Building Detection From Panchromatic and Multispectral Images With Dual-Stream Asymmetric Fusion Networks

Ziyue Huang, Qingjie Liu, *Member, IEEE*, Huanyu Zhou, Guangshuai Gao, Tao Xu, Qi Wen, and Yunhong Wang, *Fellow, IEEE*

***Abstract*—Building detection from panchromatic (PAN) and multispectral (MS) images is an essential task for many practical applications. In this article, a dual-stream asymmetric fusion network is proposed, named DAFNet. DAFNet can achieve effective information fusion at the feature level. It obtains better building detection performance from the following three perspectives: a two-stream network structure is designed to guarantee the ability to extract information from PAN and MS images; an asymmetric feature fusion module is proposed to fuse features efficiently and concisely; and two consistency regularization losses, i.e., PAN information preservation loss and cross-modal semantic consistency loss are applied to further explore the consistency between features for better fusion. The experiments are conducted on a challenging building detection dataset collected from GaoFen-2 satellite images. Comprehensive evaluations on 12 popular detection methods demonstrate the superiority of our DAFNet compared with the existing state-of-the-art fusion methods. We reveal that feature-level fusion is more suitable for building detection from PAN-MS images.**

***Index Terms*—Building detection, deep learning, multimodal fusion, remote sensing (RS) images.**

I. Introduction

BUILDING detection from remote sensing (RS) images is an important research topic since it provides basic information for a wide range of applications such as urban planning [1], earthquake disaster reduction [2], and mapping [3]. Thanks to powerful deep neural networks, research in this field has been advancing rapidly in recent years. Many of these methods draw inspiration from generic object detection approaches that aim to detect everyday objects in natural scenes.

Manuscript received 6 January 2023; revised 23 February 2023; accepted 21 March 2023. Date of publication 27 March 2023; date of current version 7 April 2023. This work was supported by the National Natural Science Foundation of China under Grant 62176017 and Grant 41871283. *(Corresponding author: Qingjie Liu.)*

Ziyue Huang, Qingjie Liu, Huanyu Zhou, Guangshuai Gao, and Yunhong Wang are with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China, and also with the Hangzhou Innovation Institute, Beihang University, Hangzhou 310051, China (e-mail: ziyuehuang@buaa.edu.cn; qingjie.liu@buaa.edu.cn; zhysora@buaa.edu.cn; gaoguangshuai1990@buaa.edu.cn; yhwang@buaa.edu.cn).

Tao Xu is with the School of Information Science and Engineering, University of Jinan, Jinan 250022, China (e-mail: xutao@ujn.edu.cn).

Qi Wen is with the National Disaster Reduction Center of China, Beijing 100124, China (e-mail: whistlewen@aliyun.com).

Code is released at https://github.com/floatingstarZ/DAFNet

Digital Object Identifier 10.1109/JSTARS.2023.3261866

However, real-world RS image processing and analysis systems usually accept inputs in two modalities: panchromatic (PAN) and multispectral (MS), because many Earth observation satellites cannot provide images with both high spatial and spectral resolutions. Alternatively, these satellites acquire images in two modalities: PAN images that are rich in spatial information with fine details and textures and MS images with rich spectrum information complementary to the PAN images. To leverage the advantages of both modalities and facilitate subsequent processing steps, researchers have developed pan-sharpening techniques that fuse PAN and MS images to generate high-resolution MS images. The pan-sharpened images have been proven to be able to achieve a fairly good recognition performance [4] because they can preserve the spatial detail and spectral information of the image content [5]. Pan-sharpening-then-understanding has become a standard pipeline for many RS image interpretation systems.

However, our experiments show that constructing RS image interpretation models (e.g., building detection model) based on pan-sharpened images might be suboptimal. Pan-sharpening methods have a significant impact on building detection performance. We test various image fusion methods, including traditional [6], [7], [8], [9] and deep learning ones [10], [11], [12], [13]. Some of them degrade detection methods significantly compared with those using only PAN images. There are two reasons for this, which are as follows:

1) these pan-sharpening methods are not optimized for the downstream tasks, such as building detection, even though they can produce good pan-sharpened images;
2) PAN and MS are considered to have equal contributions for fusion; however, our experiments show that PAN is more important than MS for object detection.

Compared with pan-sharpening methods, feature-level fusion [14], [15], [16], [17], [18], [19] combines the fusion process with the downstream tasks to alleviate the suboptimal problem. Nonetheless, there are still several concerns that must be taken into account while applying feature fusion on building detection.

1) As an effective multiscale feature fusion method, feature pyramid network (FPN) [20] is widely used in object detection. The strategy of combining multimodal fusion with FPN will affect the detection performance. The effective combination strategy remains to be studied.

2) Most current fusion modules fuse features using a symmetrical structure. However, PAN and MS are not equally helpful for building detection. There still needs to be more researches on the asymmetric fusion structure to highlight the information of PAN.
3) Attention mechanisms [14], [15], [16], [17], [18], [19] are widely employed in existing fusion methods, which pushes the model focus on essential parts and obtain the complementary information from multimodal data [14].

However, they ignore the heterogeneous gap [21] existing in multimodal features, i.e., features from different modalities located in unequal subspaces, and will weaken multimodal fusion's benefits. We propose the dual-stream asymmetric fusion (DAF) network to deal with these issues. The commonly used two-stream architecture [22], [23] in multimodal fusion is adopted to extract MS and PAN features. To better adapt to the FPN, we propose dual fusion FPN, which first performs scale fusion, and then, modality fusion. An asymmetric feature fusion (AFF) module and a PAN information preservation (PiP) loss are designed to avoid losing PAN information. Motivated by DCCA [24], a cross-modal semantic consistency (CSC) loss is introduced to alleviate the heterogeneous gap so that the fused feature does not contain noise and is more robust.

In summary, our contributions are as follows.

1) We reveal that models aiming to detect buildings from RS images should be well-designed. Performing detection from fused images may not be a good solution. Detection from joint inputs of PAN and MS images has great potential to be investigated.
2) A dual-stream asymmetric fusion network, termed DAF, is proposed for building detection. DAF takes advantage of the original information of PAN and MS images and fuses them using an AFF module. PAN information preservation (PiP) loss and cross-modal semantic consistency (CSC) loss are proposed to augment building detection further.
3) Experiments demonstrate that the proposed losses and AFF module have strong adaptability that are applicable to various detectors and can boost detectors' performance on building detection without bells and whistles.

II. Related Work

A. Generic Object Detection

Object detection is a fundamental task in computer vision. It has achieved great success thanks to powerful deep neural networks. Most of the existing detectors can be grouped into two families, namely two-stage detectors [25], [26], [27], [28], and single-stage detectors [29], [30], [31], [32], [33]. Two-stage detectors resolve the detection with a two-stage pipeline, in which the first stage generates a set of candidate proposals, and the second stage performs category classification and bounding box regression simultaneously. The second stage can be considered a refinement process. Thus, two-stage detectors generally show high detection performance; however, they always suffer from low inference speed. Single-stage methods discard the proposal generation stage and directly conduct detection from features. These methods are more computationally efficient than two-stage detectors but have lower accuracy.

In order to achieve a better performance, single-stage methods usually place a large number of preset dense anchors over images, and then, predict the final detection boxes by scoring the anchors and estimating relative offsets to them. Anchors play a similar role to proposals, thus enabling detection performance promotion. These methods [25], [29] are widely known as anchor-based detectors. However, to guarantee better performance, there might be more than $10K$ anchors required, which significantly decreases the training and inference speed, and more importantly, results in extremely unbalanced positive and negative samples during training. Anchor-free models (e.g., FSAF [31], FCOS [30]) are proposed to solve these problems. They use the center points or center areas as positive sample areas and directly predict detection boxes and categories in these areas.

B. Building Detection

Benefiting from generic object detectors, detecting ground objects in RS images has been advancing rapidly in recent years. Many methods have emerged and achieved astonishing performance. Among all concerning objects in aerial images, buildings are the most important and challenging ones. Great efforts have been devoted to solving the building detection problem.

Vakalopoulou et al. [34] propose an automatic building detection framework based on deep features and SVM classifiers. Zhang et al. [35] design a coarse-to-fine detection framework, which uses saliency maps to locate built-up regions, followed by an R-CNN [36] like pipeline to detect buildings. Li et al. [37] develop a cascaded network, where they incorporate the Hough transform to highlight the boundaries of buildings. Li et al. [38] design a multibranch network to capture contextual and structural features for better identification of buildings.

Many methods solve building extraction with instance segmentation approaches. Alshehhi et al. [39] address road and building extraction with a single-branch CNN. Hamaguchi et al. [40] address the multiscale problem with a multitask framework. The framework consists of multiple U-Net models. Each model is devoted to a specific size of building. Yang et al. [41] design a dense-attention network for building extraction. The attention mechanism can strengthen features, thereby enabling better performance. Griffiths et al. [42] argue that label quality is critical for model training. They propose to improve building footprint masks using morphological geodesic active contours. Han et al. [43] combine the advantages of traditional image processing methods and deep models. They use traditional methods to enhance the dataset, and then, employ a Mask R-CNN for building detection. Sirko et al. [44] study continental-scale building detection. They also utilize U-Net to segment buildings.

In addition to the perspective of segmentation, some works seek better representations for buildings, e.g., polygon and vector fields. Castrejon et al. [45] cast instance-level building segmentation as a contour polygon prediction task, inspiring more

subsequent building detection works. Li et al. [46] circumvent the conventional pixel-wise segmentation of aerial images and directly predict buildings and roads in a vector representation. They developed a method named PolyMapper to achieve this goal. Wei et al. [47] propose a two-step method. They first introduce an improved fully convolutional network to obtain masks of building footprints, and then, use a polygon regularization algorithm to transfer the masks into polygons. Li et al. [48] propose a hybrid model for building polygon extraction, in which they employ several networks to obtain bounding boxes, segmentation masks, and corners of buildings, and then, use Delaunay triangulation to construct building polygons. Zhu et al. [49] present an adaptive polygon generation algorithm (APGA), which first generates sequences of building vertexes and then arranges them to form polygons.

In real applications, images may come from different platforms, and thus, are with different resolutions. To bridge the resolution gap, Guo et al. [50] adopt a superresolution method to zoom them into the same resolution and perform building segmentation. Chen et al. [51] extract buildings from PAN and MS imagery to fully explore the spatial-spectral information. They propose to use a multiscale spatial-spectral contextual information mining CNN for this goal.

C. Multimodal Fusion

Different modalities captured from the same platform usually carry distinct yet complementary information. Combining them together, such as RGB-Depth [17], RGB-Thermal [17], [52], Audio-Visual [53], RGB-LiDAR [54], and RGB-Radar [53] is believed to enable considerable and consistent perception improvement compared with a single modality. Bin et al. [55], [56] use an adaptive multimodal mechanism in dealing with real-world inverse synthetic aperture radar (ISAR) object recognition problem on the level of feature and decision. Bin et al. [57] proposed deep geometric learning to strengthen the capability of the CNN in multimodal scenarios.

Multimodality fusion can be divided into three categories: early fusion, mid fusion, and late fusion. These fusion strategies happen on pixel level, feature level, and decision level. Early fusion is widely used in RS field, such as pan sharpening [10], [11]. However, pan sharpening is independent of downstream tasks; therefore, it may not be beneficial for interpretation models. Late fusion makes decisions based on the predictions obtained from each modality. However, terrible predictions from one modality are likely to damage the final performance. The mid-fusion strategy has been widely studied. The key to achieving effective multimodal fusion is to filter useless information in each modality and combine the rest. This idea coincides with attention, making attention mechanisms are widely used in multimodal fusion [17], [22], [53], [58], [59], [60].

Satellites usually carry two kinds of sensors providing two modalities: PAN and MS. In addition to combing the strengths of the two modalities using pan-sharpening techniques, researchers also explore interpreting RS images using midfusion strategies. Li et al. [22] design an attention-based heterogeneous gated fusion network to fuse the optical and SAR features for land cover classification. Kang et al. [15] propose a fully convolutional network using a cross-gate module to fuse features from optical and SAR images.

III. Method

A. Overview

The overall pipeline of our method is shown in Fig. 1. Two CNN networks are used to extract features from the input PAN and MS images. Since the modalities carry distinct information, these two networks do not share weights. Then, FPNs [20] are equipped to obtain multiscale features for better detecting buildings on various scales. Finally, AFF module is proposed to fuse features of PAN and MS images. Two loss functions are introduced to enforce fusion: PAN information preservation (PiP) loss and cross-modal semantic consistency (CSC) loss. Our fusion strategy occurs in the feature extraction stage and is independent of detection heads, making it applicable to a variety of detection methods.

B. Feature Extraction

Given a pair of PAN and MS images $\{I^p, I^m\}$, where superscript p denotes PAN and m denotes MS, ResNet50 [61] with unshared parameters are used as the backbones to extract their visual features. During the construction of the dataset, MS images are upsampled by bilinear interpolation to the same size as PAN images, which gives the same resolution to the features obtained by the two branches. ResNet50 is composed of one input block B and four stages: $\{\mathrm{R}_2, \mathrm{R}_3, \mathrm{R}_4, \mathrm{R}_5\}$. ResNet50 accepts three-channel RGB images as inputs, which is inconsistent with PAN and MS images. Modifying the input channel of the input block will destroy the pretrained parameters, whereas selecting only three channels as input will damage the multispectral information [62], [63]. Motivated by the 3-D CNN [64] used in hyperspectral image classification, we develop a sliding strategy to fill this gap. The PAN image is replicated three times and stacked together to form a three-channel input. Then, the new inputs are fed into the PAN branch

$$C_1^p = \mathrm{B}^p\left([I^p, I^p, I^p]\right) \tag{1}$$

where B^p denotes the input block of the PAN branch, and C_1^p is the obtained feature. For one MS image that contains four channels, we slide the input block of ResNet50 along its channel dimension and obtain the features of the input block in the MS branch

$$C_1^m = \sum_{i=0}^{1} \mathrm{B}^m\left(I_{i:i+3}^m\right) \tag{2}$$

where $I_{i:i+3}^m$ represents the ith to the $(i+3)$th channels of the MS image. Through the sliding strategy, the network can process MS images while preserving the pretrained weights.

After the feed-forward propagation, features in a pyramid style $\{(C_i^p, C_i^m)\}_{i=2}^5$ are obtained by

$$C_i = \mathrm{R}_i(C_{i-1}), \quad i = 2, 3, 4, 5 \tag{3}$$

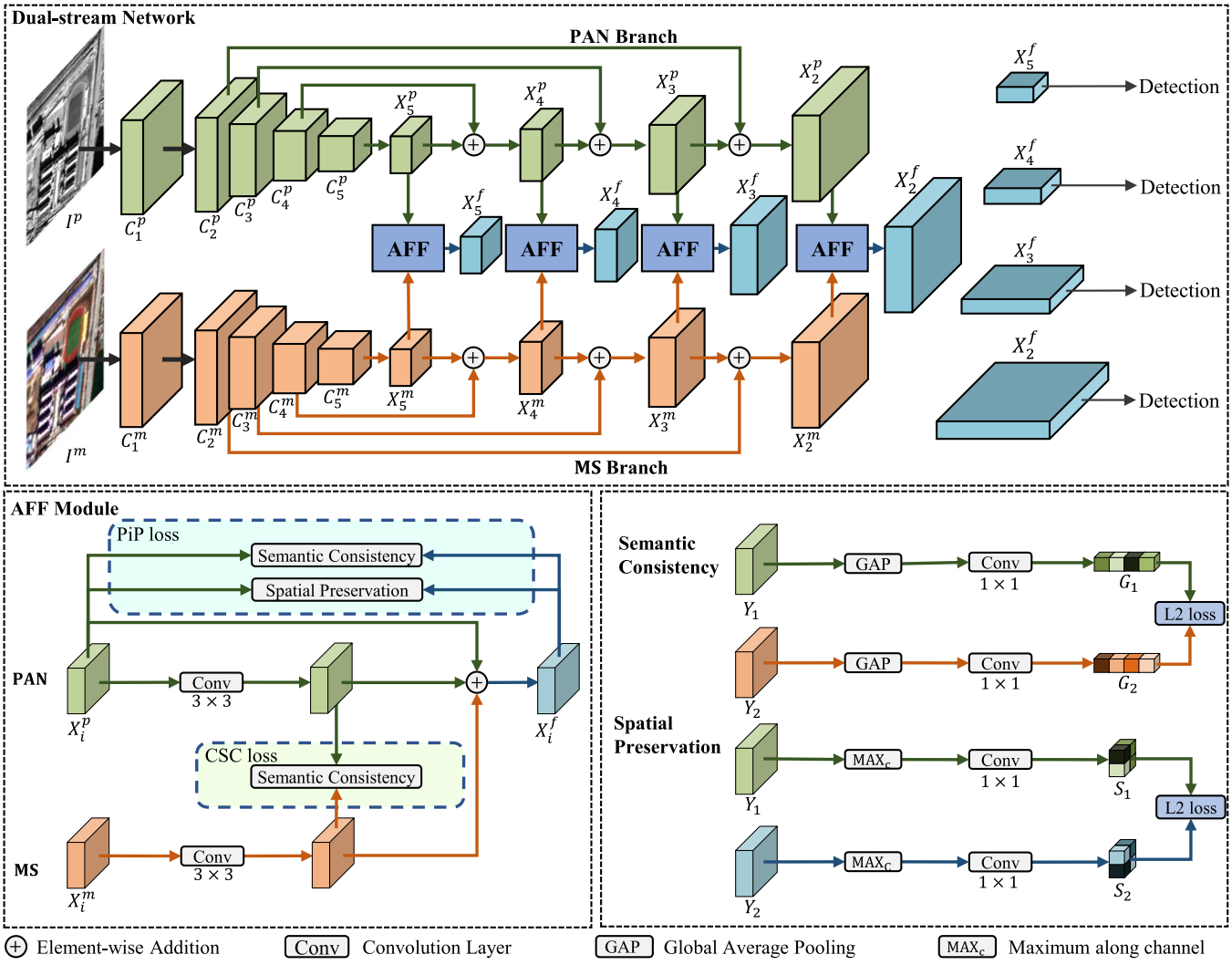

Fig. 1. Overall architecture of our dual-stream asymmetric fusion network (DAFNet). It consists of two networks responsible for extracting features of PAN and MS images, respectively. The DAFNet is equipped with two independent FPNs [20] for each modality to address the multiscale detection problem. The proposed AFF module fuses features in each scale of the two FPNs to combine the strengths of PAN and MS. Two loss functions, i.e., PiP loss and CSC loss are proposed to ensure that the fusion is achievable and the most important information of PAN can be preserved.

each is with channels of {256, 512, 1024, 2048}, respectively. R_i denotes the ith layer of backbone.

Then, features pass through two independent FPNs to obtain multiscale features. The FPN conducts multiscale feature fusion through a top-down pathway with lateral connections [20], which produces the features $\{X_i\}_{i=2}^{5}$, all with channels of 256. Finally, the two pyramidal-style features are fused through AFF modules for detection:

$$X_i^f = \mathrm{AFF}_i\left(X_i^p, X_i^m\right), \quad i = 2, 3, 4, 5 \tag{4}$$

where AFF_i denotes the ith AFF module. Considering that PAN image features account for the dominant role in the detection, a skip connection is added to ease gradients update of the PAN branch, as shown in the AFF module in Fig. 1. It is formulated as

$$\mathrm{AFF}_i\left(X_i^p, X_i^m\right) = \mathrm{Conv}(X_i^p) + \mathrm{Conv}(X_i^m) + X_i^p \tag{5}$$

where Conv is a convolutional layer with kernel size 3×3.

C. Consistency Regularization of the AFF Module

For object detection, each level of the FPN is supervised by the regression loss and classification loss [20] to learn features with semantic and spatial information. The semantic information helps the detector to distinguish objects in each region. The spatial information, such as contours and edges, helps to identify object boundaries [65]. By using global average pooling (GAP) [66], the global-level representations of the whole image could be obtained. The maximum value along channels describes the spatial information to some extent [67].

There are two consistencies essential for fusion and detection.

1) The features of the PAN and MS images should hold a semantic consistency since they are captured over the same site.
2) Both semantic and spatial information of PAN images should be preserved after fusion since PAN images play a decisive role in detection.

Two regularization losses are proposed to achieve these consistencies: semantic consistency loss and spatial preservation loss.

Semantic consistency loss ensures that the semantic information of two inputs is as close as possible. To this end, the global-level representations G_i are obtained by applying GAP [66], followed by a 1×1 convolutional layer without bias. To avoid a trivial solution, that is, features collapse to 0, an orthogonal regularization [68] is applied to constrain the parameter of convolutional layer. The parameter is marked as $W_c \in \mathbb{R}^{D \times D}$, where D is the dimension of the features. Finally, the L2 distance of features in the latent space is calculated to obtain semantic consistency loss $\mathcal{L}_c$.

$$G_i = \text{Conv}(\text{GAP}(Y_i)), \quad i = 1, 2 \tag{6}$$

$$\mathcal{L}_c(Y_1, Y_2) = ||G_1 - G_2||_2 + ||W_c^T W_c - I||_2 \tag{7}$$

where Y_i denotes the features to be constrained, and I denotes the identity matrix with ones on the diagonal and zeros elsewhere. The CSC loss is the sum of semantic consistency losses between PAN and MS features at each stage

$$\mathcal{L}_{\text{csc}} = \sum_{i=2}^{5} \mathcal{L}_c\left(X_i^p, X_i^m\right). \tag{8}$$

The spatial preservation loss measures spatial information agreement between two inputs. Considering that spatial information lies in the activations of feature maps, a max-pooling operation along the channel axis is performed to obtain the spatial feature map. Finally, the loss is calculated using L2 distance as

$$S_i = \text{Conv}\left(\max_c(Y_i)\right), \quad i = 1, 2 \tag{9}$$

$$\mathcal{L}_s(Y_1, Y_2) = ||S_1 - S_2||_2 \tag{10}$$

where S_i denotes the spatial feature map, $\mathcal{L}_s$ is the spatial preservation loss, and Y_i denotes the features to be constrained. The parameters of the spatial preservation loss will not collapse to zero since its' optimization difficulty is much lower than that of the semantic consistency loss.

The overall preservation loss between PAN and fused features, i.e., the PiP loss, could be formulated as

$$\mathcal{L}_{\text{pip}} = \sum_{i=2}^{5} \mathcal{L}_c\left(X_i^f, X_i^p\right) + \mathcal{L}_s\left(X_i^f, X_i^p\right) \tag{11}$$

D. Detection

To coordinate optimization with the detection task, the CSC loss and PiP loss are optimized during training. Let $\mathcal{L}_{\text{det}}$ be the detection loss and the total loss of our model is

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{pip}} + \mathcal{L}_{\text{csc}}. \tag{12}$$

The detection loss $\mathcal{L}_{\text{det}}$ depends only on the detector, irrelative to our method. Our experiments are performed on 12 popular object detectors, including Faster R-CNN (FRCNN) [25], FoveaNet (FvNet) [69], FSAF [31], GA Faster R-CNN (GFRCNN) [28], Grid R-CNN (GRCNN) [26], RetinaNet (RtnNet) [29], ATSS [32], Cascade R-CNN (CRCNN) [70], Dynamic R-CNN (DRCNN) [71], Reppoints [72], Sparse R-CNN (SRCNN) [73], and the newest RTMDet [33]. The first 11 models use FPN for multiscale feature extraction. These models are trained with 12 epochs, and the learning rate decays by a factor of 10 at epoch 8 and 11. Models except for Sparse R-CNN [73] are optimized with SGD optimizer with an initial learning rate of 0.01. For Sparse R-CNN [73], the SGD optimizer is replaced with AdamW optimizer and reduces the initial learning rate to 0.000025. RTMDet [33] is an efficient real-time detector equipped with an FPN. The AdamW with a 0.05 weight decay and cosine annealing [74] with a minimum learning rate of 0.0002 are adopted for optimizing RTMDet. The medium size one is chosen for our experiments among the five available model sizes in RTMDet. Warm-up strategy is adopted for the first 500 iterations with ratios of 0.33 to stabilize the training process. The gradient clipping with maximum normalized value of 35 is also utilized to avoid gradient explosion. The experiments run on a single NVIDIA 2080TI GPU with a batch size of 4. For testing, non-maximum suppression (NMS) with intersection over union (IoU) threshold of 0.3 is leveraged to remove duplicated bounding boxes. In addition, boxes with scores less than 0.05 are removed to further reduce false detections.

IV. Experiments

In this section, we first introduce the building dataset for evaluation. Then, the impacts of different fusion levels on detection is validated, revealing the disadvantages of image-level fusion and result-level fusion methods in building detection. The alternative multiscale architectures for multimodal fusion are discussed afterward. What is following is the ablation study of our proposed CSC loss and PiP loss. Finally, the proposed DAFNet is compared with other feature fusion strategies.

A. Dataset

Experiments are conducted on 5M-building dataset [75], which is comprised of images captured by GaoFen-2 satellite over Shandong province of China. This dataset contains 109 PAN images and their corresponding MS images. The spatial resolution is about 3.2 m for the 4-band MS images and 0.8 m for the PAN images. The image size ranges from 2000×2000 to 5000×5000 pixels. Buildings in 5M-building dataset are diverse in scale and shape. Some examples are shown in Fig. 2.

The MS images are up-sampled by bilinear interpolation to meet the size of their corresponding PAN images. Then, all images are cropped into 512×512 patches with an overlap of 64, constituting training and test samples. Finally, there are 3 750 images containing 62 487 buildings in the training set and 1 233 images containing 15 550 buildings in the testing set. Images captured by GaoFen-2 have a bit depth deeper than 8, so the pixels are normalized into [0, 255] by histogram equalization.

Statistics of training set w.r.t building size, aspect ratio, and instances in each sample are shown in Fig. 3. The dataset contains many small objects; 37 299 buildings are smaller than 32×32 pixels. It also can be seen that buildings vary significantly in aspect ratio; 11 341 buildings have an aspect ratio greater than 4.

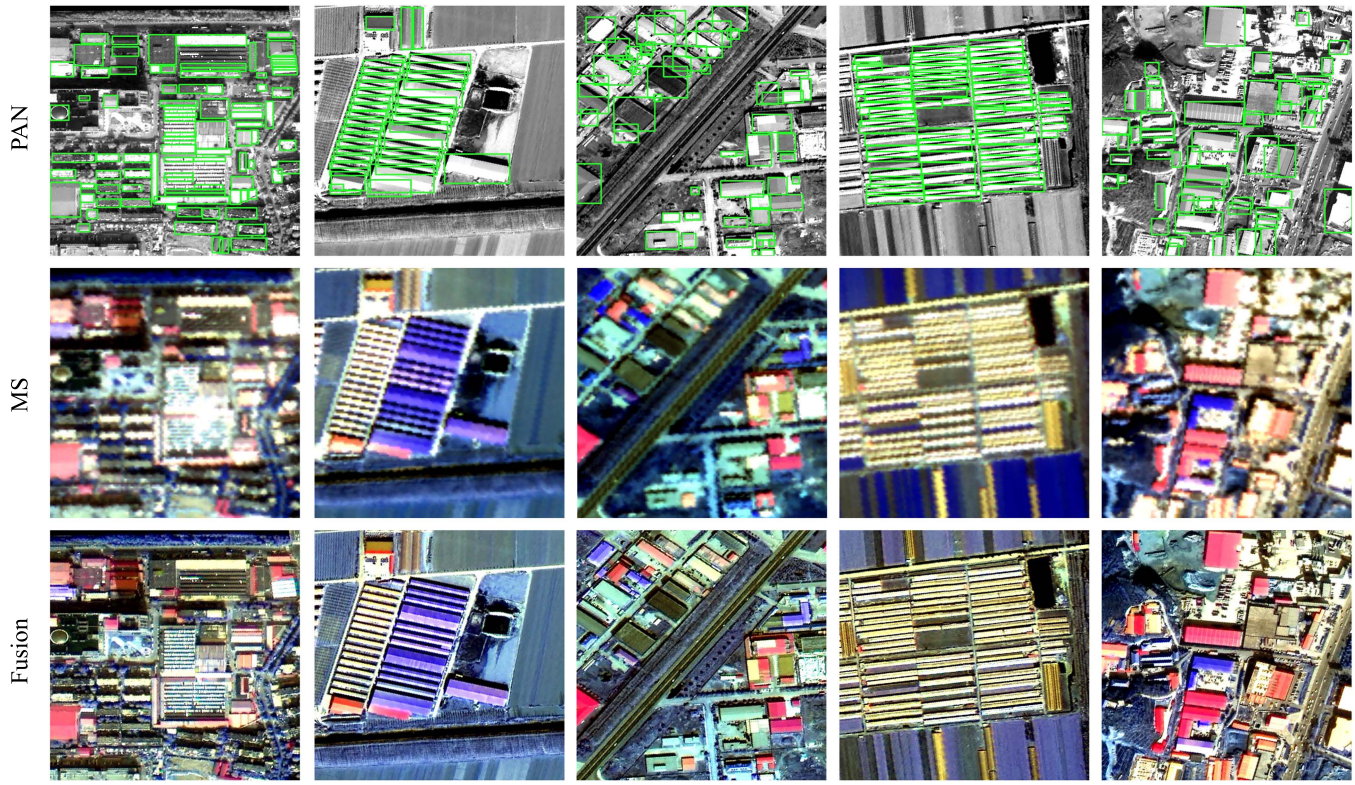

Fig. 2. Sample images in 5M-building. PAN and MS image pairs are shown in the top and middle rows. The fused images obtained by the Brovey method are illustrated in the bottom row. Annotations are marked with green rectangle boxes and are drawn on the PAN images.

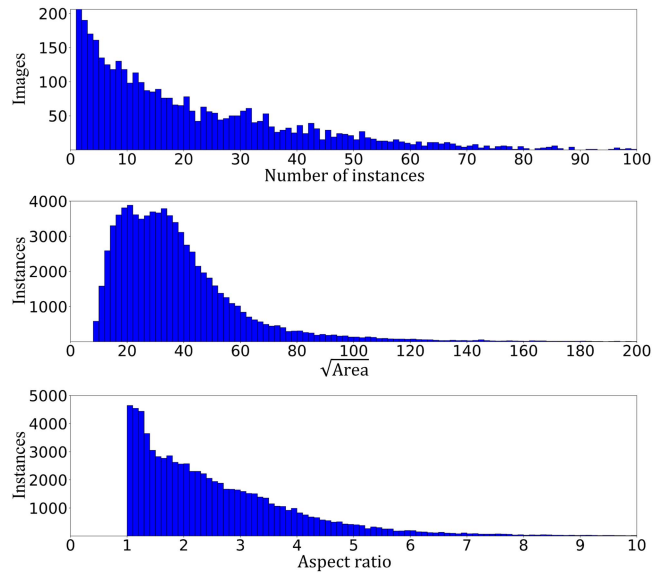

Fig. 3. Statistics of 5M-building dataset in three aspects: number of instances in each image, building area, and the aspect ratio of buildings.

B. Experiment Details

Our model is implemented with MMDetection [76]. All models use ResNet50 as the backbone network. The backbone is initialized with ImageNet [77] pretrained weights. The first layer of the backbone is frozen to match the default configuration in MMDetection [76]. During training, the MS images and PAN images are resized into 800×800 through bilinear interpolation, and then, random horizontal flips with a probability of 0.5 for data augmentation. The images are normalized with the mean and variance obtained from the ImageNet images, since the pretrained parameters are derived from the ImageNet classification task. Image preprocessing in the test phase is consistent with training, except that no data augmentation is used. The performance is measured by COCO [78] metrics, including mean average pooling (mAP) and AP50.

C. Impacts of Fusion Levels

The impacts of different fusion levels on detection are validated in this section, including image-level fusion, feature-level fusion, and decision-level fusion.

Eight pan-sharpening methods are selected for image-level fusion, including Brovey [6], fast intensity-hue-saturation (FIHS)-based [7], principal component analysis (PCA)-based [8], A Tróus wavelet transform (ATWT)-based [9], PGMAN [11], PNN [79], PanNet [13], and PSGAN [12]. These methods are either widely used in practical applications or new deep fusion approaches. In our experiments, the traditional methods are directly applied to obtain fused images without training. The CNN-based methods are trained with the open accessed GaoFen-2 images, and then, used for image fusion. Two nonreference metrics D_λ [80] and D_S [80] are used to evaluate the performance of the pan-sharpening methods. Faster R-CNN [25]

TABLE I
DETECTION PERFORMANCE WITH DIFFERENT FUSION STRATEGIES IN TERMS OF AP50 (%) ON 5M-BUILDING TEST SET

Task	Dataset	Detectors		Metrics	
		FRCNN [25]	RtnNet [29]	$D_\lambda \downarrow$	$D_S \downarrow$
Image-level	Brovey [6]	**67.5**	62.7	0.1629	0.2256
	FIHS [7]	59.9	54.7	0.2691	0.2201
	PCA [8]	65.3	61.4	0.2100	0.2692
	PGMAN [11]	63.1	57.9	0.0171	0.0647
	PNN [79]	62.6	57.6	0.0179	0.0748
	ATWT [9]	50.5	46.7	0.0088	0.0512
	PSGAN [12]	52.7	55.7	0.0139	0.0985
	PANNet [13]	52.1	55.7	0.0109	0.1102
	[PAN, MS]	64.9	59.5	-	-
Single-modality	MS	48.7	45.2	-	-
	PAN	67.0	**63.0**	-	-
Decision-level	NMS	65.5	60.4	-	-

and RetinaNet [29] are used for evaluation. The results are shown in Table I.

It can be seen that PAN images are far better than MS images for building detection, indicating that spatial information is essential for the detection task. Simply concatenating PAN and MS images together is not a good solution. The results degenerate compared to that of using PAN images. A possible reason may be that MS channels dominate the input, which makes the network hard to learn textural and structural features that are mostly within PAN. The decision-level fusion based on the detections of PAN and MS images is also investigated. We use PAN and MS as training data, and train two independent detectors based on Faster R-CNN. Then, the results of the two detectors are merged by using nonmaximum suppression (NMS) algorithm. The results are shown in the NMS row of Table I, which are worse than PAN but slightly better than [PAN, MS]. Since the performance gap between PAN and MS images is huge, MS drags the performance of NMS. Also, pan-sharpening has a significant impact on building detection. As shown in Table I, although CNN-based methods achieve much better fusion results in terms of D_λ and D_S, the performance on detection is completely opposite. ATWT based is the worst among the eight methods. Brovey-based produces the best results, although PAN is slightly better for the RetinaNet detector. Detections from the rest pan-sharpening methods are not as good as PAN images, indicating substantial information loss during pan sharpening.

D. Multiscale Multimodal Fusion

An FPN [20] is widely used in object detection to address the multiscale problem. In this work, the strategies of combining multiscale features for multimodal fusion are discussed.

A straightforward approach is simultaneously performing multiscale multimodal feature fusion, termed SiMM, as shown in Fig. 4(b). At each scale, except for the lowest one, the fusion module accepts features from the multimodal features in the same scale and the fused futures of the lower scale. The fusion process can be described as follows:

$$X_5^f = \text{Fusion}_i\left(\text{L}_5^p(C_5^p), \text{L}_5^m(C_5^m)\right) \tag{13}$$

$$X_i^f = \text{Fusion}_i\left(\text{L}_i^p(C_i^p), \text{L}_i^m(C_i^m), X_{i+1}^f\right),\ i = 2, 3, 4 \tag{14}$$

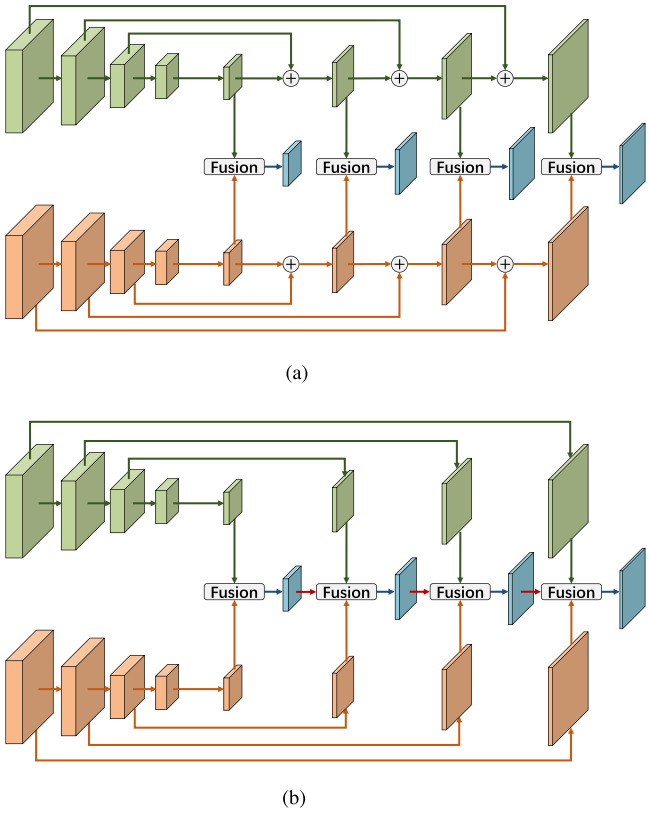

Fig. 4. Two multiscale architectures for fusion. (a) Dual FPN fusion (DuFF), which first obtains multiscale features by the FPN, and then, fuses features in each scale. (b) SiMM, which performs simultaneous multiscale and multimodal feature fusion.

where C_i^p and C_i^m denote features from PAN and MS after the ith stage, respectively; L_i denotes the ith lateral connection; X_i^f denotes the fused feature; and Fusion_i denotes the fusion module. During fusion, the last obtained feature is upsampled by a factor of two.

The second architecture is dual FPN fusion (DuFF), which is also implemented in our DAF network. DuFF first builds feature pyramids for different modalities, and then, performs fusion in each scale, as shown in Fig. 4(a).

The experiments evaluate two fusion strategies, i.e., element-wise addition (ADD) and concatenation (CAT). The ADD applies element-wise addition to combine features and uses a 3×3 convolutional layer to obtain fused features. The CAT concatenates features and then compresses the dimension through a 3×3 convolutional layer. Both of these two operations are tested in SiMM and DuFF. The results are shown in Table II, with the "Source" column indicates the data source utilized for training. DuFF delivers better performance than SiMM, and ADD operation is better than CAT. The reason is that DuFF is a progressive fusion method that first performs multiscale feature fusion, and then, completes the multimodal feature fusion, while SiMM accomplishes multiscale and multimodal fusion simultaneously, the network would be confused about what is important and what should be preserved when fusion. On the other hand, the large semantic gaps between different modalities and different scales

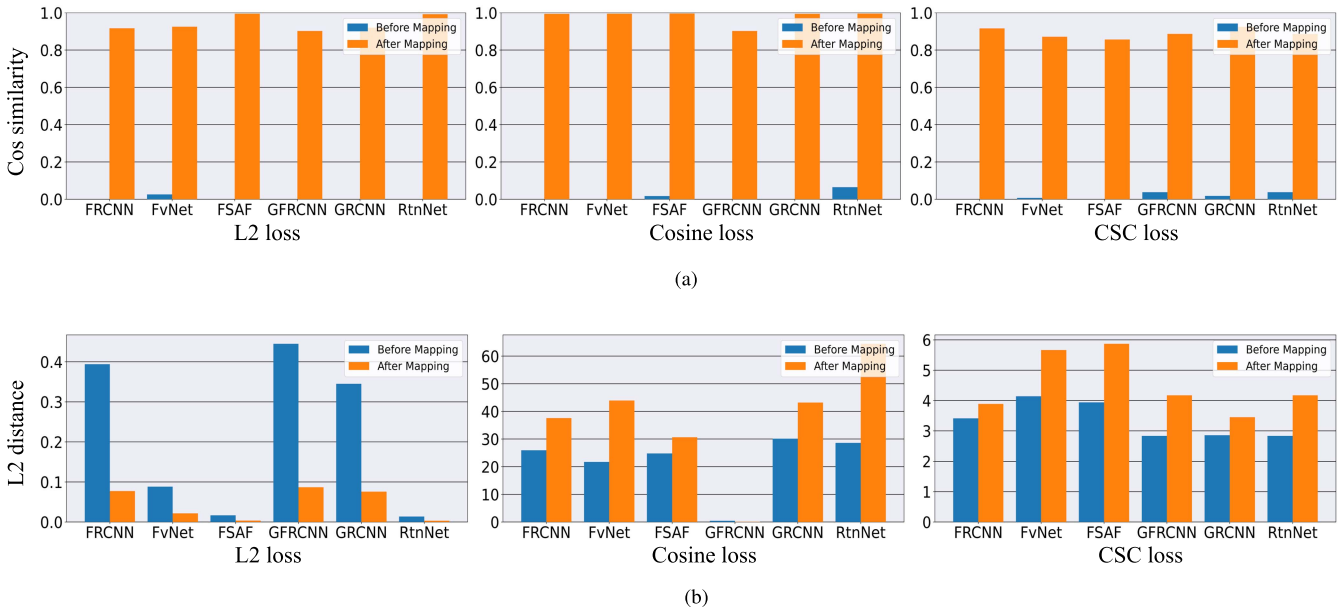

Fig. 5. Influence of consistency loss on feature learning. The semantic agreement before and after mapping is measured by cosine distance and diversity using L2 distance. A good consistency loss must ensure that the semantic information of features should be held and does not harm the diversity of the features. (a) Semantic agreement before and after mapping. (b) Diversity changes before and after mapping.

TABLE II
PERFORMANCE COMPARISON OF MULTISCALE MULTIMODAL FUSION STRATEGIES BASED ON FASTER R-CNN [25] ON 5M-BUILDING

Methods	Fusion Strategy	Source	AP	AP50
FPN	-	PAN	39.6	67.0
FPN	-	MS	20.5	48.7
SiMM	ADD	PAN + MS	39.5	67.2
SiMM	CAT	PAN + MS	39.1	66.5
DuFF	ADD	PAN + MS	**39.8**	**68.0**
DuFF	CAT	PAN + MS	39.6	66.5
SiMM	ADD	PAN + PAN	39.4	67.1
SiMM	CAT	PAN + PAN	38.7	66.3
DuFF	ADD	PAN + PAN	39.2	67.0
DuFF	CAT	PAN + PAN	38.4	65.3

make fusion difficult. Additionally, we perform experiments using PAN+PAN as the data source, as shown in the bottom four rows in Table II. It is found that the performance of the dual-stream network using PAN+PAN as the data source is lower than that of the single network using only PAN. This phenomenon indicates that the performance improvement brought by DuFF is due to the spectral information from MS images rather than the extra computation of the multibranch structure. Thus, DuFF with ADD is chosen as our multiscale architecture for fusion.

E. Consistency Loss

In addition to our CSC loss, there are two options for imposing consistencies between two features. The first is simply minimizing L2 distance between features [81], and the second is maximizing cosine distance between the two modalities [82]. All losses are computed in a latent space where features are projected with a linear mapping. The results are shown in Table III. The baseline simply adds the two features element-wise

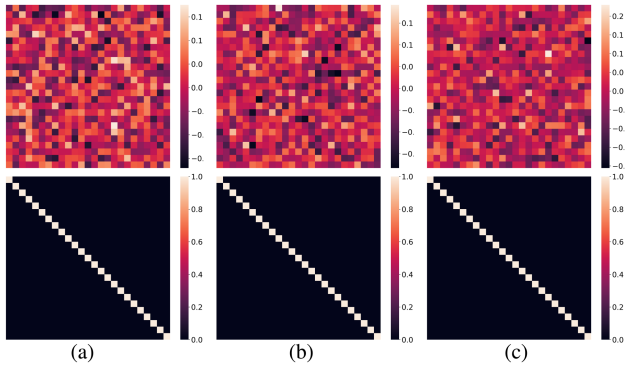

Fig. 6. Visualization of W_c and $W_c^T W_c$. The first 25 rows and 25 columns of the parameters used in the first three layers are visualized. The first row indicates the parameters W_c of 1×1 convolutional layer used in semantic consistency loss, and the second row indicates $W_c^T W_c$.

without imposing any constraint. It can be seen that L2 loss decreases the detection performance. In particular, FSAF [31] and RetinaNet [29] do not converge. Cosine loss and ours improve the performance, while ours obtains the best results.

The semantic agreement and diversities of features before and after the linear mapping are computed to further investigate why this happens. Also, the cosine metric is used to measure semantic agreement and L2 distance to measure the diversity of features. If two features have strong semantic consistency, their cosine similarity should be close. Features should also be diverse so the model can learn good decision boundaries. The results are shown in Fig. 5.

As can be observed, all losses improve semantic consistency. Cosine loss obtains the best results since it imposes the cosine similarity directly. However, it boosts the diversity of features,

TABLE III
COMPARISON WITH OTHER CONSISTENCY LOSS ON 5M-BUILDING DATASET

Fusion module	Consistency loss	FRCNN [25]	FvNet [70]	FSAF [31]	GFRCNN [28]	GRCNN [26]	RtnNet [29]	Avg. Improv.
AFF	-	68.0	66.5	**68.0**	67.2	68.5	63.0	
	L2 loss	65.6	62.6	NaN	64.3	66.8	NaN	NaN
	Cosine loss	68.1	66.0	67.4	**67.9**	**69.4**	63.7	+0.24
	CSC loss	**68.5**	**67.7**	67.5	67.4	**69.4**	**64.0**	**+0.55**

The consistency constraints are imposed on six models, and AP50 (%) is used as the metric. The last column is the average improvement compared with training without consistency loss. "NaN" means the training does not converge.

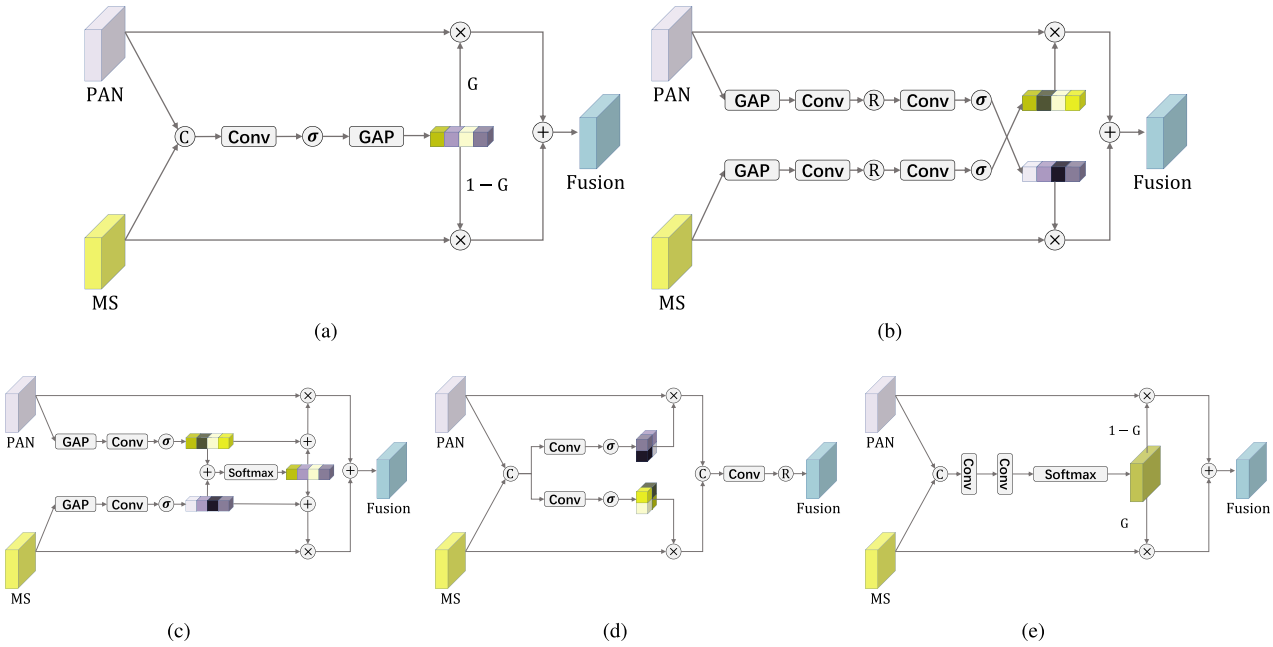

Fig. 7. Five popular fusion approaches for comparison. (a) Channel-wise weighted feature fusion (CWF) [14]. (b) Cross gates (GRSs) [15]. (c) Cross reference module (CRM) [16]. (d) Gated information fusion (GIF) [17]. (e) Adaptive feature fusion modules (AFFM) [18] Ⓒ: Concatenate operation. Ⓡ: ReLU. ⊗: Tensor product. ⊕: Element-wise addition. Ⓢ: Sigmoid activation. "GAP": Global average pooling.

TABLE IV
ABLATION STUDY OF AFF MODULE ON 5M-BUILDING USING SIX DETECTORS, AP50 (%) IS USED AS THE METRIC

Fusion module	CSC	PiP	FRCNN [25]	FvNet [70]	FSAF [31]	GFRCNN [28]	GRCNN [26]	RtnNet [29]	Avg. Improv.
AFF			68.0	66.5	68.0	67.2	68.5	63.0	
	✓		**68.5**	**67.7**	67.5	67.4	69.4	**64.0**	+0.55
		✓	68.4	67.3	**68.6**	**67.6**	69.6	63.4	+0.62
	✓	✓	68.2	**67.7**	68.4	**67.6**	**70.0**	63.7	**+0.73**

which may increase model instability, as shown in the bar graph. L2 loss significantly decreases the diversity, which hampers the detection. The proposed loss improves the semantic consistency while still maintaining an appropriate diversity of the features.

In addition, the parameters W_c and $W_c^T W_c$ in (7) are visualized. The first 25 rows and 25 columns of the parameters used in the first three layers are selected for visualization, as shown in Fig. 6. It can be found that the orthogonal loss of our matrix can well force the matrix to meet the orthogonality, so the mapping is only used to transform the feature into a new space to complete the constraints, and there will be no feature collapse. We also study the effectiveness of each term of our loss. The results are shown in Table IV. The ADD strategy is considered as the baseline. It can be seen that the CSC loss increases AP50 by an average of 0.55, the PiP loss increases by an average of 0.62, and the combination achieves the best, increasing by an average of 0.73.

F. Overall Results

To further demonstrate the effectiveness of our fusion method, we compare it with other fusion strategies that are widely used in RGB-Depth and RGB-Thermal perception tasks, including: channel-wise weighted feature fusion (CWF) [14], cross gates (CRGs) [15], cross reference module (CRM) [16], gated information fusion (GIF) [17], and a fusion method for PAN and MS data fusion, i.e., the adaptive feature fusion module

TABLE V
PERFORMANCE COMPARISON ON 5M-BUILDING IN TERMS OF AP50 (%)

Methods	PAN (Baseline)	Brovey [6]	ADD	CWF [14]	CRGs [84]	CRM [16]	GIF [17]	AFMM [18]	DAFNet (Ours)
FRCNN [25]	67.0	67.5	68.0	66.0	66.1	66.2	67.0	67.6	**68.2** (+1.2)
FvNet [69]	66.5	66.4	66.5	66.4	66.8	67.0	67.0	66.6	**67.7** (+1.2)
FSAF [31]	67.5	67.1	68.0	67.9	67.0	67.3	67.1	67.5	**68.4** (+0.9)
GFRCNN [28]	66.1	66.6	67.2	66.3	65.4	66.1	66.8	67.3	**67.6** (+1.5)
GRCNN [26]	68.5	68.8	68.5	68.2	68.1	68.4	69.3	68.8	**70.0** (+1.5)
RtnNet [29]	63.0	62.7	63.0	63.5	63.7	63.5	63.0	62.8	**63.7** (+0.7)
ATSS [32]	66.9	**68.0**	66.9	66.8	66.7	66.8	67.0	67.1	67.7 (+0.8)
CRCNN [70]	66.6	67.3	66.9	65.8	66.3	66.6	67.0	66.8	**67.9** (+1.3)
DNRCNN [71]	65.0	64.3	64.7	64.1	64.0	64.2	64.6	64.9	**65.1** (+0.1)
Reppoints [72]	67.8	68.3	68.2	**68.8**	68.1	68.0	68.3	68.2	68.4 (+0.6)
SRCNN [73]	51.1	50.5	50.2	49.9	48.3	50.2	50.9	50.0	**54.8** (+3.7)
RTMDet [33]	66.4	66.3	67.1	67.3	66.7	66.3	66.8	65.9	**67.6** (+1.2)
Avg. Improv.	-	+0.12	+0.24	-0.18	-0.51	-0.18	+0.21	+0.06	**+1.27**

(AFFM) [18]. For CWF, CRGs, and GIF, we reimplement them in strict accordance with the article; for SCA and AFMM, we use the codes the authors provided.

The detailed architecture of each fusion module is shown in Fig. 7. For fair comparisons, our AFF module is replaced with these modules and an extra skip connection is added on the PAN side so that the valuable PAN information could be preserved for detection. Twelve popular object detectors are employed for evaluation, including Faster R-CNN (FRCNN) [25], FoveaNet (FvNet) [69], FSAF [31], GA Faster R-CNN (GFRCNN) [28], Grid R-CNN (GRCNN) [26], RetinaNet (RtnNet) [29], ATSS [32], Cascade R-CNN (CRCNN) [70], Dynamic R-CNN (DRCNN) [71], Reppoints [72], Sparse R-CNN (SRCNN) [73], and the newest RTMDet [33].

CWF [14] first concatenates features of PAN and MS, and then, fuses them using a convolutional layer. Afterward, a weight vector is generated from the fused features using GAP, which will be used to reweight the PAN and MS features, as shown in Fig. 7(a). Finally, the fused features are obtained by element-wise addition of the weighted PAN and MS features.

CRGs [15] generates channel weights for PAN and MS modalities, respectively, and then, applies them crosswise, as shown in Fig. 7(b).

CRM [16] first obtains channel attention vectors for each modality, and then, mines the most discriminative features among them through element-wise addition. Finally, the channel features are fused according to the weights of each mode and the common important region, as shown in Fig. 7(c).

GIF [17] uses a spatial gate fusion mechanism. It generates spatial weight maps for each modality based on their concatenated features. Fusion is achieved through weighted concatenation, as shown in Fig. 7(d).

AFFM [18] generates weights from the concatenation of features after two convolutional layers. A softmax operation will then normalizes the weights along the channel. After that, AFFM computes element-wise weighted sum to fuse spatial and spectral features, as shown in Fig. 7(e).

Detection results using PAN images are taken as the baseline and compared with the ADD fusion strategy described in Section IV-D and detections based on Brovey pan-sharpened images. The performance of these methods is illustrated in Table V. In general, the performance of all detectors on the 5M-Building dataset does not exceed 70% AP50. This is mainly because 5M-Building dataset covers complex scenes and has diverse building styles, and large scale variations than other building datasets, as shown in Figs. 2 and 8. These diversities make 5M-Building dataset more challenging, so the detection performance is relatively lower.

It can be seen that, according to the average improvements, ADD, Brovey pan-sharpening, and GIF slightly improve the detection performance. All the other three fusion approaches decrease the detection. In sum, these four fusion approaches do not contribute much to the detection. Our method achieves an average improvement of 1.27% AP50. Furthermore, we promote Grid R-CNN to achieve 70% AP50, which performs the best in 5M-Building dataset, and improve Sparse R-CNN by 3.7% AP50.

Table VI shows all methods' running time and complexity based on Faster R-CNN [25]. Our DAFNet has fewer parameters than other feature fusion methods and achieves better performance. In particular, DAFNet reaches 68.2% AP50 with 74.2 M parameters and 12.5 FPS during inference, indicating that it is a effective way to realize feature fusion compared with other methods. In addition, the CSC loss and PiP loss introduced by

TABLE VI
COMPARISON OF PERFORMANCE (AP50), PARAMETER (PARAM), TRAINING TIME (TIME), AND INFERENCE SPEED (SPEED) OF DAFNET WITH OTHER METHODS BASED ON FASTER R-CNN [25]

Methods	AP50	Param (M)	Time (h)	Speed (FPS)
PAN	67.0	**41.4**	**1.39**	**21.6**
Brovey [6]	67.5	**41.4**	1.44	20.8
ADD	68.0	74.2	2.59	12.5
CWF [14]	66.0	77.9	2.81	11.6
CRGs [83]	66.1	74.3	2.60	11.7
CRM [16]	66.2	75.5	2.72	11.8
GIF [17]	67.0	74.9	2.72	10.9
AFFM [86]	67.6	83.1	2.97	10.2
DAFNet (Ours)	**68.2**	74.2	2.82	12.5

The inference speed is tested using 800 × 800 size images on a single NVIDIA 2080TI. Time of postprocess (i.e., NMS) is included to reflect the most realistic performance of the models.

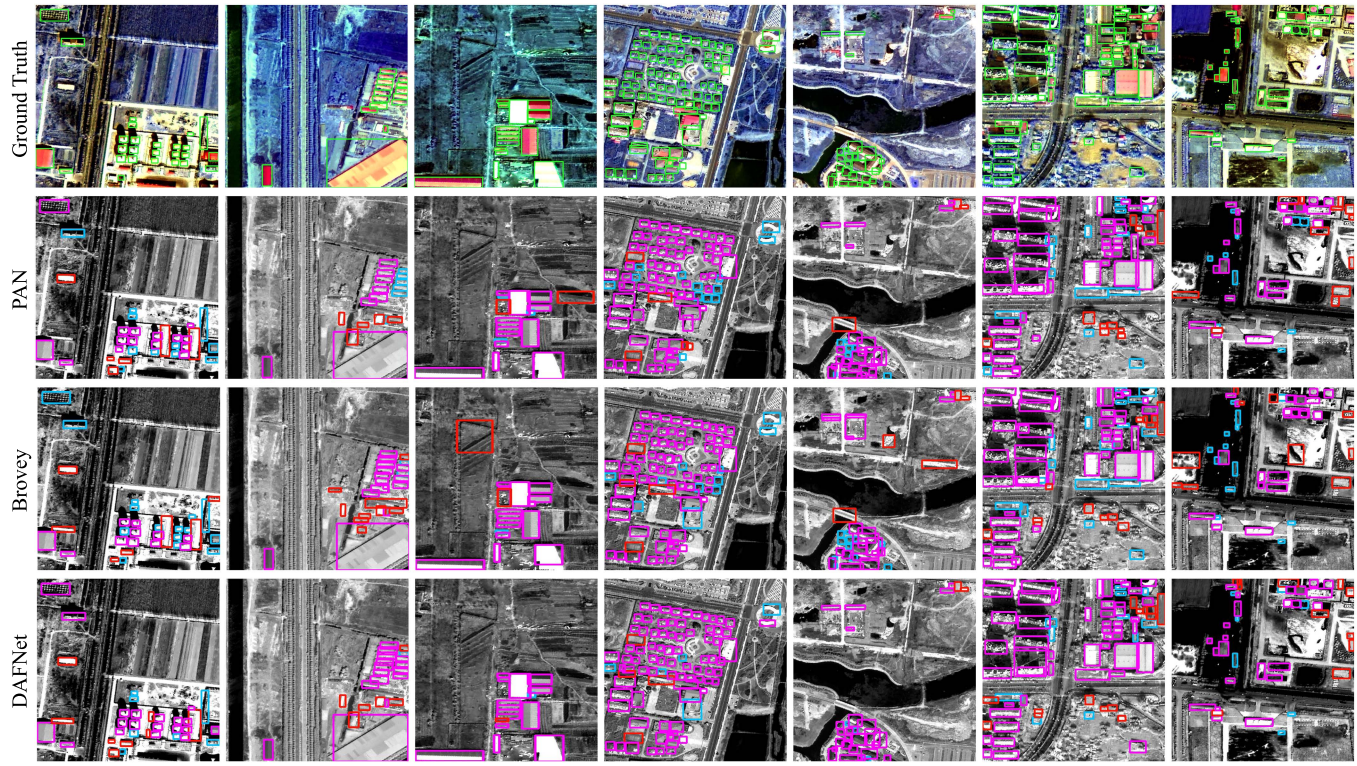

Fig. 8. Example detection results on 5M Building. Some Faster R-CNN detection results from PAN images and Brovey pan-sharpened images are given in the second and third rows for comparison. Note: Pink boxes refer to detection results, green boxes refer to the ground truth, blue boxes refer to miss detections, and red boxes refer to false positives.

DAFNet can effectively improve the fusion effect without adding extra computational costs during inference.

Some results are visualized in Fig. 8. Detection results from PAN and Brovey pan-sharpened images using Faster R-CNN are shown in the second and third rows. As can be observed, our model has fewer miss detections and false positives. The proposed fusion method effectively combines the strengths of PAN and MS images, enabling augmented features of buildings, thus leading to more accurate localization and classification.

V. Conclusion

In this article, we have conducted in-depth studies of building detection from remote sensing images. We reveal that pan sharpening may degenerate the building detection performance. The building detection problem is resolved from a multimodality feature fusion view and a dual-stream asymmetric fusion network is proposed to effectively fuse and augment PAN and MS features for building detection. The fusion is realized with an AFF module and two consistency regularization losses, i.e., CSC loss and PiP loss. Extensive experiments on 5M-building demonstrate the effectiveness and superiority of the proposed approach.

Although the proposed DAFNet was motivated by the PAN and MS fusion problem in remote sensing, the method is a general framework that can be applied to other data sources, such as optical images and photogrammetric point clouds [84]. Additionally, we noticed that the independent dual-branch structure would bring too many parameters. A future direction is to use siamese networks combined with joint learning [85] to achieve a tradeoff between speed and accuracy.

References

[1] C. Kamusoko, "Importance of remote sensing and land change modeling for urbanization studies," in *Urban Development in Asia and Africa*. Singapore: Springer, 2017, pp. 3–10.

[2] H. Ma, Y. Liu, Y. Ren, and J. Yu, "Detection of collapsed buildings in post-earthquake remote sensing images based on the improved YOLOv3," *Remote Sens.*, vol. 12, no. 1, 2020, Art. no. 44.

[3] N. Casagli et al., "Spaceborne, UAV and ground-based remote sensing techniques for landslide mapping, monitoring and early warning," *Geoenvironmental Disasters*, vol. 4, no. 1, pp. 1–23, 2017.

[4] Q. Hu et al., "Exploring the use of Google earth imagery and object-based methods in land use/cover mapping," *Remote Sens.*, vol. 5, no. 11, pp. 6026–6042, 2013.

[5] G.-S. Xia et al., "Dota: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.

[6] A. R. Gillespie, A. B. Kahle, and R. E. Walker, "Color enhancement of highly correlated images. II. channel ratio and 'chromaticity' transformation techniques," *Remote Sens. Environ.*, vol. 22, no. 3, pp. 343–365, 1987.

[7] T.-M. Tu, P. S. Huang, C.-L. Hung, and C.-P. Chang, "A fast intensity-hue-saturation fusion technique with spectral adjustment for IKONOS imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 4, pp. 309–312, Oct. 2004.

[8] H. R. Shahdoosti and H. Ghassemian, "Spatial PCA as a new method for image fusion," in *Proc. 16th CSI Int. Symp. Artif. Intell. Signal Process.*, 2012, pp. 090–094.

[9] R. B. Gomez, A. Jazaeri, and M. Kafatos, "Wavelet-based hyperspectral and multispectral image fusion," in *Proc. Geo-Spatial Image Data Exploitation II*, 2001, vol. 4383, pp. 36–42.

[10] J. Ma, W. Yu, C. Chen, P. Liang, X. Guo, and J. Jiang, "Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion," *Inf. Fusion*, vol. 62, pp. 110–120, 2020.

[11] H. Zhou, Q. Liu, and Y. Wang, "PGMAN: An unsupervised generative multiadversarial network for pansharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6316–6327, Jun. 2021.

[12] Q. Liu, H. Zhou, Q. Xu, X. Liu, and Y. Wang, "PSGAN: A generative adversarial network for remote sensing image pan-sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10227–10242, Dec. 2021.

[13] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PANNET: A deep network architecture for pan-sharpening," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5449–5457.

[14] Q. Zhang, S. Zhao, Y. Luo, D. Zhang, N. Huang, and J. Han, "ABMDRNet: Adaptive-weighted bi-directional modality difference reduction network for RGB-T semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2633–2642.

[15] W. Kang, Y. Xiang, F. Wang, and H. You, "CFNet: A cross fusion network for joint land cover classification using optical and SAR image," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1562–1574, Jan. 2022.

[16] W. Ji et al., "Calibrated RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9471–9481.

[17] J. Kim, J. Koh, Y. Kim, J. Choi, Y. Hwang, and J. W. Choi, "Robust deep multi-modal learning based on gated information fusion network," in *Proc. Asian Conf. Comput. Vis.*, Springer, 2018, pp. 90–106.

[18] K. Zhang, A. Wang, F. Zhang, W. Diao, J. Sun, and L. Bruzzone, "Spatial and spectral extraction network with adaptive feature fusion for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2022, Art. no. 5410814.

[19] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019.

[20] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[21] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63373–63394, 2019.

[22] X. Li, L. Lei, Y. Sun, M. Li, and G. Kuang, "Collaborative attention-based heterogeneous gated fusion network for land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 3829–3845, May 2021.

[23] X. Liu, Q. Liu, and Y. Wang, "Remote sensing image fusion based on two-stream fusion network," *Inf. Fusion*, vol. 55, pp. 1–15, 2020.

[24] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Adv. Neural Inf. Process. Syst.*, vol. 28, pp. 91–99, 2015.

[26] X. Lu, B. Li, Y. Yue, Q. Li, and J. Yan, "Grid R-CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7363–7372.

[27] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards balanced learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 821–830.

[28] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, "Region proposal by guided anchoring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2965–2974.

[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[30] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.

[31] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 840–849.

[32] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9759–9768.

[33] C. Lyu et al., "RTMDet: An empirical study of designing real-time object detectors," 2022, *arXiv:2212.07784*.

[34] 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.*, 2015, pp. 1873–1876.

[35] Q. Zhang, Y. Wang, Q. Liu, X. Liu, and W. Wang, "CNN based suburban building detection using monocular high resolution Google earth images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 661–664.

[36] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[37] Q. Li, Y. Wang, Q. Liu, and W. Wang, "Hough transform guided deep feature extraction for dense building detection in remote sensing images," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 1872–1876.

[38] C. Li, X. Huang, J. Tang, and K. Wang, "A multi-branch feature fusion network for building detection in remote sensing images," *IEEE Access*, vol. 9, pp. 168511–168519, 2021.

[39] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. Dalla Mura, "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 130, pp. 139–149, 2017.

[40] R. Hamaguchi and S. Hikosaka, "Building detection from satellite imagery using ensemble of size-specific detectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 187–191.

[41] H. Yang, P. Wu, X. Yao, Y. Wu, B. Wang, and Y. Xu, "Building extraction in very high resolution imagery by dense-attention networks," *Remote Sens.*, vol. 10, no. 11, 2018, Art. no. 1768.

[42] D. Griffiths and J. Boehm, "Improving public data for building segmentation from convolutional neural networks (CNNs) for fused airborne lidar and image data using active contours," *ISPRS J. Photogrammetry Remote Sens.*, vol. 154, pp. 70–83, 2019.

[43] Q. Han, Q. Yin, X. Zheng, and Z. Chen, "Remote sensing image building detection method based on Mask R-CNN," *Complex Intell. Syst.*, pp. 1–9, 2021.

[44] W. Sirko et al., "Continental-scale building detection from high resolution satellite imagery," 2021, *arXiv:2107.12283*.

[45] L. Castrejon, K. Kundu, R. Urtasun, and S. Fidler, "Annotating object instances with a polygon-RNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5230–5238.

[46] Z. Li, J. D. Wegner, and A. Lucchi, "Topological map extraction from overhead images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1715–1724.

[47] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using CNN and regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2178–2189, Mar. 2019.

[48] Z. Li, Q. Xin, Y. Sun, and M. Cao, "A deep learning-based framework for automated extraction of building footprint polygons from very high-resolution aerial imagery," *Remote Sens.*, vol. 13, no. 18, 2021, Art. no. 3630.

[49] Y. Zhu, B. Huang, J. Gao, E. Huang, and H. Chen, "Adaptive polygon generation algorithm for automatic building extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4702114.

[50] Z. Guo et al., "Super-resolution integrated building semantic segmentation for multi-source remote sensing imagery," *IEEE Access*, vol. 7, pp. 99381–99397, 2019.

[51] Y. Chen, L. Tang, X. Yang, M. Bilal, and Q. Li, "Object-based multimodal convolution neural networks for building extraction using panchromatic and multispectral imagery," *Neurocomputing*, vol. 386, pp. 136–146, 2020.

[52] H. Zhang, E. Fromont, S. Lefàvre, and B. Avignon, "Guided attentive feature fusion for multispectral pedestrian detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 72–80.

[53] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "MMTM: Multimodal transfer module for cnn fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13289–13299.

[54] R. Cheng, R. Razani, E. Taghavi, E. Li, and B. Liu, "2-S3Net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12547–12556.

[55] B. Xue and N. Tong, "Real-world ISAR object recognition using deep multimodal relation learning," *IEEE Trans. Cybern.*, vol. 50, no. 10, pp. 4256–4267, Oct. 2020.

[56] B. Xue, W. Yi, F. Jing, and S. Wu, "Complex ISAR target recognition using deep adaptive learning," *Eng. Appl. Artif. Intell.*, vol. 97, 2021, Art. no. 104025.
[57] B. Xue, Y. He, F. Jing, Y. Ren, L. Jiao, and Y. Huang, "Robot target recognition using deep federated learning," *Int. J. Intell. Syst.*, vol. 36, no. 12, pp. 7754–7769, 2021.
[58] S. Li, C. Zou, Y. Li, X. Zhao, and Y. Gao, "Attention-based multi-modal fusion network for semantic scene completion," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 07, pp. 11402–11409.
[59] J. H. Yoo, Y. Kim, J. Kim, and J. W. Choi, "3D-CVF: Generating joint camera and lidar features using cross-view spatial feature fusion for 3 d object detection," in *Proc. 16th Eur. Conf., Comput. Vis.*, Glasgow, U.K., Aug. 2020, pp. 720–736.
[60] B. Duan, H. Tang, W. Wang, Z. Zong, G. Yang, and Y. Yan, "Audio-visual event localization via recursive fusion by joint co-attention," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 4013–4022.
[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[62] B. Pan, Z. Shi, X. Xu, T. Shi, N. Zhang, and X. Zhu, "CoinNet: Copy initialization network for multispectral imagery semantic segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 816–820, May 2018.
[63] X. He, Y. Chen, and P. Ghamisi, "Heterogeneous transfer learning for hyperspectral image classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3246–3263, May 2020.
[64] M. Ahmad, A. M. Khan, M. Mazzara, S. Distefano, M. Ali, and M. S. Sarfraz, "A fast and compact 3-D CNN for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Dec. 2022, Art. no. 5502205.
[65] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., Aug. 2020, pp. 213–229.
[66] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*.
[67] 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.
[68] A. Brock, T. Lim, J. M. Ritchie, and N. Weston, "Neural photo editing with introspective adversarial networks," in *Proc. Int. Conf. Learn. Representations*, 2017.
[69] X. Li et al., "FoveaNet: Perspective-aware urban scene parsing," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 784–792.
[70] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.
[71] H. Zhang, H. Chang, B. Ma, N. Wang, and X. Chen, "Dynamic R-CNN: Towards high quality object detection via dynamic training," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 260–275.
[72] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "Reppoints: Point set representation for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9657–9666.
[73] P. Sun et al., "Sparse R-CNN: End-to-End object detection with learnable proposals," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14454–14463.
[74] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Representations*, 2017.
[75] Z. Lu, T. Xu, K. Liu, Z. Liu, F. Zhou, and Q. Liu, "5m-building: A large-scale high-resolution building dataset with CNN based detection analysis," in *Proc. IEEE 31st Int. Conf. Tools With Artif. Intell.*, 2019, pp. 1385–1389.
[76] K. Chen et al., "MMdetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.
[77] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 25, 2012.
[78] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[79] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, 2016, Art. no. 594.
[80] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, "Multispectral and panchromatic data fusion assessment without reference," *Photogrammetric Eng. Remote Sens.*, vol. 74, no. 2, pp. 193–200, 2008.
[81] B. Xue and N. Tong, "DIOD: Fast and efficient weakly semi-supervised deep complex ISAR object detection," *IEEE Trans. Cybern.*, vol. 49, no. 11, pp. 3991–4003, Nov. 2019.
[82] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[83] Q. Ming, L. Miao, Z. Zhou, and Y. Dong, "CFC-Net: A critical feature capturing network for arbitrary-oriented object detection in remote sensing images," *IEEE Trans. Geosci Remote Sens.*, vol. 60, pp. 1–14, 2022.
[84] Y. Xie, J. Tian, and X. X. Zhu, "A co-learning method to utilize optical images and photogrammetric point clouds for building extraction," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 116, 2023, Art. no. 103165.
[85] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, and C. Zhu, "Siamese network for RGB-D salient object detection and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5541–5559, Sep. 2022.
[86] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

Ziyue Huang received the B.S. degree in information and computational science from the School of Mathematics and Statistics, Beijing Institute of Technology, Beijing, China, in 2018. He is currently working toward the Ph.D. degree in computer science with the Laboratory of Intelligent Recognition and Image Processing, School of Computer Science and Engineering, Beihang University, Beijing.

His research interests include computer vision and object detection.

Qingjie Liu (Member, IEEE) received the B.S. degree in computer science from Hunan University, Changsha, China, in 2007, and the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2014.

He is currently an Associate Professor with the School of Computer Science and Engineering, Beihang University. He is also a Distinguished Research Fellow with the Hangzhou Institute of Innovation, Beihang University, Hangzhou, China. His current research interests include image fusion, object detection, image segmentation, and change detection.

Huanyu Zhou received the B.S. and M.S. degrees in computer science from the School of Computer Science and Engineering, Beihang University, Beijing, China, in 2020 and 2013, respectively.

His research interests include computer vision and pattern recognition.

Guangshuai Gao received the B.S. degree in applied physics and M.S. degree in applied physics and signal and information processing from the Zhongyuan University of Technology, Zhengzhou, China, in 2014 and 2017, respectively, and the Ph.D. degree in computer science from School of Computer Science and Engineering, Beihang University, Beijing, China, in 2022.

He is currently a Lecturer with the School of Electronics and Information, Zhongyuan University of Technology. His research interests include image processing, digital machine learning, and remote sensing imagery interpretation.

Tao Xu received the M.S. degree in computer science from the China University of Petroleum, Beijing, China, in 2007, and the Ph.D. degree in computer science from Beihang University, Beijing, in 2016.

He is an Associate Professor with the School of Information Science and Engineering, University of Jinan, Jinan, China. His research interests include remote sensing image analysis, pattern recognition, and human–computer interaction.

Qi Wen received the B.S. degree in in mechanic and electronic engineering from the Beijing Institute of Technology, Beijing, China, in 2004, and the Ph.D. degree in signal and information processing from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, in 2009.

He is currently a Professor with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences. His research interests include remote-sensing information extraction and high-resolution remote sensing for disaster reduction.

Yunhong Wang (Fellow, IEEE) received the B.S. degree in electronic engineering from Northwestern Polytechnical University, Xi'an, China, in 1989, and the M.S. and Ph.D. degrees in electronic engineering from the Nanjing University of Science and Technology, Nanjing, China, in 1995 and 1998, respectively.

She was with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, from 1998 to 2004. Since 2004, she has been a Professor with the School of Computer Science and Engineering, Beihang University, Beijing, where she is also the Director of the Laboratory of Intelligent Recognition and Image Processing. Her research interests include biometrics, pattern recognition, computer vision, data fusion, and image processing.

Dr. Wang is a Fellow of the International Association for Pattern Recognition and China Computer Federation.