f_{5-i}) + ddm_i(Y_{i-1}) i = 2, 3 \tag{23}$$

where Y_i represents the output of the ith decoder block and ddm_i represents the ith DDM output. Moreover, we get our auxiliary output Out_{aux} as follows:

$$Out_{aux} = Conv_{1\times1}(Y_3). \tag{24}$$

E. Joint Learning Module

To suppress noise in low-level detailed features, we present the JLM based on cosine similarity, as shown in Fig. 5. First, we calculate the cosine similarity of the auxiliary output and the low-level fusion feature to segment the redundancy of detail information, and the similarity weight S_{DF} obtained is as follows:

$$S_{DF} = Cos(Conv_{1\times1}(f_2), Out_{aux}) \tag{25}$$

where Cos denotes the cosine similarity function. We then use a pair of reverse similarity weights to denoise and enhance the cross-source feature maps of the lower levels as follows:

$$JL = Conv_{1\times1}(Cat((f_2 \times S_{DF}), (f_2 \times (1 - S_{DF})))). \tag{26}$$

In the JLM, the S_{DF} calculates the similarity between the Out_{aux} and f_2, whereas the reverse weight $(1 - S_{DF})$ measures the difference between them. We found that the low-level fusion feature contains redundant information or interference noise in the mixed details. Hence, the reverse weight can be multiplied by the low-level cross-source feature to suppress noise in the details. The features refined by the JLM are denoted as JL. Out_{aux} and JL are combined to obtain the final full-resolution prediction map Out_{prim} after upsampling four times using a convolution operation with a 3×3 kernel as follows:

$$Out_{prim} = Conv_{3\times3}(Up_4(JL \times Out_{aux} + Out_{aux})). \tag{27}$$

F. Loss Function

The loss function used by MISNet is the most extensively applied cross-entropy loss function L, used to supervise the auxiliary and final outputs

$$L = -\sum_{i}^{Q} V_i \times \log(P_i), i = 0, 1, 2, 3, 4, 5 \tag{28}$$

where V_i is used to indicate whether the predicted class is consistent with the sample class and Q denotes the number of classes; if so, it is 1; otherwise, it is 0. P_i denotes the predicted probability that the sample belongs to class i. The total loss functions include the primary loss function L_{prim} and the auxiliary loss function L_{aux}, which jointly supervise the model output

$$L_{total} = (L_{prim} + L_{aux})/2. \tag{29}$$

IV. Experimental Results and Analysis

This section describes the experimental setup and results of quantitative and qualitative evaluations and an ablation study.

A. Datasets

The Vaihingen and Potsdam benchmark datasets from the ISPRS 2-D Semantic Labeling Contest [40] were used to validate the MISNet. These benchmark datasets contain six scene parsing classes: buildings (blue), impervious surfaces (white), clutter/background (red), low vegetation (cyan), trees (green), and cars (yellow). The Vaihingen benchmark dataset contains 33 VHR images showing 9 cm/pixel in images with a 2500 $\times$ 2000 resolution. Each HRAI shows the IRRG bands and corresponding nDSMs [18]. As the dataset is publicly available, we considered the settings used in the contest. Specifically, 17 ground-truth images were applied as the test set, five HRAIs (ID 11, 15, 28, 30, and 34) were applied as the validation set, and the remaining 11 HRAIs were applied as the training set.

The Potsdam benchmark dataset contains 38 HRAIs showing 5 cm/pixel in images with a 6000 $\times$ 6000 resolution. Each HRAI shows the IRRG bands and corresponding nDSMs. The ground-truth images of the Potsdam dataset were applied as the test set, and the remaining seven HRAIs (ID 2_11, 2_12, 4_10, 5_11, 6_7, 7_8, and 7_10) were applied as the validation set.

B. Performance Measures

To quantitatively verify the effectiveness and robustness of the proposed MISNet, we selected intersection over the union (IoU), mean intersection over the union (mIoU), class accuracy (Acc), and mean class accuracy (mAcc) as the performance measures [50], [51], [52].

C. Implementation Details

All relevant experiments were conducted on the PyTorch framework and a 12 GB NVIDIA TITAN Xp GPU. In the

TABLE I
SCENE PARSING RESULTS ON THE VAIHINGEN DATASET

	Imp.surf		Tree		Low veg.		Building		Car		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
FCN-8S	89.66	79.71	89.22	75.58	75.83	64.33	93.22	86.80	45.12	40.16	78.61	69.32
SegNet	89.88	80.93	88.96	75.96	78.66	64.07	90.88	86.54	43.93	43.16	78.46	70.13
U-Net	91.68	80.90	91.30	77.86	77.97	65.91	89.84	86.50	75.80	71.22	79.75	71.34
HRCNet	91.62	81.60	90.55	78.58	79.24	67.38	91.72	88.01	70.69	68.73	84.76	76.86
DeepLabv3+	90.06	81.11	88.60	76.64	76.65	64.44	87.04	82.7	42.51	43.10	76.97	69.49
RedNet	91.49	84.62	91.41	78.27	78.67	66.59	94.81	91.07	59.77	56.06	83.23	75.32
TSNet	87.93	78.98	**94.26**	**81.26**	71.62	57.03	**95.81**	91.47	67.63	66.86	83.54	75.12
SA-Gate	90.99	**85.70**	89.06	79.15	**84.95**	68.68	93.85	91.72	84.27	78.07	88.62	80.67
ACNet	91.95	85.34	91.20	78.55	78.64	66.87	95.45	91.82	83.12	76.81	88.07	79.88
ESANet	92.09	85.18	92.35	77.65	75.72	65.48	94.93	91.16	75.92	70.11	86.20	77.92
Ours	**92.46**	**85.70**	90.15	79.51	81.66	**68.81**	95.10	**92.05**	**88.77**	**81.86**	**89.63**	**81.59**

TABLE II
SCENE PARSING RESULTS ON THE POTSDAM DATASET

	Imp.surf		Tree		Low veg.		Building		Car		Cluter		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
FCN-8S	89.47	79.77	82.86	71.23	85.13	71.12	90.69	83.60	91.02	81.53	49.05	36.49	78.61	69.32
SegNet	90.18	80.46	82.49	70.68	85.88	71.63	90.21	84.18	93.15	89.72	51.76	37.21	82.28	72.31
U-Net	90.03	80.27	84.06	72.05	85.82	71.60	88.71	82.92	93.89	90.24	50.30	36.26	82.13	72.22
HRCNet	90.03	81.68	82.02	71.32	88.17	73.18	90.87	85.75	93.94	89.82	56.72	40.03	83.63	73.63
DeepLabv3+	91.57	82.49	85.45	73.32	87.36	73.63	91.78	87.59	93.89	90.04	53.80	**43.54**	83.97	75.10
RedNet	**92.19**	82.83	83.00	71.77	87.00	73.22	93.61	**90.13**	93.36	90.08	56.74	43.51	84.32	75.26
TSNet	85.22	76.85	78.75	67.49	**88.52**	67.98	91.85	86.65	78.22	76.85	37.49	30.85	76.68	67.78
SA-Gate	85.84	80.64	85.70	72.89	86.46	72.71	93.65	88.51	92.18	89.39	**62.70**	40.59	84.42	74.12
ACNet	91.32	82.74	86.03	72.87	86.16	73.53	93.83	90.06	93.79	90.43	54.51	41.65	84.27	75.21
ESANet	91.38	**82.92**	82.48	70.81	87.10	73.16	93.69	89.82	93.08	88.53	55.68	43.38	83.90	74.77
Ours	91.10	82.50	**87.88**	**74.90**	85.84	**74.11**	**93.86**	89.73	**94.70**	**90.53**	55.04	42.10	**84.73**	**75.64**

experiment, the weights of the encoder of the proposed MISNet used Res2Net-50 [42] to initialize. The input size of the proposed method is 256 × 256. The ground truth and input images are then enhanced by applying cropping and random scaling, a counterclockwise 90°, 180°, 270° rotation, and vertical and horizontal flipping [53], [54], [55].

During the training stage, we apply the Adam method for optimization. Its initial learning rate is 0.0001, weight decay is 0.0005, and momentum parameter is 0.9, which is reduced to 0.1 times of the original every 20 epochs. It took approximately 100 epochs for the proposed MISNet to converge.

D. Comparisons With SOTA Models

The proposed MISNet was compared with 10 SOTA scene parsing models: FCN [19], U-Net [20], DeepLabV3+ [22], SegNet [21], RedNet [31], HRCNet [23], ACNet [32], SA-Gate [33], ESANet [35], and TSNet [34]. For a fair comparison, the scene parsing results were retrieved from the original papers or generated by running the available source codes. FCN, UNet, SegNet, DeepLabV3+, and HRCNet belong to single-source scene parsing, and the rest belong to RGB-D scene parsing. We chose these models for comparison, given the scarcity of scene parsing models for HRAIs and because existing models do not have publicly available codes.

Quantitative Experiments: Tables I and II list the obtained Acc, mAcc, IoU, and mIoU for the evaluated methods. The measurements for the proposed MISNet in the multiclass labels indicate excellent performance, demonstrating the effectiveness of our proposal. In general, single-source scene parsing is inferior to the multisource approach. Therefore, auxiliary data nDSM are introduced to capture 3-D spatial information for scene parsing of HRAIs, and the scene parsing performance is significantly improved.

Qualitative Experiments: To further evaluate the excellent performance of our proposed MISNet, we show the results of scene parsing in Fig. 6. Owing to the limitation of space, we chose HRCNet [23] as the typical representative single-source approach and selected all multisource approaches for comparison. The results in lines 4 and 6 in Fig. 6 show that the multisource approach using nDSM data can easily extract staggered buildings; the single-source approach identifies large impervious surface areas as buildings. The results in lines 1 and 7 show that the proposed MISNet is superior to other approaches because the other approaches tend to confuse similar areas, such as low vegetation and trees; moreover, it can easily extract small objects, such as cars that were hidden under the shadows of buildings. Furthermore, our method of segmentation has a sharper profile, as listed in lines 2 and 3.

E. Ablation Studies

We performed ablation studies on the Potsdam and Vaihingen benchmark datasets to investigate the contribution of the different modules to this approach, as listed in Table III.

1) Effectiveness of FFOM: Based on the backbone network, the FFOM is added to the encoding part. The DDM and JLM are replaced by a simple convolution operation and upsampling

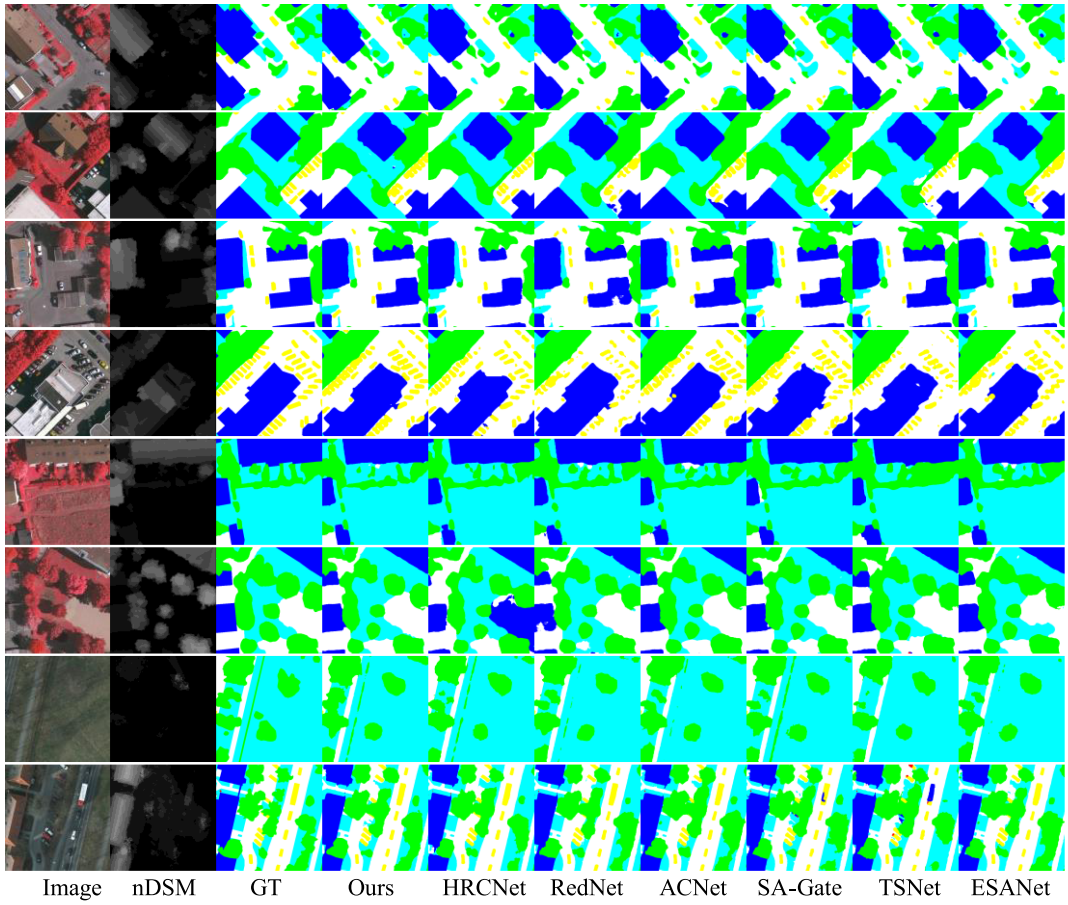

Fig. 6. Scene parsing maps obtained from the MISNet and SOTA approaches.

TABLE III
RESULTS OF ABLATION STUDIES ON MISNET

Methods	Vaihingen		Potsdam	
	mAcc	mIoU	mAcc	mIoU
Backbone	86.48	78.20	83.52	74.21
Backbone + FFOM	86.68	78.58	83.70	74.39
Backbone + MCAM	87.63	79.13	84.18	74.95
Backbone + DDM	87.82	80.03	84.10	74.64
Backbone + JLM	87.85	79.43	84.45	74.51
Ours	**89.63**	**81.59**	**84.73**	**75.64**

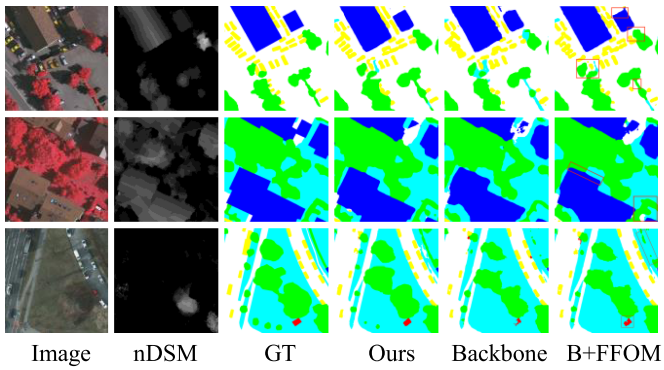

Fig. 7. Scene parsing maps obtained from backbone, backbone add FFOM(B+FFOM) and complete MISNet.

operation in the decoding part, where the convolution kernel is 1×1. It is used to adjust the channel number to discard the interactivity between cross-layer features, so that decoding blocks and fusion features are concatenated in series layer by layer, and the prediction graph is finally obtained. The cyan low vegetation and green trees in the forecast diagram of the scenario analysis shown in Fig. 7 are similar in color. Without the FFOM to fuse IRRG and nDSM, incorrectly classifying trees as low vegetation would be easy. In addition, in the hyperspectral image, the top area of the building resembles the shaded area. The addition of the FFOM can make the contour of the building flat and smooth.

2) Effectiveness of MCAM: Based on the backbone network, we use simple addition instead of the FFOM in the encoder to merge the two kinds of image features and retain the MCAM. In the decoder, we use convolution and upsampling instead of the DDM and JLM to obtain the final output. Fig. 8 shows that if the global multiscale context information processing of

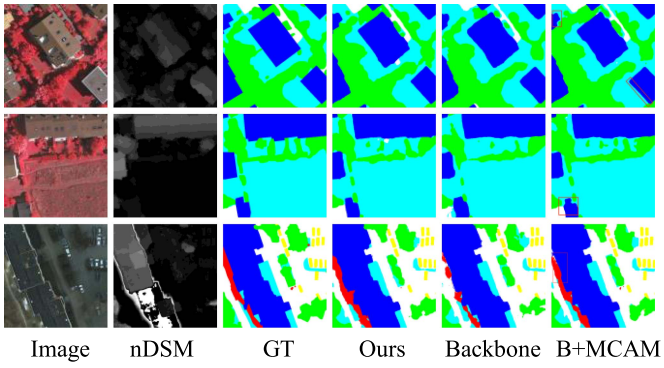

Fig. 8. Scene parsing maps obtained from backbone, backbone add MCAM(B+MCAM) and complete MISNet.

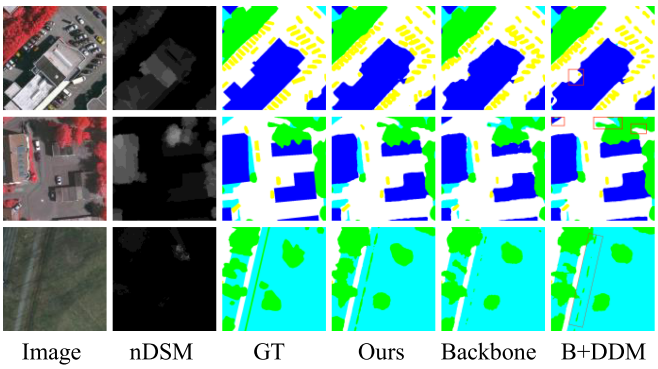

Fig. 9. Scene parsing maps obtained from backbone, backbone add DDM(B+DDM) and complete MISNet.

the MCAM is omitted, objects of different scales cannot be completely divided, for example, details such as the zigzag texture of a building cannot be accurately divided. Further, with the addition of the MCAM, some areas of the impervious surface can be segmented. As shown in Table III, the MCAM facilitates accurate scenario resolution.

3) Effectiveness of DDM: Based on the backbone network, we replace the FFOM with simple element addition at the corresponding nDSM and IRRG flow level, add the DDM in the decoding part, retain the decoding operation in the original model, and eliminate the JLM at the same time. Fig. 9 shows that the segmentation effect becomes significantly worse after the DDM is removed. We use the DDM to integrate and refine multiscale features from top to bottom, solving segmentation errors and incomplete problems, such as confusion of low vegetation and trees and unclear segmentation of jagged textures of buildings.

4) Effectiveness of JLM: Based on the backbone network, the FFOM is replaced by simple element addition in the encoding part; 1×1 convolution and upsampling are used in the decoding part to replace the original DDM, and the JLM and its related operations are retained. Fig. 10 shows that without the addition of the JLM, the area of the trees and impervious surfaces could not be completely and correctly segmented, and segmentation errors occurred, such as impervious surfaces and buildings. The

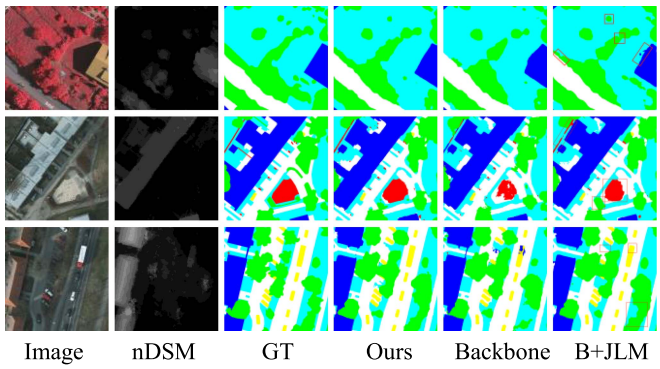

Fig. 10. Scene parsing maps obtained from backbone, backbone add JLM(B+JLM) and complete MISNet.

JLM can be seen to improve the integrity and correctness of segmentation.

V. Conclusion

We propose the MISNet, a cross-source interaction model that exploits the dependence between two kinds of aerial image data in different convolution layers. The FFOM performs cross-source fusion and guides interactive features layer by layer. The MCAM establishes a transition between the encoder and decoder, and the DDM refines and denoises details in low-level contextual features based on high-level semantics. To further improve the MISNet, the JLM based on similarity learning is introduced, and auxiliary supervision is added to optimize the scene parsing performance. Experimental results indicate that the proposed MISNet is superior to 10 SOTA approaches in terms of various evaluation measures on two benchmark datasets.

References

[1] W. Zhou, J. Jin, J. Lei, and J.-N. Hwang, "CEGFNet: Common extraction and gate fusion network for scene parsing of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5405110.

[2] 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.

[3] J. Jin, W. Zhou, R. Yang, L. Ye, and L. Yu, "Edge detection guide network for semantic segmentation of remote-sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Jan. 2023, Art. no. 5000505, doi: 10.1109/LGRS.2023.3234257.

[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. 2021.

[5] J. Ma, W. Zhou, J. Lei, and L. Yu, "Adjacent bi-hierarchical network for scene parsing of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, doi: 10.1109/LGRS.2023.3241648.

[6] W. Zhou, C. Liu, J. Lei, L. Yu, and T. Luo, "HFNet: Hierarchical feedback network with multilevel atrous spatial pyramid pooling for RGB-D saliency detection," *Neurocomputing*, vol. 490, pp. 347–357, 2022.

[7] W. Zhou, C. Liu, J. Lei, and L. Yu, "RLLNet: A lightweight remaking learning network for saliency redetection on RGB-D images," *Sci. China Inf. Sci.*, vol. 65, no. 6, 2022, Art. no. 160107.

[8] 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. Photogrammetry Remote Sens.*, vol. 167, pp. 12–23, Sep. 2020.

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

[10] Y. Cai, W. Zhou, L. Zhang, L. Yu, and T. Luo, "DHFNet: Dual-decoding hierarchical fusion network for RGB-thermal semantic segmentation," *Vis. Comput.*, 2023, doi: 10.1007/s00371-023-02773-6.

[11] T. Gong, W. Zhou, X. Qian, J. Lei, and L. Yu, "Global contextually guided lightweight network for RGB-thermal urban scene understanding," *Eng. Appl. Artif. Intell.*, vol. 117, 2023, Art. no. 105510.

[12] J. Wu, W. Zhou, X. Qian, J. Lei, L. Yu, and T. Luo, "MFENet: Multitype fusion and enhancement network for detecting salient objects in RGB-T images," *Digit. Signal Process.*, vol. 133, 2023, Art. no. 103827.

[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] W. Zhou, J. Jin, J. Lei, and L. Yu, "CIMFNet: Cross-layer interaction and multiscale fusion network for semantic segmentation of high-resolution remote sensing images," *IEEE J. Sel.Topics Signal Process.*, vol. 16, no. 4, pp. 666–676, Jun. 2022.

[15] W. Zhou, E. Yang, J. Lei, and L. Yu, "FRNet: Feature reconstruction network for RGB-D indoor scene parsing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 4, pp. 677–687, Jun. 2022.

[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. 1, Dec. 2017, Art. no. 52.

[17] W. Zhou, Y. Yue, M. Fang, X. Qian, R. Yang, and L. Yu, "BCINet: Bilateral cross-modal interaction network for indoor scene understanding in RGB-D images," *Inf. Fusion*, vol. 94, pp. 32–42, 2023.

[18] Y. Liu, D. M. Nguyen, N. Deligiannis, W. Ding, and A. Munteanu, "Hourglass-shapenetwork based semantic segmentation for high resolution aerial imagery," *Remote Sens.*, vol. 9, no. 6, May 2017, Art. no. 522.

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

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

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

[22] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[23] Z. Xu, W. Zhang, T. Zhang, and J. Li, "HRCNet: High-resolution context extraction network for semantic segmentation of remote sensing images," *Remote Sens.*, vol. 13, no. 1, Dec. 2020, Art. no. 71.

[24] L. Mou, Y. Hua, and X. X. Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7557–7569, Nov. 2020.

[25] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 181, pp. 84–98, Nov. 2021.

[26] Q. Zhao, J. Liu, Y. Li, and H. Zhang, "Semantic segmentation with attention mechanism for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5403913.

[27] D. Zhao, C. Wang, Y. Gao, Z. Shi, and F. Xie, "Semantic segmentation of remote sensing image based on regional self-attention mechanism," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8010305.

[28] C. Couprie, C. Farabet, L. Najman, and Y. LeCun, "Indoor semantic segmentation using depth information," 2013, *arXiv:1301.3572*.

[29] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 345–360.

[30] D. Lin, G. Chen, D. Cohen-Or, P. A. Heng, and H. Huang, "Cascaded feature network for semantic segmentation of RGB-D images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1311–1319.

[31] J. Jiang, L. Zheng, F. Luo, and Z. Zhang, "RedNet: Residual encoder-decoder network for indoor RGB-D semantic segmentation," 2018, *arXiv:1806.01054*.

[32] X. Hu, K. Yang, L. Fei, and K. Wang, "ACNet: Attention based network to exploit complementary features for RGBD semantic segmentation," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1440–1444.

[33] X. Chen et al., "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 561–577.

[34] W. Zhou, J. Yuan, J. Lei, and T. Luo, "TSNet: Three-stream self-attention network for RGB-D indoor semantic segmentation," *IEEE Intell. Syst.*, vol. 36, no. 44, pp. 73–78, Jul./Aug. 2021.

[35] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, and H. M. Gross, "Efficient RGB-D semantic segmentation for indoor scene analysis," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2021, pp. 13525–13531.

[36] Y. Qian, L. Deng, T. Li, C. Wang, and M. Yang, "Gated-residual block for semantic segmentation using RGB-D data," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 11836–11844, Aug. 2022.

[37] H. Zhou, L. Qi, H. Huang, X. Yang, Z. Wan, and X. Wen, "CANet: Co-attention network for RGB-D semantic segmentation," *Pattern Recognit.*, vol. 124, Apr. 2022, Art. no. 108468.

[38] W. Zhou, E. Yang, J. Lei, J. Wan, and L. Yu, "PGDENet: Progressive guided fusion and depth enhancement network for RGB-D indoor scene parsing," *IEEE Trans. Multimedia*, to be published, doi: 10.1109/TMM.2022.3161852.

[39] K. Bittner, F. Adam, S. Cui, M. Körner, and P. Reinartz, "Building footprint extraction from VHR remote sensing images combined with normalized DSMs using fused fully convolutional networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2615–2629, Aug. 2018.

[40] C. Peng, Y. Li, L. Jiao, Y. Chen, and R. Shang, "Densely based multi-scale and multi-modal fully convolutional networks for high-resolution remote-sensing image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2612–2626, Aug. 2019.

[41] X. Zheng, X. Wu, L. Huan, W. He, and H. Zhang, "A gather-to-guide network for remote sensing semantic segmentation of RGB and auxiliary image," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5404915.

[42] W. Zhou, Y. Zhu, J. Lei, J. Wan, and L. Yu, "CCAFNet: Crossflow and cross-scale adaptive fusion network for detecting salient objects in RGB-D images," *IEEE Trans. Multimedia*, vol. 24, pp. 2192–2204, 2022.

[43] O. Russakovsky et al., "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Apr. 2015.

[44] L. Deng, M. Yang, T. Li, Y. He, and C. Wang, "RFBNet: Deep multimodal networks with residual fusion blocks for RGB-D semantic segmentation," 2019, *arXiv: 1907.00135*.

[45] W. Zhou, Q. Guo, J. Lei, L. Yu, and J.-N. Hwang, "IRFR-Net: Interactive recursive feature-reshaping network for detecting salient objects in RGB-D images," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Aug. 20, 2021, doi: 10.1109/TNNLS.2021.3105484.

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

[47] J. Wu, W. Zhou, X. Qian, J. Lei, L. Yu, and T. Luo, "MENet: Lightweight multimodality enhancement network for detecting salient objects in RGB-Thermal images," *Neurocomputing*, vol. 527, pp. 119–129, 2023.

[48] N. Huang, Y. Luo, Q. Zhang, and J. Han, "Discriminative unimodal feature selection and fusion for RGB-D salient object detection," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108359.

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

[50] W. Zhou, J. Liu, J. Lei, J.-N. Hwang, and L. Yu, "GMNet: Graded-feature multilabel-learning network for RGB-thermal urban scene semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 7790–7802, Sep. 2021.

[51] W. Zhou and J. Hong, "FHENet: Lightweight feature hierarchical exploration network for real-time rail surface defect inspection in RGB-D images," *IEEE Trans. Instrum. Meas.*, vol. 72, Jan. 2023, Art. no. 5005008.

[52] W. Zhou, Y. Lv, J. Lei, and L. Yu, "Embedded control gate fusion and attention residual learning for RGB–thermal urban scene parsing," *IEEE Trans. Intell. Transp. Syst.*, to be published, doi: 10.1109/TITS.2023.3242651.

[53] W. Zhou, Y. Zhu, J. Lei, R. Yang, and L. Yu, "LSNet: Lightweight spatial boosting network for detecting salient objects in RGB-thermal images," *IEEE Trans. Image Process.*, to be published, doi: 10.1109/TIP.2023.3242775.

[54] W. Zhou, L. Yu, Y. Zhou, W. Qiu, M.-W. Wu, and T. Luo, "Local and global feature learning for blind quality evaluation of screen content and natural scene images," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2086–2095, May 2018.

[55] W. Zhou, Q. Guo, J. Lei, L. Yu, and J.-N. Hwang, "ECFFNet: Effective and consistent feature fusion network for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1224–1235, Mar. 2022.