Tang and Dong
EURASIP Journal on Advances in Signal Processing (2022) 2022:56
https://doi.org/10.1186/s13634-022-00891-2

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Bridge-over-water detection via modulated deformable convolution and attention mechanisms

Rui Tang and Ganggang Dong*

*Correspondence: dongganggang@xidian.edu.cn

National Lab of Radar Signal Processing, Xidian University, Xi'an, China

Abstract

Bridge-over-water detection plays vital role in urban surveillance and military reconnaissance. Bridges have arbitrary orientations and extreme aspect ratios in remote sensing images, and the preceding works cannot adequately extract bridge-related features. Small bridges are difficult to detect accurately in optical remote sensing images. The oriented bounding box annotations are required by previous deep-learning-based methods for detecting rotated objects. But obtaining the annotations is a laborious task. Though widely studied previously, they are still challenging problems. To address these problems, modulated deformable convolution and attention mechanisms were introduced in this paper. Modulated deformable convolution made the receptive field more flexible. The feature extraction capability of the network was enhanced. A new weighted structure was designed to quantify the contributions of channel and spatial attention mechanisms. A selective attention usage strategy was proposed to improve the detection performance. To locate bridge-over-water more precisely, a new bounding box conversion module was presented. There was no need for oriented bounding box annotations, and the process only relied on bridge-related prior knowledge. Multiple experiments were performed to verify the effectiveness of proposed methods.

Keywords: Bridge detection, Modulated deformable convolution, Attention mechanism, Oriented bounding box, Optical remote sensing images

1 Introduction

Bridges over water bodies are the key node of the transportation network. Monitoring them is valuable for city surveillance, disaster assessment, and military reconnaissance [1, 2]. In remote sensing images, bridges over water bodies appear with different scales, shapes and textures. They have a high diversity of orientations and backgrounds. It is not easy to detect them accurately in aerial images.

Before the emergence of deep-learning technologies, prior knowledge and hand-crafted features are widely used by classical methods to detect bridges in remote sensing images [3–5]. Two-step approach is usually adopted. Waterbody regions are separated from land regions firstly, and then, bridges are further extracted from the waterbody

regions. The system proposed by Loménie et al. [6] categorizes terrain pixels according to their semantic meaning, and then, manually produced spatial decision rules are applied to locate bridges. Han et al. [7] analyze textures based on the gray level co-occurrence matrix. Correlation, entropy and homogeneity features are chosen to distinguish the rivers, and Hough transformation is applied for extracting the bridges. Luo et al. [8] use the Gauss Markov Random Field (GMRF)-Support Vector Machine (SVM) classification method to extract water bodies. And then, the bridge's main trunk is detected by static rules in the thinned image. The method proposed by Zhao et al. [9] gets bridge region candidates by saliency detection in compressed domain firstly. And then, these candidates are validated through Extreme Learning Machine (ELM) classification with Local Binary Patterns (LBP) feature; the final detection results are obtained. Chen et al. [10] apply the histogram-based threshold segmentation method for extract water bodies. On the basis of direction-augmented linear structuring elements, mathematical morphology method is utilized by them to extract the bridges. Gedik et al. [11] propose an algorithm for detecting bridges over water bodies in NDWI [12] images. Water regions are obtained through thresholding the NIR and clustering the NDWI images. Certain geometric constraints are adopted to identify river and water canals, and morphological operations are applied to locate bridges. Classical methods usually are clear in physical meaning. However, bridges are of varying backgrounds, textures and shapes in complex scenarios. Only relying on hand-crafted features and manually introduced decision rules is less robust. For example, it is difficult for classical methods to detect bridges with extreme aspect ratios in water bodies with different reflective spectral characteristics.

The past decade has witnessed the tremendous advances of deep-learning technologies [13, 14]. Since AlexNet was proposed in 2012, extensive researches have been devoted to designing advanced convolutional neural networks (CNNs). To have more powerful feature extraction abilities, networks are designed to be deeper and more complex [15–17]. But it is nontrivial to holistically redesign the network architecture. Thus, researchers propose some modules that can be easily embedded into existing networks to enhance network capabilities. These modules have played a positive role in various tasks and achieved impressive results. For instance, deformable convolutional modules are presented to solve the problem that regular CNNs are inherently limited to model geometric transformation [18, 19]. They have flexible receptive field shapes to fit the contours of objects, which are very effective for computer vision tasks. Attention modules implement adaptive weight adjustment within multi-dim feature maps. Channel attention modules model the channel-wise relationships and weight each channel. The pioneering work is Squeeze-and-excitation (SE) block [20]. Global average pooling is used to squeeze spatial information of the input feature. Full-connected layers and activation function layers are used to generate channel-wise weights. Most of the subsequent works follow the idea of SE to further enhance the power of channel modeling [21–23]. Different from channel attention modules, spatial attention modules focus on key spatial regions. RAM [24] is built on Recurrent Neural Network, and it can be trained by reinforcement learning methods. GENet [25] first aggregates feature responses across spatial neighborhoods and then modulates the input feature map according to the aggregated result. In addition, spatial attention mechanism can also be implemented by self-attention modules [26, 27].

These embedded modules perform well in various computer vision tasks such as image classification, object detection and semantic segmentation [28]. It is valuable to apply them to remote sensing tasks including bridge detection.

In computer vision community, deep-learning-based detection methods generally can be grouped into two categories: region proposal-based detectors and regression-based detectors. Region proposal-based detectors frame the detection as a "coarse-to-fine" process [29]. R-CNN [30] first adopts the convolutional neural network to extract features. The Region Proposal Network of Faster R-CNN [31] greatly improves the efficiency of region proposal. Cascade R-CNN [32] consists of a sequence of detectors trained with increasing intersection over union (IoU) thresholds to achieve better detection performance. While regression-based detectors achieve the detection task in one step [29]. YOLO series [17, 33, 34] are characterized by strong real-time performance and fast inference speed. CornerNet [35] obtains bounding boxes based on the prediction of top-left and bottom-right corner keypoints. CenterNet [36] returns the properties of the objects from center points. New deep-learning-based algorithms are constantly being proposed. But today, the classic Faster R-CNN and YOLO series are still widely adopted in industry because of their reliability.

Many detection algorithms based on CNNs have been adopted in the field of earth observation. For instance, YOLT [37] is proposed to detect small objects in satellite images. CAD-Net [38] combines global and local contexts, and spatial-and-scale-aware module is designed on the basis of feature pyramid structure. Especially, Nogueira *et al.* [39] compare the performances of several deep-learning-based detectors on their bridge detection datasets. Some works pay attention to detect bridges in SAR images [40, 41]. But only horizontal bounding boxes of bridges are predicted by these methods. Some algorithms achieve oriented object detection, such as SCRDet [42], SCRDet++ [43], and R^3*det* [44]. But a large number of oriented bounding box annotations are required by them.

Inspired by the progress in the fields of remote sensing and computer vision, this paper explores a efficient bridge-over-water detection scheme, as shown in Fig. 1. The scheme is driven by data and knowledge sequentially. Firstly, the backbone network of deep-learning-based detector is optimized by the proposed method of combining modulated deformable convolution (Mdconv) and attention mechanisms. Then, a post-processing guided by prior knowledge is presented. Horizontal bounding boxes (HBBs) are converted to oriented bounding boxes (OBBs).

Specifically, the main contributions of this paper can be summarized as follows:

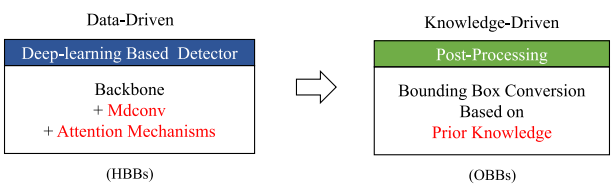


Fig. 1 Principle of the proposed scheme

1 The modulated deformable convolution is introduced in bridge-over-water detection. The flexible receptive field can adapt to various bridge shapes and contours.
2 The channel and spatial attention mechanism is further applied to improve the detection performance especially for small bridges. We design a weighted channel and spatial attention (WCSA) structure. The contributions of channel and spatial attention mechanisms at different stages of different backbone networks can be demonstrated. Moreover, we design a selective channel-spatial attention (SCSA) usage strategy. Our experiments demonstrate that the SCSA strategy can perform well in different detection architectures. The attention redundancies within backbone networks are effectively reduced.
3 To locate bridge-over-water more precisely, a post-processing is proposed to convert the HBBs to the OBBs, and we name it bounding box conversion module (BBCM). Our experiments demonstrate that prior knowledge is sufficient to guide the conversion process in most situations.

The remainder of the paper is organized as follows: Sect. 2 describes the proposed bridge-over-water detection scheme in detail. Sect. 3 reports and discusses the experimental results. Sect. 4 concludes this paper.

2 The proposed method

Figure 2 shows an overview of our bridge-over-water detection scheme. Data-driven method and knowledge-driven method are employed by our scheme successively. Firstly, the deep-learning-based detector with the improved backbone network predicts the HBBs of bridges. Deep network stack convolutional layers to obtain low/mid/high-level features, such as ResNet [15], DarkNet [17]. They naturally form in several stages. In this paper, the feature maps from the last 4 stages are further processed by other structures of the deep-learning-based detector. We embed modulated deformable convolutional

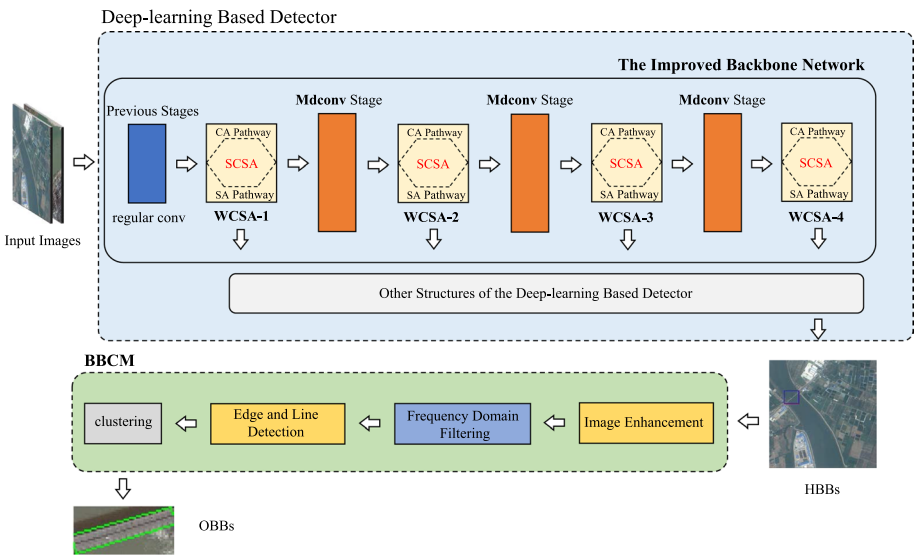


Fig. 2 Overview of the proposed scheme for bridge-over-water detection

layers in the last 3 stages. At the tail of the last 4 stages, we attempt to employ attention mechanism according to the SCSA strategy. The OBBs and the rotation angles of bridges are calculated by the presented BBCM, which contains frequency domain filtering, spatial domain operations and clustering.

2.1 Improved backbone network

In this paper, the proposed approaches are applicable to most backbone networks. For verification, we focus on the widely used DarkNet-53 (d53) in YOLOv3 architecture and ResNet-50 (r50) in Faster R-CNN with FPN [45] architecture. DarkNet-53 and ResNet-50 both stack convolutional layers to extract features. They both adopt the method of residual connection to mitigate the degradation problem of deep network. The hierarchical structures of our improved DarkNet-53 and improved ResNet-50 are illustrated in Fig. 3. The 5 stages of DarkNet-53 have 1, 2, 8, 8, and 4 building blocks, respectively. The 4 stages of ResNet-50 have 3, 4, 6, and 3 building blocks, respectively. The input image size is $H \times W$. We annotate the details of each layer including: the number of repetitions of the building blocks, the type of layer, the size of the output feature map, the number of convolutional filters, the kernel size and the stride. Compared with the original backbones, we replace some regular convolutional layers with the Mdconv layers and add some attention layers in this paper.

2.1.1 Modulated deformable convolution

The sampling grid of regular convolution is fixed shape. However, bridges over water bodies have arbitrary orientations and aspect ratios in remote sensing images. Too many irrelevant background pixels are included in the square receptive field of regular convolution. It becomes difficult for neural networks to learn object-related features.

DarkNet-53						ResNet-50					
	layer type	kernel size	stride	filters	output size	output size	filters	stride	kernel size	layer type	
	Conv	3×3	1	32	H×W	H/2×W/2	64	2	7×7	Conv	
	Conv	3×3	2	64	H/2×W/2	H/4×W/4	64	2	3×3	Max Pool	
1×	Conv	1×1	1	32			64	1	1×1	Conv	
	Conv	3×3	1	64			64	1	3×3	Conv	
	Residual				H/2×W/2		256	1	1×1	Conv	×3
	Conv	3×3	2	128	H/4×W/4	H/4×W/4	256			Residual	
	Conv	1×1	1	64		H/4×W/4				**Attention**	
2×	Conv	3×3	1	128			128	1 (2, first block)	1×1	Conv	
	Residual				H/4×W/4		128	1	3×3	**Mdconv**	
	Attention				H/4×W/4		512	1	1×1	Conv	×4
	Conv	3×3	2	256	H/8×W/8	H/8×W/8	512			Residual	
	Conv	1×1	1	128		H/8×W/8				**Attention**	
8×	**Mdconv**	3×3	1	256			256	1 (2, first block)	1×1	Conv	
	Residual				H/8×W/8		256	1	3×3	**Mdconv**	
	Attention				H/8×W/8		1024	1	1×1	Conv	×6
	Conv	3×3	2	512	H/16×W/16	H/16×W/16	1024			Residual	
	Conv	1×1	1	256		H/16×W/16				**Attention**	
8×	**Mdconv**	3×3	1	512			512	1 (2, first block)	1×1	Conv	
	Residual				H/16×W/16		512	1	3×3	**Mdconv**	
	Attention				H/16×W/16		2048	1	1×1	Conv	×3
	Conv	3×3	2	1024	H/32×W/32	H/32×W/32	2048			Residual	
	Conv	1×1	1	512		H/32×W/32				**Attention**	
4×	**Mdconv**	3×3	1	1024							
	Residual				H/32×W/32						
	Attention				H/32×W/32						

Fig. 3 The hierarchical structures of the improved DarkNet-53 and improved ResNet-50

The performance of deep-learning-based detector is inhibited. Therefore, modulated deformable convolution is introduced into our approach. On the basis of the regular convolution, the modulated deformable convolution adds an extra convolutional branch to regress offsets and modulation scalars of all sampling points. The process can be illustrated as:

$$f(k) = \sum_{n=1}^{N} w_n \cdot x(k + k_n + \Delta k_n) \cdot \Delta m_n \tag{1}$$

where N is the number of convolution sampling points. x and f, respectively, are input feature map and output feature map. k is the center position of all sampling points. $k + k_n$ denotes the n-th sampling position of the regular convolution. w_n represents the weight of convolutional kernel at n-th position. 3×3 convolution corresponds to $N = 9$ and $k_n \in (-1, 1), (-1, 0), ..., (1, 1)$. Δk_n denotes the adaptive offset, and the value of $x(k + k_n + \Delta k_n)$ is obtained by bilinear interpolation. Δm_n is the modulation scalar for the n-th position.

Though the performance obtained by Mdconv is better than the regular convolution mode, the computational cost is huge. It cannot be ignored that the computational cost of Mdconv is higher than that of regular convolution. Considering a trade-off between the detection performance and the computational cost, we decide to selectively use Mdconv in the network instead of using it in all stages. Some experiments about the Mdconv on COCO [46] benchmark were conducted by Zhu et al. [19]. According to their results, adopting the Mdconv in the last 3 stages is the best choice for Faster R-CNN architecture with the ResNet backbone. No additional improvement is observed by further replacing the regular convolutional layers in the first stage of ResNet. It is therefore insufficient to implement Mdconv only in the last stage. For bridge-over-water detection, we have also performed experiments using the Mdconv at different stages of ResNet in Faster R-CNN with FPN architecture. The results show that adopting the Mdconv in the last 3 stages is also the best choice. To maintain consistency, we follow the setting for the improved DarkNet-53 and also use the Mdconv in its last 3 stages. The replaced building blocks of DarkNet-53 and ResNet-50 are demonstrated in Fig. 4

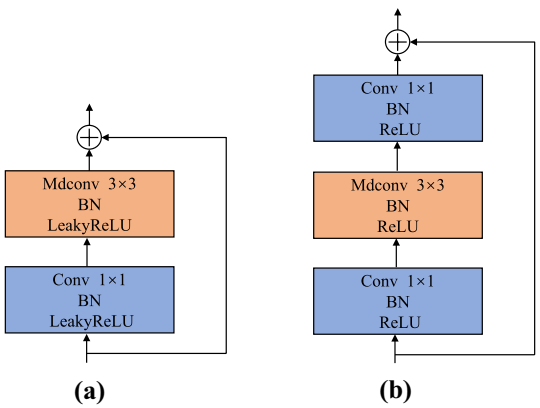


Fig. 4 The replaced building blocks: **a** is the building block embedded with Mdconv of the improved DarkNet-53; **b** is the building block embedded with Mdconv of the improved ResNet-50

a, b. We replace all 3×3 regular convolutional layers in the last 3 stages of the backbone with 3×3 Mdconv layers and keep the settings of other layers consistent with the original DarkNet-53 and ResNet-50, such as pointwise convolutional layers and activation functions.

2.1.2 WCSA structure and SCSA strategy

Various attention mechanisms have been proposed recently, and they can usually be expressed as

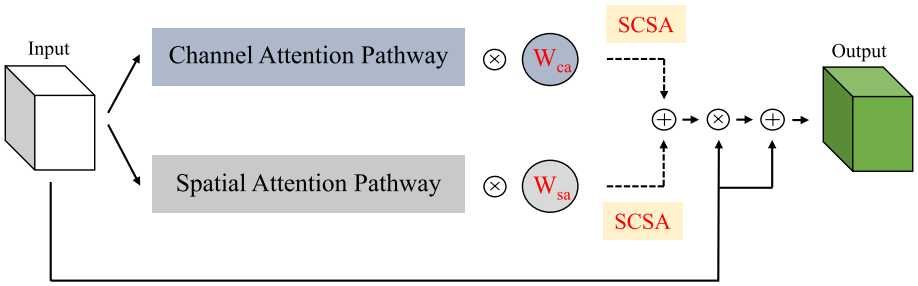
$$a = p(g(x), x) \tag{2}$$

where x represents the input of the attention module, a represents the output feature map. $g(x)$ denotes the process of generating attention weights, and $g(x)$ is adaptively adjusted through gradient backpropagation. $p(g(x), x)$ means that input x is processed by the attention weights.

Both channel and spatial attention mechanisms are beneficial for enhancing the network capabilities, and some researchers work on combining them in one module. CBAM and BAM both are widely used channel and spatial attention modules. CBAM [47] sequentially infers channel and spatial attention maps. Average-pooling and max-pooling are both adopted to describe feature. The channel-wise relationships are built by a shared multi-layer perceptron with one hidden layer, and the spatial-wise relationships are built by one convolutional layer with the kernel size of 7×7. Different from CBAM, BAM [48] arranges channel attention module and spatial attention module in parallel. Only average-pooling is used in the channel attention branch of BAM. Dilated convolution is applied by the spatial attention branch of BAM.

In this paper, for channel and spatial attention modules, we attempt to explore the necessity of two attention mechanisms at different stages of different backbone networks in this task. We hope that the attention mechanisms paired with the Mdconv can further improve the detection performance. The weighted channel and spatial attention structure is presented. The WCSA also sets the channel attention pathway and the spatial attention pathway in parallel. We add an adaptive weight to each pathway. Therefore, the network can more explicitly adjust the proportion of two attention mechanisms. By observing the changes of the pathway-weights, the roles of channel and spatial attention mechanisms in this task are transparent. As shown in Fig. 5, we build WCSA based on CBAM and BAM, respectively. For example, implementing WCSA based on CBAM, the two pathways of WCSA(CBAM) are, respectively, composed of the corresponding attention parts of CBAM. Because of the different design ideas and arrangements, CBAM and BAM may perform differently in various tasks. It is hard to judge theoretically which of them is better. And it is necessary to conduct experiments to explore their performance. We experimented with initializing all pathway-weights to 0, 0.5, and 1, respectively. The best results were obtained when they were initialized to 0.5, and this initial value was used in all our experiments.

We also design a selective channel-spatial attention usage strategy to eliminate redundancies and avoid interferences. The SCSA strategy can be described as:



Channel & Attention modules	CBAM (Serial)	BAM (Parallel)
Channel Attention Pathway	Avg Pool, Max Pool, MLP	Avg Pool, MLP
Spatial Attention Pathway	Avg Pool, Max Pool, Conv	Conv, Dilated Conv, Conv

Fig. 5 The WCSA structure and the SCSA strategy

1 when the training is completed, only the pathways with the corresponding weight values greater than 0.5 can be reserved.
2 If all pathway-weight values are less than 0.5, the 4 pathways with higher values can be retained.

In particular, for SCSA(CBAM), if both channel and spatial attention mechanisms are reserved at a certain stage, we choose to continue using the original serial CBAM at that stage. Because we set the initial value of all pathway-weights to 0.5 in our experiments, we choose 0.5 as the threshold in the SCSA strategy. We consider that the contribution of a attention pathway is negative if its pathway-weight value gradually becomes smaller than its initial value during training. Otherwise the contribution of this pathway is positive and the pathway should be preserved. Therefore, it is reasonable and intuitive to use 0.5 as the threshold in the SCSA strategy. The proposed strategy decides how to use or whether to use the attention mechanisms in each SCSA module according to the changes of all pathway-weight values. Under the constraint of the strategy, appending a SCSA module at the tail of each stage of ResNet-50 will not cause a large computational cost. Therefore, we choose employ the SCSA module after all 4 stages of ResNet-50. To maintain consistency, we also adopt the SCSA module after the last 4 stages of DarkNet-53.

2.2 Bounding box conversion

Because of the diversity of bridges' orientations in remote sensing images, using the axis-aligned bounding box to locate bridges is not accurate enough. Our scheme further converts them to OBBs. The deep-learning-based detector has greatly reduced the regions of interest in our scheme. The HBBs usually do not contain overly complex terrain textures in this task. In most cases, water with a variety of colors and textures is

the only element besides the bridge. Therefore, it is reasonable that a knowledge-driven method is sufficient for the position fine-tuning.

In the proposed BBCM, the edge characteristic of bridge-over-water is chosen as the key prior knowledge. Edges correspond to the high-frequency components in frequency domain and the pixels of large gradient in spatial domain. The edges of bridges often appear as spatial parallel lines. Therefore, frequency domain filtering and spatial line detection are combined in the BBCM. As shown in Fig. 6, we first convert the horizontal slices to grayscale and sharpen them. Next, they are converted to frequency domain, and some low-frequency components are removed. Concretely, we perform Fast Fourier Transform (FFT) and shift the zero-frequency component to the center of the spectrum firstly. And then, we remove the low-frequency components from the center region of the spectrum, which has an area of $\frac{H_s}{3} \times \frac{W_s}{3}$. Here, H_s and W_s are the height and width of the spectrum, respectively. After that, standard canny edge detection and Hough line detection are performed sequentially. To exclude the lines that do not belong to the edges of the bridge, the weighted K-means clustering is adopted. The lines from bridge edges should have a uniform angle. Slices usually only contain water bodies and the bridge, and longer lines are more likely to belong to bridge edges. So, we cluster the angles of detected lines, and the lengths of detected lines are used as the weights of corresponding angles. The lines of the main class are retained, and the OBB of the bridge is accessible according to the retained lines.

3 Results and discussion

Our experiments were conducted on a high-resolution optical remote sensing image dataset proposed for bridge-over-water detection. The resolution of each remote sensing image is in the range of 1–4 m. Each image of the dataset contains at least one bridge object, covering railway bridge, highway bridge, road-rail bridge, pedestrian bridge, water-carrying bridge, etc. There are about 2000 images and 4000 bridge-over-water instances. All instances are annotated by HBBs. In our experiments, training set, validation set, and test set were divided randomly in the proportion: 8:1:1. All experiments were conducted on an NVIDIA RTX3060. Minibatch and epoch were set to 2 and 84 in training. Stochastic gradient descent was chosen as the optimizer. The initial learning

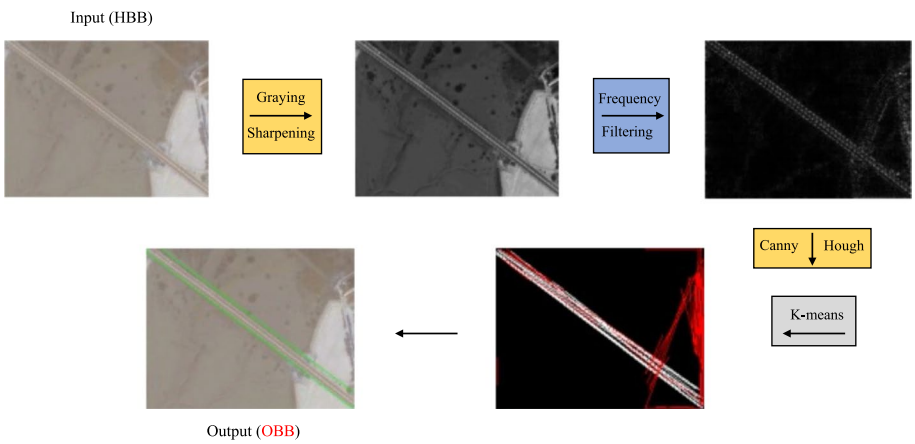


Fig. 6 Pipeline of the BBCM

rate was set to 0.0025, and learning rate decayed by a factor of 10 at the 48th and 68th epochs.

Average precision (AP), recall (Rec), and f1-score (F1S) were used to quantitatively evaluate the performance of all deep-learning-based detectors. We obeyed the standard COCO [46] evaluation method. For AP and Rec, the superscripts: "s", "m" and "l" mean the corresponding indicators for small, medium and large objects, respectively. It is worth noting that all AP indicators were calculated in $IoU = .50 : .05 : .95$, which means that the IoU threshold ranges from 0.5 to 0.95 with a step size of 0.05. These AP indicators calculated in $IoU = .50 : .05 : .95$ are the primary metrics in COCO evaluation. They are more objective than metrics based on a certain threshold, and a certain threshold would introduce bias in the evaluation. The results of the experiments performed on DarkNet-53 and YOLOv3 are shown in Tables 1 and 2. Tables 3 and 4 demonstrate the results of the experiments conducted on ResNet-50 and Faster R-CNN with FPN.

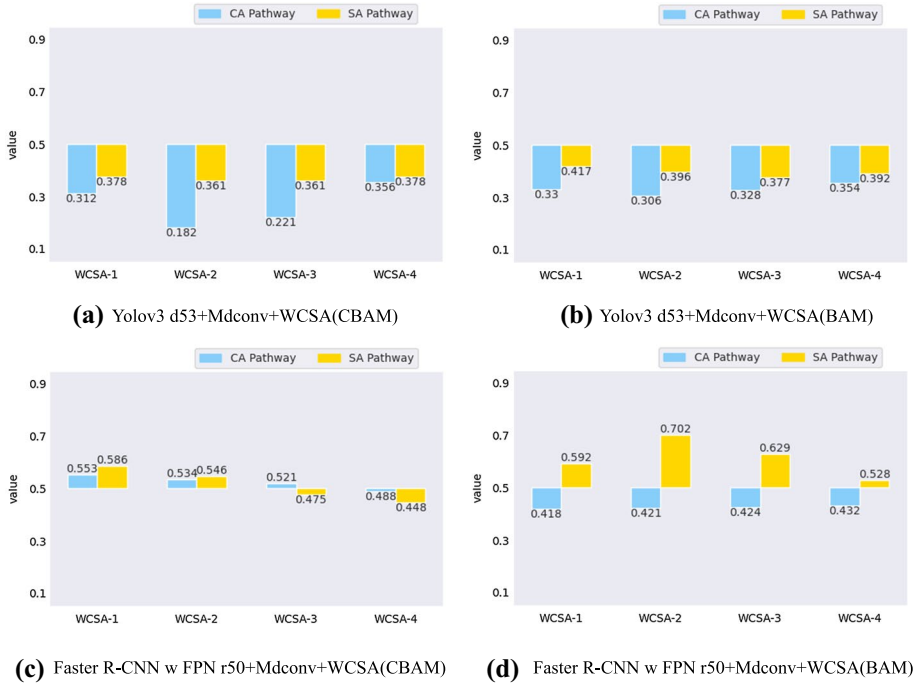
3.1 Effect of modulated deformable convolution on bridge detection performance

We first verify the effectiveness of modulated deformable convolution. It is obvious that modulated deformable convolution is able to greatly improve the performance of both architectures in bridge-over-water detection. Modulated deformable convolution dramatically boosts YOLOv3's ability to detect large objects, the AP^l and Rec^l increase by 29.6% and 12.6%, respectively. Its effect on Faster R-CNN with FPN is more moderate. The AP^m and Rec^m increase by 1.8% and 1.7%.

3.2 Effect of WCSA structure and SCSA strategy on bridge detection performance

On the basis of CBAM and BAM, the proposed WCSA modules are obtained by parallelizing and adding the pathway-weights. We first separately tested the original CBAM and BAM paired with Mdconv as benchmarks. Compared with only using Mdconv, attention mechanisms paired with Mdconv further improves the detection performance especially for small bridges. Subsequently, we performed experiments about our WCSA. We visualize the changes of all pathway-weight values in Fig. 7. As shown in Fig. 7a and b, we find that all pathway-weight values in YOLOv3 architecture have dropped. The weight values of channel attention pathways have dropped more severely. According to the proposed SCSA strategy, we only keep the spatial attention branches for both SCSA(CBAM) and SCSA(BAM) in YOLOv3. As for Faster R-CNN with FPN architecture, we find that WCSA(CBAM)-1, WCSA(CBAM)-2, and the channel attention pathway of WCSA(CBAM)-3 contribute more to detection performance in Fig. 7c; they are retained. According to Fig. 7d, the architecture tends to use the spatial attention mechanism instead of the channel attention mechanism in WCSA(BAM). Especially, the spatial attention pathway-weight value of WCSA(BAM)-2 increases from 0.5 to 0.702. Therefore, only spatial attention branches are reserved in SCSA(BAM).

We further conducted experiments to verify the proposed SCSA strategy. For YOLOv3, the combination of Mdconv and SCSA achieved the best results in terms of both AP and Rec, SCSA(CBAM) improves AP^l and Rec^l by 2% and 1.6%, respectively. SCSA(BAM) improves AP^s and Rec^s by 1.3% and 0.6%, respectively. For Faster R-CNN with FPN, SCSA strategy did not degrade the comprehensive performance, which proves that using channel and spatial attention mechanism at every stage of the backbone is



(a) Yolov3 d53+Mdconv+WCSA(CBAM) **(b)** Yolov3 d53+Mdconv+WCSA(BAM)

(c) Faster R-CNN w FPN r50+Mdconv+WCSA(CBAM) **(d)** Faster R-CNN w FPN r50+Mdconv+WCSA(BAM)

Fig. 7 The changes of pathway-weights. CA: channel attention; SA: spatial attention; d53: DarkNet-53; r50: ResNet-50

Table 1 Comparisons of average precision (AP) on DarkNet-53 and YOLOv3

Backbones	AP **(%)**	AP^s **(%)**	AP^m **(%)**	AP^l **(%)**
Plain	48.9	36.3	63.3	46.4
Plain+Mdconv	58.7	37.7	64.1	76.0
Plain+Mdconv+CBAM	62.1	43.4	66.1	78.9
Plain+Mdconv+WCSA(CBAM)	61.6	**44.2**	64.7	78.2
Pain+Mdconv+SCSA(CBAM)	**62.8**	43.3	**66.6**	**80.9**
Plain+Mdconv+BAM	62.1	43.2	**66.1**	78.7
Plain+Mdconv+WCSA(BAM)	62.1	42.7	66.0	**80.2**
Plain+Mdconv+SCSA(BAM)	**62.6**	**44.5**	66.0	79.2

The significance for bold values is only highlighted on the best performance

Table 2 Comparisons of recall (rec) and F1-score (F1S) on DarkNet-53 and YOLOv3

Backbones	Rec**(%)**	Rec^s**(%)**	Rec^m**(%)**	Rec^l**(%)**	$F1S$**(%)**
Plain	60.4	43.6	68.8	67.6	51.7
Plain+Mdconv	64.5	45.2	70.2	80.2	58.7
Plain+Mdconv+CBAM	68.2	51.3	**72.4**	83.1	**63.8**
Pain+Mdconv+WCSA(CBAM)	68.0	**52.3**	71.3	83.2	62.5
Plain+Mdconv+SCSA(CBAM)	**68.6**	51.3	**72.4**	**84.7**	62.7
Plain+Mdconv+BAM	68.1	51.7	72.2	82.6	62.4
Plain+Mdconv+WCSA(BAM)	68.1	50.6	72.2	**84.2**	63.0
Plain+Mdconv+SCSA(BAM)	**68.5**	**52.3**	**72.3**	83.5	**63.1**

The significance for bold values is only highlighted on the best performance

Table 3 Comparisons of average precision (AP) on ResNet-50 and Faster R-CNN with FPN

Backbones	AP**(%)**	AP^s**(%)**	AP^m**(%)**	AP^l**(%)**
Plain	68.4	48.4	72.1	85.8
Plain+Mdconv	69.3	48.6	**73.9**	87.2
Plain+Mdconv+CBAM	**70.1**	50.8	73.5	**88.6**
Plain+Mdconv+WCSA(CBAM)	69.7	**51.2**	73.6	86.4
Plain+Mdconv+SCSA(CBAM)	70.0	51.0	73.1	**88.6**
Pain+Mdconv+BAM	69.6	**50.6**	73.1	87.5
Plain+Mdconv+WCSA(BAM)	69.6	49.7	73.4	87.2
Plain+Mdconv+SCSA(BAM)	**69.8**	50.3	73.3	**87.7**

The significance for bold values is only highlighted on the best performance

Table 4 Comparisons of recall (rec) and F1-score (F1S) on ResNet-50 and faster R-CNN with FPN

Backbones	Rec**(%)**	Rec^s**(%)**	Rec^m**(%)**	Rec^l**(%)**	$F1S$**(%)**
Plain	72.7	54.5	76.9	89.5	71.8
Plain+Mdconv	73.7	54.7	**78.6**	90.4	72.8
Plain+Mdconv+CBAM	**75.0**	57.8	**78.6**	91.6	74.5
Plain+Mdconv+WCSA(CBAM)	74.5	57.8	78.4	89.9	**74.7**
Plain+Mdconv+SCSA(CBAM)	74.9	**58.4**	77.9	**91.9**	**74.7**
Plain+Mdconv+BAM	**74.6**	**57.2**	**78.6**	90.8	74.5
Plain+Mdconv+WCSA(BAM)	74.2	57.0	78.3	89.9	74.1
Plain+Mdconv+SCSA(BAM)	**74.6**	**57.2**	78.4	**90.9**	**74.6**

The significance for bold values is only highlighted on the best performance

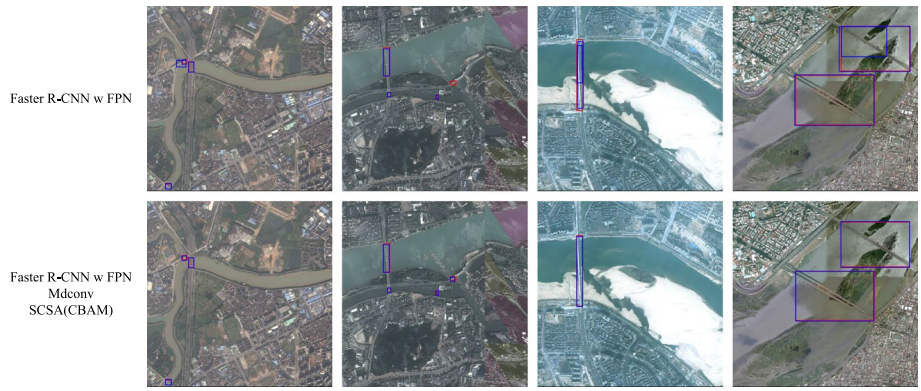


Fig. 8 Examples of test results

redundant. On the whole, SCSA strategy simplifies the network structure. It does not damage the comprehensive performance and achieves breakthroughs in some indicators.

3.3 Visualization of bridge detection results

To intuitively demonstrate the validity of the combination of Mdconv and SCSA, partial experimental results in Faster R-CNN with FPN architecture are shown in Fig. 8. The red rectangles represent the ground truth HBBs, and the predicted results are shown in blue rectangles. It is obvious that our method is able to accurately detect bridges over water bodies. The missing detection rate and false detection rate are reduced effectively.

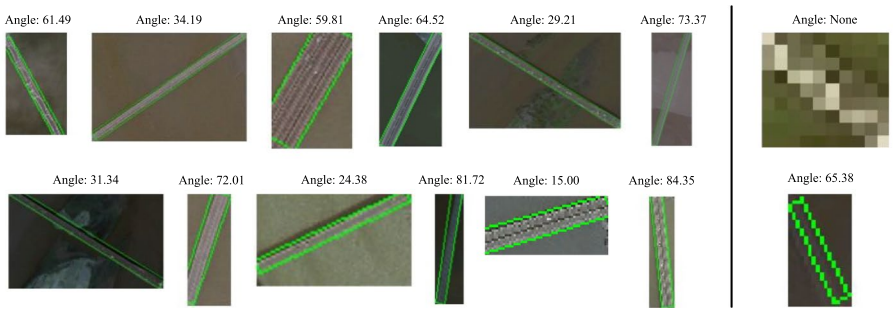


Fig. 9 Examples of bounding box conversion results

Partial bounding box conversion results of the proposed BBCM are demonstrated in Fig. 9. The green rectangles are the OBBs predicted by our BBCM, and the bridge rotation angles relative to the horizontal axis are indicated in degrees. Most of the calculated OBBs are high quality. However, it is difficult for our knowledge-driven method to handle a few tiny bridges, as shown in the right part of Fig. 9. The lower resolution provides very limited information.

4 Conclusion

In this paper, we propose a new bridge-over-water detection scheme driven by data and prior knowledge. An approach of improving the backbone network and a post-processing for bounding box conversion are presented. Modulated deformable convolution and attention mechanisms are introduced to enhance the detection performance of the deep-learning-based detector. A weighted channel and spatial attention structure is designed to analyze the degree of dependence on channel and spatial attention mechanisms within the backbone network. It can be concluded from our experiments that channel attention mechanism is less important than spatial attention mechanism for DarkNet-53 with Mdconv in YOLOv3. The proposed selective channel-spatial attention usage strategy is able to effectively eliminate the redundancy of attention mechanisms while maintaining the comprehensive performance of the detector. In addition, there is no need for oriented bounding box annotations, and our scheme can predict the precise position of bridge-over-water through the proposed post-processing module in most scenarios. The validity and generality of the proposed scheme are verified by our experiments. However, detecting tiny bridges in environments with drastic changes in color and surface texture remains a challenge.

Abbreviations

IoU:	Intersection over union
Mdconv:	Modulated deformable convolution
HBB:	Horizontal bounding box
OBB:	Oriented bounding box
WCSA:	Weighted channel and spatial attention
SCSA:	Selective channel-spatial attention
BBCM:	Bounding box conversion module
CBAM:	Convolutional block attention module
BAM:	Bottleneck attention module
AP:	Average precision
Rec:	Recall
F1S:	F1-score

Acknowledgements
Not applicable.

Author contributions
All authors have contributed toward this work as well as in compilation of this manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by National Natural Science Fund of China under Grant 61971324 and 61525105, the fund of National Lab of Radar Signal Processing. E-mail: dongganggang@xidian.edu.cn.

Availability of data and materials
The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 28 March 2022 Accepted: 5 June 2022
Published online: 29 June 2022

References

1. F. Biondi, P. Addabbo, S.L. Ullo, C. Clemente, D. Orlando, Perspectives on the structural health monitoring of bridges by synthetic aperture radar. Remote Sens. **12**(23), 3852 (2020)
2. C. Chen, J. Fu, Y. Gai, J. Li, L. Chen, V.S. Mantravadi, A. Tan, Damaged bridges over water: using high-spatial-resolution remote-sensing images for recognition, detection, and assessment. IEEE Geosci. Remote Sens. Mag. **6**(3), 69–85 (2018)
3. H. Biao, L. Ying, J. Licheng, Segmentation and recognition of bridges in high resolution sar images. In: 2001 CIE International Conference on Radar Proceedings (Cat No. 01TH8559), pp. 479–482 (2001). IEEE
4. F. Wu, C. Wang, H. Zhang, *Recognition of bridges by integrating satellite sarand optical imagery* Recognition of bridges by integrating satellite sarand optical imagery. (IEEE, 2005), pp. 3939–3940
5. G. Sithole, G. Vosselman, Bridge detection in airborne laser scanner data. ISPRS J. Photogramm. Remote Sens. **61**(1), 33–46 (2006)
6. N. Loménie, J. Barbeau, R. Trias-Sanz, *Integrating textural and geometric information for an automatic bridge detection system* In: IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings. (IEEE, 2003), pp. 3952–3954
7. Y. Han, H. Zheng, Q. Cao, Y. Wang, An effective method for bridge detection from satellite imagery. In: 2007 2nd IEEE Conference on industrial electronics and applications, pp. 2753–2757 (2007). IEEE
8. J. Luo, D. Ming, W. Liu, Z. Shen, M. Wang, H. Sheng, Extraction of bridges over water from ikonos panchromatic data. International journal of remote sensing **28**(16), 3633–3648 (2007)
9. Y. Zhao, S. Yu, J. Wu, L. Han, Z. Chen, Yang, X. Zhao, B. Data-driven bridge detection in compressed domain from panchromatic satellite imagery. In: international symposium on neural networks, pp. 449–458 (2014). Springer
10. C. Chen, Q. Qin, N. Zhang, J. Li, L. Chen, J. Wang, X. Qin, X. Yang, Extraction of bridges over water from high-resolution optical remote-sensing images based on mathematical morphology. Int. J. Remote Sens. **35**(10), 3664–3682 (2014)
11. E. Gedik, U. Cinar, E. Karaman, Y. Yardimci, U. Halici, K. Pakin, A new robust method for bridge detection from high resolution electro-optic satellite images. In: Proceedings of the 4th GEOBIA, 298–302 (2012)
12. B.-C. Gao, Ndwi-a normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens. Environ. **58**(3), 257–266 (1996)
13. A. Li, X. Zhu, S. He, J. Xia, Water surface object detection using panoramic vision based on improved single-shot multibox detector. EURASIP J. Adv. Signal Process. **2021**(1), 1–15 (2021)
14. L. Guanglong, Z. Zhu, B. Yongqiang, L. Tingna, X. Zhibo, Psenet-based efficient scene text detection. EURASIP Journal on Advances in Signal Processing **2021**(1) (2021)
15. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp. 770–778 (2016)
16. S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp. 1492–1500 (2017)
17. J. Redmon, A. Farhadi, Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
18. J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks. In: Proceedings of the IEEE International Conference on computer vision, pp. 764–773 (2017)
19. X. Zhu, H. Hu, S. Lin, J. Dai, Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pp. 9308–9316 (2019)

20. J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on computer vision and pattern recognition, pp. 7132–7141 (2018)
21. Z. Gao, J. Xie, Q. Wang, P. Li, Global second-order pooling convolutional networks. In Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pp. 3024–3033 (2019)
22. Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, Eca-net: efficient channel attention for deep convolutional neural networks, 2020 ieee. In: CVF Conference on computer vision and pattern recognition (CVPR). IEEE (2020)
23. H. Lee, H.-E. Kim, H. Nam, Srm A style-based recalibration module for convolutional neural networks. In Proceedings of the IEEE/CVF International conference on computer vision, pp. 1854–1862 (2019)
24. V. Mnih, N. Heess, A. Graves, Recurrent models of visual attention. In Advances in neural information processing systems, pp. 2204–2212 (2014)
25. J. Hu, L. Shen, S. Albanie, G. Sun, Vedaldi, A. Gather-excite Exploiting feature context in convolutional neural networks. arXiv preprint arXiv:1810.12348 (2018)
26. X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks. In Proceedings of the IEEE Conference on computer vision and pattern recognition, pp. 7794–7803 (2018)
27. Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, W. Liu, Ccnet Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International conference on computer vision, pp. 603–612 (2019)
28. M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R.R. Martin, M.-M. Cheng,S.-M. Hu, Attention mechanisms in computer vision A survey. arXiv preprint arXiv:2111.07624 (2021)
29. Z. Zou, Z. Shi, Y. Guo, J. Ye, Object detection in 20 years A survey. arXiv preprint arXiv:1905.05055 (2019)
30. R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on computer vision and pattern Recognition, pp. 580–587 (2014)
31. S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2016)
32. Z. Cai, N. Vasconcelos, Cascade r-cnn Delving into high quality object detection. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp. 6154–6162 (2018)
33. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp. 779–788 (2016)
34. A. Bochkovskiy, C.-Y. Wang, H.-Y.M. Liao, Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
35. H. Law, J. Deng, Cornernet Detecting objects as paired keypoints. In: Proceedings of the European Conference on computer vision (ECCV), pp. 734–750 (2018)
36. X. Zhou, D. Wang, P. Krähenbühl, Objects as points. arXiv preprint arXiv:1904.07850 (2019)
37. A. Van Etten, You only look twice Rapid multi-scale object detection in satellite imagery. arXiv preprint arXiv:1805.09512 (2018)
38. G. Zhang, S. Lu, W. Zhang, Cad-net A context-aware detection network for objects in remote sensing imagery. IEEE Trans. Geosci. Remote Sens. **57**(12), 10015–10024 (2019)
39. K. Nogueira, C. Cesar, P.H. Gama, G.L. Machado, dos Santos, J.A. A tool for bridge detection in major infrastructure works using satellite images. In: 2019 XV Workshop de Visão Computacional (WVC), pp. 72–77 (2019). IEEE
40. L. Chen, T. Weng, J. Xing, Z. Pan, Z. Yuan, X. Xing, P. Zhang, A new deep learning network for automatic bridge detection from sar images based on balanced and attention mechanism. Remote Sens. **12**(3), 441 (2020)
41. L. Chen, T. Weng, J. Xing, Z. Li, Z. Yuan, Z. Pan, S. Tan, R. Luo, Employing deep learning for automatic river bridge detection from sar images based on adaptively effective feature fusion. Int. J. Appl. Earth Observ. Geoinform. **102**, 102425 (2021)
42. X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, X. Sun, K. Fu, Scrdet Towards more robust detection for small, cluttered and rotated objects. In Proceedings of the IEEE/CVF International Conference on computer vision, pp. 8232–8241 (2019)
43. X. Yang, J. Yan, X. Yang, J. Tang, W. Liao, T. He, Scrdet++ Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. arXiv preprint arXiv:2004.13316 (2020)
44. X. Yang, Q. Liu, J. Yan, A. Li, Z. Zhang, G. Yu, R3det Refined single-stage detector with feature refinement for rotating object. arXiv preprint arXiv:1908.05612**2**(4) (2019)
45. T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on computer vision and pattern recognition, pp. 2117–2125 (2017)
46. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco Common objects in context. In: European Conference on computer vision, pp. 740–755 (2014). Springer
47. S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam Convolutional block attention module. In Proceedings of the European Conference on computer vision (ECCV), pp. 3–19 (2018)
48. J. Park, S. Woo, J.-Y. Lee, I.S. Kweon, Bam Bottleneck attention module. arXiv preprint arXiv:1807.06514 (2018)

Publisher's Note