Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 6317691, 11 pages
https://doi.org/10.1155/2022/6317691

Research Article

A Surface Target Recognition Algorithm Based on Coordinate Attention and Double-Layer Cascade

Runze Guo, Zhen Zuo, Shaojing Su, and Bei Sun

College of Intelligence Science Technology, National University of Defense Technology, Changsha 410000, China

Correspondence should be addressed to Zhen Zuo; z.zuo@nudt.edu.cn

Received 28 February 2022; Accepted 16 April 2022; Published 10 May 2022

Academic Editor: Kuruva Lakshmanna

As a branch of target recognition, surface target recognition plays an irreplaceable role in both military and civilian applications. However, the large target size variation, low image resolution, and high real-time requirements pose challenges to existing algorithms. To address the issues, we take YOLOv5 as a backbone and adopt coordinate attention and a double-layer cascade structure to enhance both the recognition performance and speed. Specifically, coordinate attention is introduced to guide the corresponding network to focus on discriminative features by capturing channel and location information. Meanwhile, the double-layer cascade structure is designed for finely extracting and aggregating semantic features and spatial features at different scales. We test the model on the COCO dataset, the VOC dataset, and self-built surface target dataset. Experimental results show that proposed coordinate attention module and multiscale module improve the recognition effect of multiscale surface targets and meet the requirement of real time.

1. Introduction

With the growing development of marine technology, the types and numbers of ships are increasing, and the performed tasks are more dangerous and complex. Both the processing of maritime emergencies and the construction of intelligent shipping put forward higher requirements for surface target recognition [1, 2]. At present, surface target identification has been the key technology to environmental sensing for unmanned surface vehicles (USVs) [3]. In the military field, surface target recognition is an important part of marine environment reconnaissance, precision targeting, and other tasks. In the civilian field, accurate recognition plays an irreplaceable role in water personnel rescue, obstacle detection, etc.

Surface target recognition essentially belongs to target recognition. In recent years, deep convolutional neural networks (CNNs) have made great progress in the field of object detection and recognition [4] and have been successfully applied in medical diagnosis, face recognition, etc. Continuously updated network and more sophisticated big data technology perform increasingly well on public datasets. However, it is still difficult to recognize surface targets in complex climatic situations. First, there are many types of surface targets with various postures. Their interclass differences are small and intraclass differences are large [5]. Simply extracting traditional features is no longer sufficient for practical needs. Second, due to the inconsistent size and sampling distance of the surface target, the scale of targets spans a large. The aspect ratio of the bounding box under different angles is variable. In addition, the water surface is a highly reflective surface. The quality of images is more susceptible to the influence of weather conditions and background, resulting in low resolution, blurred edges, and easy confusion. Finally, surface target recognition technology is mainly used in the environment perception tasks that have more stringent requirements on the real time [6, 7]. As shown in Figure 1, the three images indicated by the orange arrows are actually the same target, yet the sizes are completely different.

In fact, the difficulty of surface target recognition exists mainly in the extraction of discriminative features and the recognition of multiscale targets. The key to extracting discriminative features is to focus on the detailed information that is beneficial to classification. The main method is attention

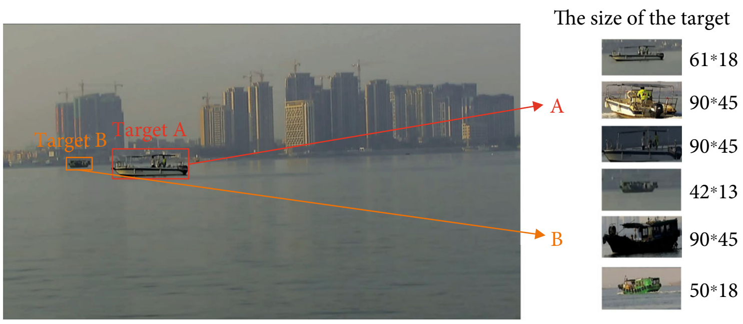

Figure 1: Real water scenes. The first three images pointed to by the red arrows are different samples of the target A (yachts). The last three images are samples of the target B (fish boats). The sizes of targets are marked out.

mechanisms [8]. Currently, there are two main types of existing methods to solve the multiscale problem: image pyramid and feature pyramid [9]. However, direct use will increase the network overhead and cannot meet the requirement of real-time algorithm.

Therefore, YOLOv5 is used as the basic framework to ensure the speed of this algorithm in this paper. First, coordinated attention is introduced to effectively capture channel and location information. Unlike previous attention methods, it learns correlations between channels without additional computational overhead. Then, a double-layer cascade is used to stitch and enhance the feature maps at different scales through maximum pooling and parameter aggregation. To prove the superiority of this algorithm, extensive experiments are conducted on the COCO dataset, the VOC dataset, and self-built surface target dataset. Experimental results show that our network performs better than other methods on multiscale surface targets and meet the real-time requirements.

The contributions of this paper are summarized as three points: (1) in order to improve the real-time performance of the algorithm, the single-stage detection algorithm YOLOv5 is used as the basic framework. (2) Coordinate attention is introduced to capture the channel and location information of the network. (3) The double-layer cascade structure is used to improve the recognition ability of multiscale targets. The organisation of the proposed work is as follows: the introduction is presented in Section 1. The related works is described in Section 2. The implementation method of this paper is illustrated in Section 3. The experimental results and analysis are described in Section 4. Section 5 describes the conclusion and prospect of this paper.

2. Related Works

As the core and key technology of USV sensing, surface target recognition is a very challenging task. Especially, the study under complex climate conditions has more theoretical and practical significance [10]. The process of surface recognition is shown in Figure 2. It needs to return the classification of the target contained on the image or video and needs the support from big data. The difference is that it also requires the target's position, whereas target recognition does not. In this section, we give a literature review, including target detection and recognition methods based deep learning, attention mechanisms, and multiscale methods.

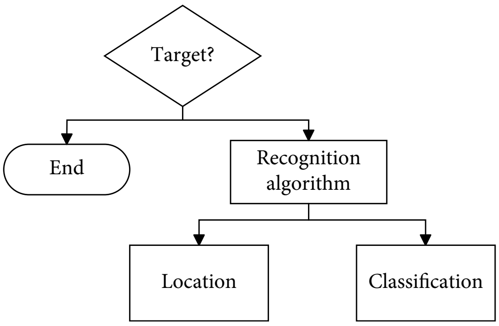

Figure 2: The process of surface target recognition.

2.1. Target Detection and Recognition Methods. Compared with traditional methods, target detection and recognition methods based on deep learning have significantly improved in terms of accuracy and generalizability. According to the number of stages, the deep learning-based target detection and recognition algorithms are divided into the two-stage method and the single-stage method. The former first extracts the borders of possible candidate regions and then inputs them into the region of interest (ROI) pooling layer together with the feature map, the advantage of which is high accuracy. The latter can directly regress the target category by delineating the selected frames according to the feature map, the advantage of which is fast speed.

The study was started with the two-stage approach. R-CNN [11] was the first method to introduce deep learning into the field of object detection and achieve adaptive learning, which was subsequently improved by many researchers. SPP-Net [12] introduces spatial pyramid pooling into R-CNN to reduce the impact of the size on the network. Fast R-CNN [13] uses ROI pooling based on the layer of SPP and achieves end-to-end training, which mainly improves the speed of the model. R-FCN [14] reduces the workload required for each ROI by constructing location-sensitive score maps to achieve speedup.

On the other hand, SSD [15] and YOLO [16–18] are typical one-stage approaches, also known as classification regression-based models. They are designed to directly

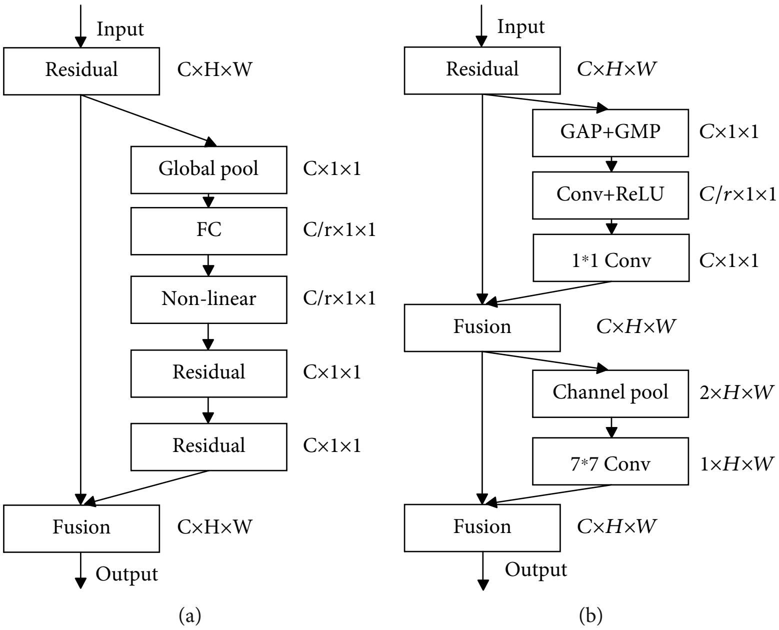

FIGURE 3: The structure of SE and CBAM. (a) SE. (b) CBAM.

classify and train predefined anchors without a proposal generation step. SSD draws on the anchor mechanism and regression idea of faster R-CNN in its design, with six (or four) default boxes at each pixel point of the feature map. YOLO divides the feature map into a 7×7 grid and then regresses the corresponding default boxes directly, so the speed is fast. However, YOLO [16] does not introduce multiscale information. It is difficult to obtain sufficiently rich target localization information when dealing with multitarget recognition. YOLOv2 [17] and YOLOv3 [18] introduce the anchor point mechanism, which improves the detection and recognition accuracy. To ensure the real time of the algorithm, we take YOLOv5 as the main framework.

2.2. Attention Mechanisms. Like human vision mechanisms, attention mechanisms in deep learning tend to focus on key information and ignore irrelevant information. They have been proven to be beneficial to a range of computer vision tasks. SENet [19] and CBAM [20] are typical networks applying attention mechanisms, the structures of which are illustrated in Figure 3. SENet focuses on the channel features of targets. It compresses the feature map and learns the interrelationships between channels. CBAM uses convolution with large size kernels, and combines spatial and channel features. The reason for the popularity of the self-attention networks, including NLnet [21], GCNet [22], and A2Net [23], is that they have the ability to compute in parallel and learn better about distant dependencies. Nonlocal mechanisms are also important to critical information. Later works, such as Genet [24], AA [25], and TA [26], continue to progress by designing different attention modules or fusion of different information.

However, SE and CBAM do not learn the importance of positional relationships and correlations between different channels. Self-attention is not applicable to surface target recognition task due to its large computational effort. Therefore, we choose an attention method that learns channel relations and channel dependencies called coordinate attention.

2.3. Multiscale Methods. Multiscale is one of the major differences between surface target recognition tasks and other vision tasks. Large-scale targets are generally easier to detect and recognize due to their large area and enriched feature. Small-scale targets, with fewer features and less resolution, are more difficult to locate and recognize accurately, but they occupy a proportion in images. In a practical application scenario, the scale of the target is measured by the ratio of the target size to the image size.

As a challenging problem in target detection and recognition, the variation of target scales affects the accuracy and speed of the model. There are two main types of methods to deal with the multiscale problem in vision tasks: image pyramid and feature pyramid. In image pyramid, images are scaled at different scales and then directly input to the detector. Based on the image pyramid approach, SNIP [27] selects different proposals for different resolutions to perform gradient propagation in the multiscale training process. SNIPER [28] crops images around the ground truth box on the feature map and selects the context region. However, SNIP and SNIPER still suffer from an inevitable increase in inference time during use.

The idea of feature pyramid is to approximate the image pyramid directly at the feature level. At the beginning, MS-CNN [29] handles objects of different sizes directly on different downsampled layers. Subsequently, TDM [30] and FPN [9] add new top-down branches to supplement the lack of semantic information at the bottom layer, both of which are the continuation of the feature pyramid approach. PANET [31] enhances the feature hierarchy representation

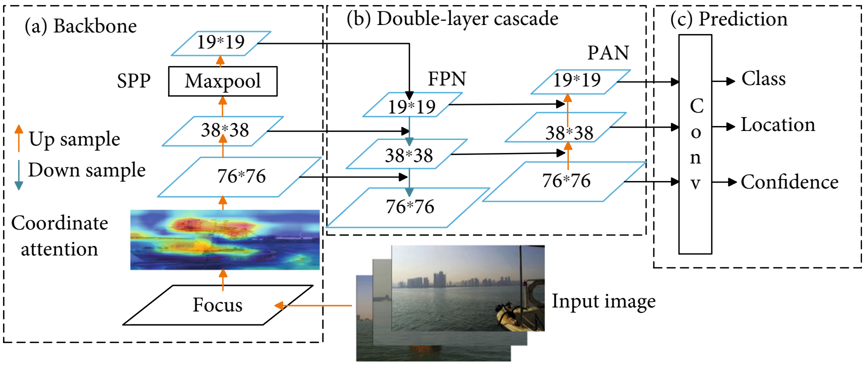

FIGURE 4: Illustration of our model. (a) The backbone designed for extracting multiscale features. (b) The double-layer cascade for the fusion of information at different scales. (c) the prediction block for the output of class, location, and confidence.

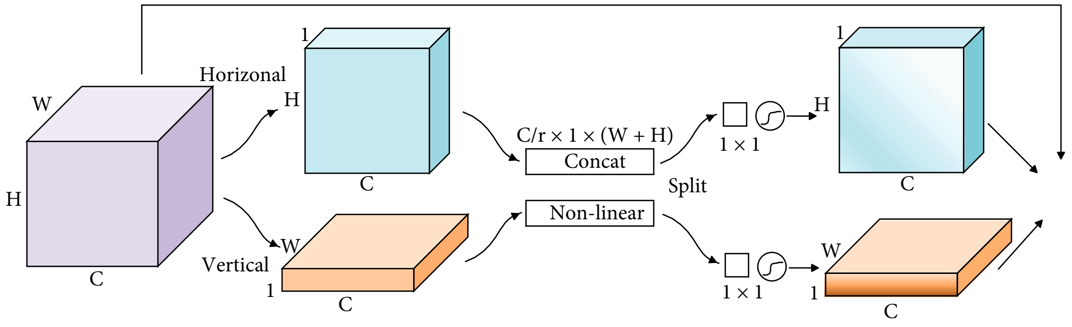

FIGURE 5: The structure of coordinate attention.

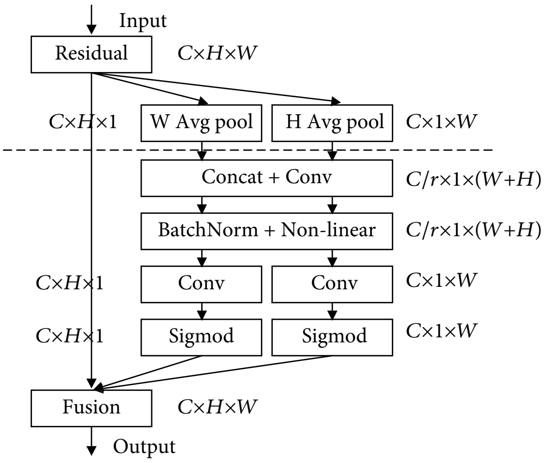

FIGURE 6: The illustration of coordinate attention. The dimensions of each step are marked.

with additional bottom-up paths and proposes adaptive feature pooling to aggregate features from various scales.

3. Methodology

To meet the real-time requirements of the algorithm, this paper uses YOLOv5 as the main framework. First, coordinate attention is added to focus on key information. Then, a double-layer cascade is added to solve the problem caused by multiscale and low-resolution targets. As shown in Figure 4, the structure of our model is divided into three parts, the backbone, the double-layer cascade, and the prediction.

3.1. Coordinate Attention. Many attention mechanisms are used in deep CNNs and bring great improvement on the performance of the network, but these mechanisms are significantly lagging when used in small networks. The reason is that the computational overhead of most attention mechanisms is not affordable for small networks [32]. The common attention mechanisms are SE, BAM, and CBAM. SE only considers the internal channel information and ignores the importance of location information. BAM and CBAM try to introduce location information as the basic of SE, but they fail in learning correlation through channels that is critical in identification tasks [33]. Therefore, this paper introduces an efficient and lightweight attention mechanism called coordinate attention which embeds location information into channel attention. Figure 5 illustrates the structure of the coordinate attention.

Generally, the attention module can be considered as a computational unit that is used to enhance the feature representation of the network. In coordinate attention, the channel attention is split into two parallel one-dimensional features encoding processes along the vertical and horizontal directions, respectively, to mitigate the loss of location information caused by global pooling. These two feature maps are

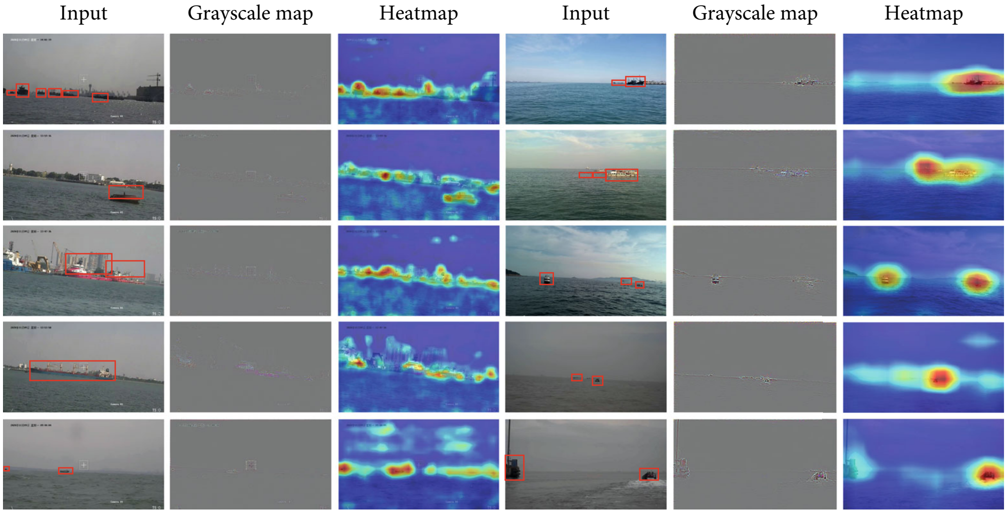

Figure 7: The visualization of our proposed model including saliency maps and heat map. Targets are marked out by red rows in original images.

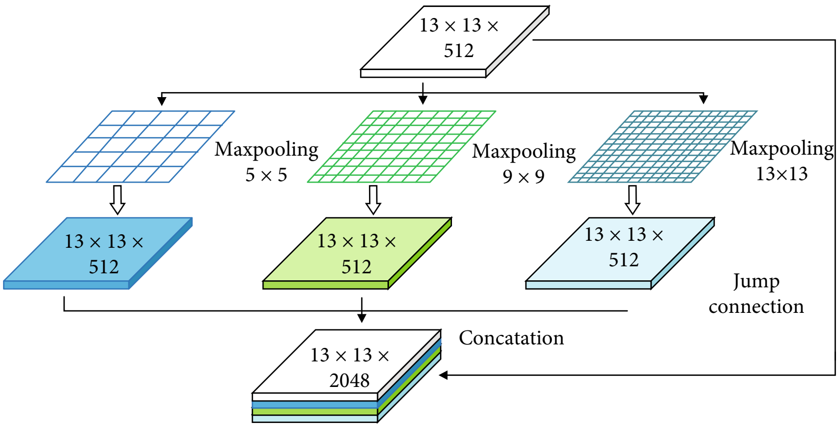

Figure 8: The structure of SPP block. Here, 4 different paths are concated.

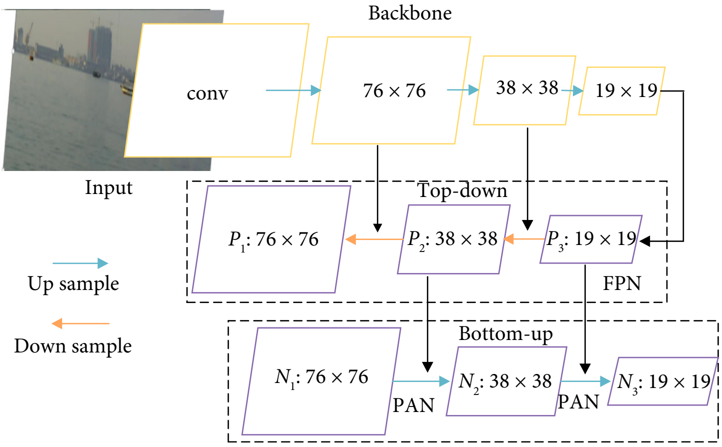

Figure 9: The double-layer network. Arrows of different colors and directions are used to indicate up sampling and downsampling, respectively.

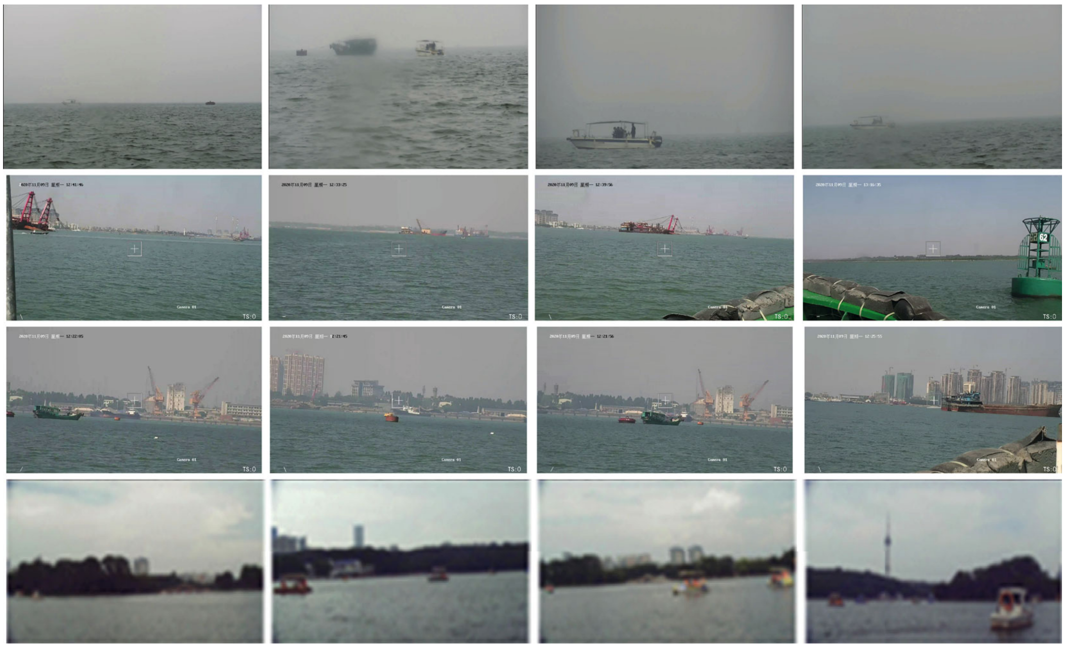

Figure 10: Part of our self-built surface target dataset. From top to bottom, the scenes are the distant sea, the near sea, the coast, and the lake in order.

Table 1: The definition of targets in different sizes.

Types	COCO	Self -built surface target (1920 ∗ 1080)
Small targets	≤32 ∗ 32	w ∗ h ≤ 1867
Medium targets	32 ∗ 32 ~ 96 ∗ 96	1867 < w ∗ h < 7465
Large targets	≥96 ∗ 96	w ∗ h ≥ 7465

embedded with different orientation information and encoded as the attention map. Thus, the location information can be stored in the attention map. We divide the process of coordinate attention encoding into two steps: coordinate information embedding and coordinate attention generation. As shown in Figure 6, we mark the dimensions of the tensor in each step.

3.1.1. Coordinate Information Embedding. To encode channel relations and position correlations, the global pooling as formulated in Equation (1) is divided into two one-dimensional encoding operations. Given the input feature $X \in \mathbb{R}^{C\times H\times W}$, the vertical and horizontal coordinates are encoded separately by using different kernels $(H, 1)$ and $(1, W)$, and the outputs are denoted as

$$z_c = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_c(i, j), \tag{1}$$

$$z_c^h(h) = \frac{1}{W}\sum_{0\le i<W} x_c(h, i), \tag{2}$$

$$z_c^w(w) = \frac{1}{H}\sum_{0\le j<H} x_c(j, w). \tag{3}$$

The above two variations (2) and (3) are along two directions, and a pair of feature maps is generated. These two feature maps allow the attention module to learn feature correlations in one spatial direction while retaining position information in the other, which helps to accurately identify and locate surface targets.

3.1.2. Coordinate Information Generation. The $z_c^h(h)$ and $z_c^w(w)$ generated in the first step are concatenated and fed into a shared 1×1 convolutional transformation function C_1, generating

$$c = \delta\left(C_1\left(\left[z^h, z^w\right]\right)\right). \tag{4}$$

Here $[\cdot, \cdot]$ denotes the concatenation operation, δ denotes the nonlinear activation function, and $c \in \mathbb{R}^{C/r\times(W+H)}$ represents the feature map containing two directions of encoded information. We use r as the scaling ratio to control the network overhead. Through experiments, r is set as 24, which can balance the accuracy and the speed. After that, c is decomposed into two independent tensors $c^h \in \mathbb{R}^{C/r\times H}$ and $c^w \in \mathbb{R}^{C/r\times W}$. These two are converted to tensors (g^h, g^h) with the same channel number as X by two 1×1 the convolutional functions C_h and C_w, respectively.

$$g^h = \sigma\left(C_h\left(c^h\right)\right),$$
$$g^w = \sigma(C_w(c^w)). \tag{5}$$

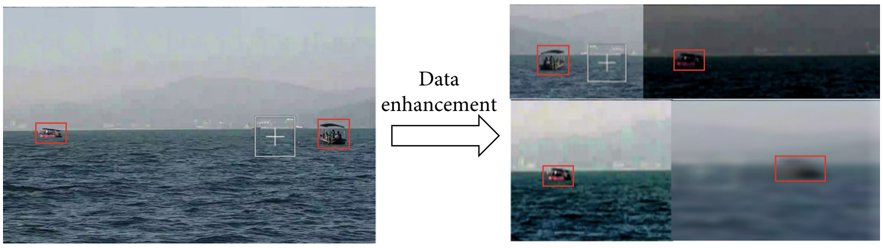

Figure 11: The illustration of our method of data enhancement. The size of the two images before and after the transformation is the same.

Table 2: Result comparisons under different settings.

Setting	r	mAP	AP_S	AP_M	AP_L	FPS
Baseline	/	78.9	65.5	82.4	89.2	26.5
+SE	24	77.5	62.3	82.1	90.5	15.9
+CBAM	24	80.6	67.2	86.7	88.6	16.3
+HA	24	79.6	66.4	86.3	90.3	22.6
+VA	24	80.2	67.9	85.2	89.2	23.7
+CA	24	80.4	68.5	86.1	89.7	22.4
+DC	/	81.2	68.2	87.5	88.5	21.5
CA+DC	24	81.9	69.2	87.3	89.5	20.8

Here, σ is the sigmoid function. g^h and g^h are expanded and treated as attention weights. Finally, the output of coordinate attention $C(i, j)$ is represented as

$$C(i, j) = x(i, j) \times g_c^h(i) \times g_c^w(j). \tag{6}$$

In order to have an intuitive understanding of the coordinate attention, we visualize the greyscale maps and heat maps. As can be seen in Figure 7, the greyscale map roughly reflects the overall profile of the surface target. On the other hand, the heat map filters out a large amount of irrelevant information and focuses on useful information. The darker the color on the map, the more significant the role of the area for classification.

3.2. Double-Layer Cascade. Multiscale means sampling the target at different granularities. In general, smaller and denser sampling in granularity allows more details to be seen, while larger and sparser sampling allows the overall contour and shape to be seen. Distance variation and sensor zoom are the main physical reasons for the variable scale properties of the target over the image domain. With the distance of target from near to far, the high-frequency detail texture information gradually declines and the regional scale of the target in the image continues to decrease. First, it is necessary to ensure that the network can extract features at multiple scales, so we improve the spatial pyramid pooling (SPP). Second, a double network is used to aggregate multiscale features.

3.2.1. Improved SPP-Net. SPP-Net is a general CNN framework, which breaks the limitation that the input image must have a fixed size in traditional CNNs. In order to make the model adaptive and capable of handling images of different sizes, the SPP is introduced in this paper and placed after the backbone. SPP has the following significant features: (1) SPP generates fixed-size outputs regardless of the size of the input image, which is convenient for subsequent network processing. (2) Multilevel pooling makes SPP more adaptive to the change of the size. (3) Due to the flexibility of image input size, SPP is more effective in detecting and recognizing multiscale targets. The key point of SPP is that fixed-size feature vectors can be extracted from multiscale features. Therefore, SPP also shows great strength in target detection and recognition.

Compared with the previous YOLO, our proposed model adds an SPP module between convolutional layers. Figure 8. illustrates the structure of SPP block. Unlike the original SPP module, the SPP module in this paper consists of four parallel branches with kernel sizes of 5×5, 9×9, and 13×13 maximum pooling and a jump connection. The three different sizes of pooling are used to achieve the extraction of local features at different scales. The jump connection is used to preserve the original global features. Finally, the dimensionality of the tensor is expanded by the process of concating to achieve the fusion of local and global features, which enriches the expressiveness of the feature map and facilitates the case of various scales in water scene images. Compared with the way of using $k \times k$ maximum pooling alone, SPP module is more effective to increase the reception range of the main features and significantly separates the contextual features.

3.2.2. Double-Layer Network. Surface target recognition not only needs to local features with small receptive field to get the detail information but also needs to global features with large receptive field to get the global coarse-grained information, such as the shape and contour. As the CNN deepens, the network keeps downsampling. The semantic information becomes richer, and the spatial information is sparser. The last layer may even have a downsampling rate of 16 or 32. This result is that small targets on the original image have less effective information on the feature map. The performance of object recognition decreases sharply. In surface target recognition, small-scale targets have few pixels corresponding to them in the original image, and it is more difficult to find the corresponding information after downsampling.

Improved SPP-Net ensures that the network can receive information at multiscales; the key to the next strep lies in

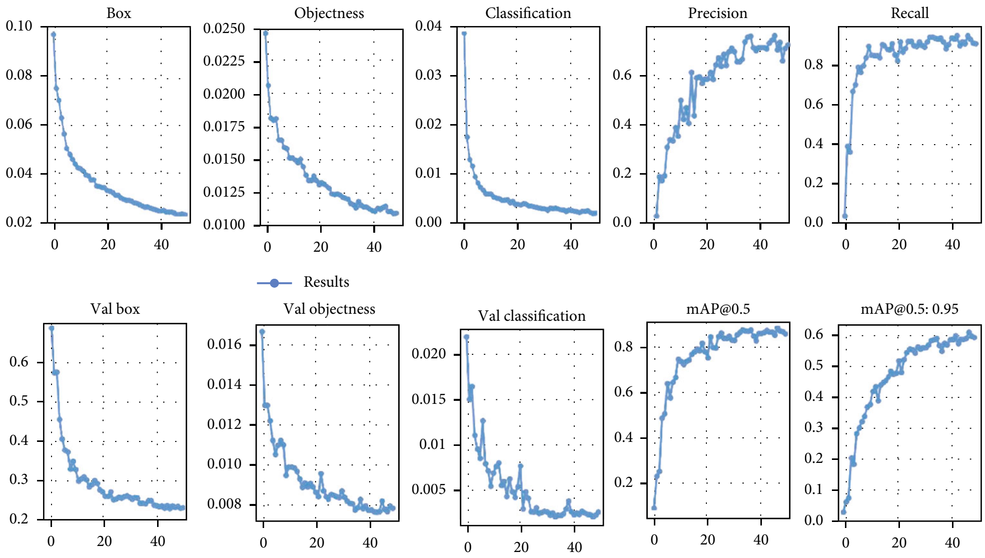

Figure 12: Graphs of the parameters during training.

Table 3: Comparison with other methods on COCO.

Model	Backbone	AP	$\mathrm{AP_S}$	$\mathrm{AP_M}$	$\mathrm{AP_L}$
Fast R-CNN	VGG-16	31.9	15.7	36.5	45.5
Faster R-CNN	ResNet-101	34.9	15.6	38.7	50.9
SSD300	VGG-16	24.4	6.6	25.9	41.4
DSSD513	ResNet-101	33.2	13.0	35.4	51.1
RetinaNet	ResNet-50	32.5	13.9	35.8	46.7
YOLOv3	DarkNet-53	33	18.3	35.4	41.9
YOLOv4	CSPDarkNet-53	43.5	26.7	46.7	53.3
SNIP	ResNet-101	44.4	27.3	47.4	56.9
SNIPER	ResNet-101	46.1	29.6	48.9	58.1
Our model	CSPDarkNet-53	47.5	32.9	48.3	57.5

how to extract multiscale features. Experiments show that neurons at higher levels respond strongly to the global features, while other neurons are more susceptible to local textures and contours. That means that networks at shallow levels are more related to detailed information and the networks at higher layers are more related to semantic information. The feature pyramid network (FPN) was created from this starting point, which has greatly promoted the subsequent work of object detection and recognition. FPN mainly consists of four operational processes: bottom-up path, top-down path, lateral connection, and convolutional fusion, through which models obtain strong semantic features. However, focusing only on the semantic features from deep levels is not sufficient in that this approach tends to ignore the detailed information contained in the shallow features. The introduction of path aggregation network (PAN) is aimed at enhancing the detail information in the shallow features from top to down.

Different from the direct use of FPN layers, our network adds two bottom-up feature pyramids (PAN) after the FPN layer, as shown in Figure 9. FPN layer passes high-level semantic features from the top to down. Although it enhances the whole pyramid, it only enhances the semantic information rather than the detail information. In this paper, we address this point by adding PAN, which conveys detailed localization features from the bottom to top. $\{P_1, P_2, P_3\}$ is used to represent the feature maps generated by FPN. $\{N_1, N_2, N_3\}$ is used to represent the newly generated high feature maps by the augmented paths and corresponding $\{P_1, P_2, P_3\}$. N_{i+1} is generated by a higher resolution N_i and laterally connected P_{i+1}. The spatial size of N_i is first reduced by 3×3 convolution. Then, P_{i+1} and downsampled feature maps are summed by lateral concatenation. The obtained feature maps repeat the above steps once again until the iterative step is terminated.

The purpose of this module is to transmit the semantic features from the deep layer to the shallow layer through FPN and the localization information from the shallow layer to the deep layer through PAN. The above two are combined to aggregate parameters at different scales to further improve the feature extraction capability of the model for multiscale targets.

4. Experiments

In this section, experiments are conducted on the COCO dataset, the VOC dataset, and our self-built surface target dataset. We describe the setup of our experiments and the way that the dataset is processed in Section 4.1. Then, a series of ablation experiments are performed to demonstrate the contribution of each proposed component to the performance of the model in Section 4.2. Finally, we compare our approach with state-of-the-art approaches on object detection and recognition.

4.1. Experimental Setup and Data Processing. The experiments in this paper are conducted on Ubantu 16.04 LTS operating system and Pytorch 1.2.0 deep learning framework. The model

Table 4: Accuracy for each category on VOC2012.

Model	Backbone	GPU	Aero	Bike	Boat	Bus	Car	mAP
Fast R-CNN	VGG-16	Yes	82.3	78.4	52.3	77.8	71.6	68.4
Faster R-CNN	ResNet-101	Yes	84.9	79.8	53.9	77.5	75.9	70.4
YOLO	DarkNet-53	Yes	77.0	67.2	38.3	68.9	55.9	57.9
SSD300	VGG-16	Yes	85.6	80.1	57.6	79.4	76.1	76.8
SSD513	ResNet-101	Yes	87.4	82.3	59.0	81.7	81.5	74.9
YOLOv3	DarkNet-53	Yes	81.7	85.7	65.5	87.2	87.5	74.5
Our model	CSP DarkNet-53	Yes	85.3	81.9	76.2	88.9	83.6	78.2

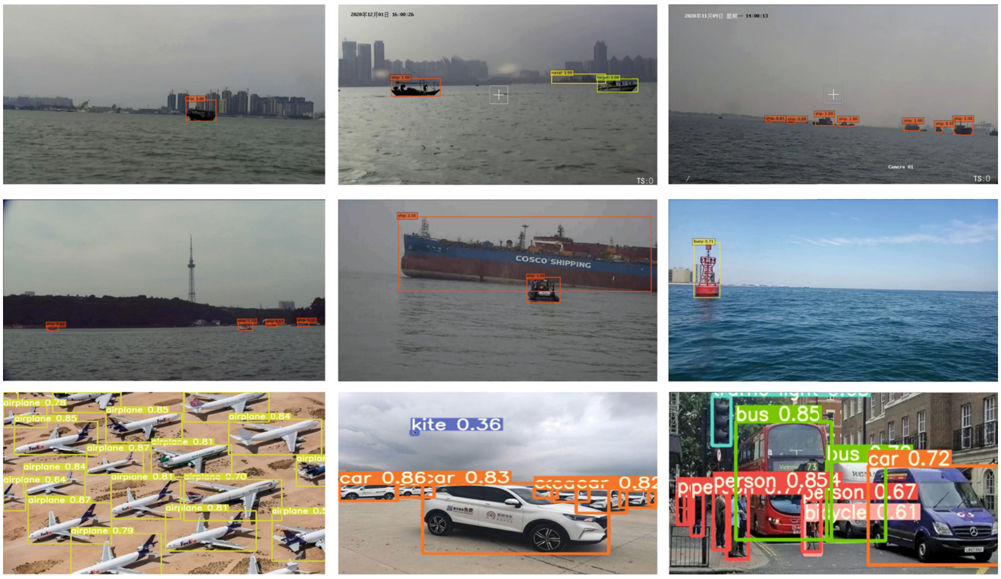

Figure 13: The test results on our self-built dataset and other public images.

is trained and debugged using GPUs with an NVIDIA RTX graphics card. On the public datasets, such as COCO and VOC, the input image size is uniformly cropped to 640 × 640 and the epoch is set to 300. A batch of 16 images are processed per iteration. The initial learning rate is 10^{-2}, adjusted to 10^{-3} for 100 epochs and 10^{-4} for 200 epochs until the end of training. The optimizer uses SGD optimizer.

Since the size of the bounding box is completely different from the COCO dataset, this paper uses the k-means algorithm to recalculate the anchors for optimization. The size of the input image is 1920 × 1080, and the epoch is set to 600. A batch of 8 images are processed into each iteration. The initial learning rate is 10^{-1}, adjusted to 10^{-2} when training 100 epochs, and adjusted to 10^{-3} until the end of training. The optimizer uses SGD optimizer.

In order to verify the recognition effect of the model on surface targets, a dataset is established by visible light sensor acquisition and manual annotation [5]. As can be seen in Figure 10, the dataset contains different scenes and a total of different classes of surface targets such as fishing boats, yachts, buildings, bridge piers, and water drums. First, the bounding boxes and categories of surface targets are labeled using a software called "Lableimg." Then, they are transformed into "txt" files like YOLO for processing, which include five types of information: the category of the target, x and y coordinates, and the width and height of the image.

The self-built surface target dataset consists of a total of 2229 images with 5731 labeled targets, and the size of the images is 1920 × 1080 pixels. Similar to the COCO, we divide targets into small, medium, and large, as seen in Table 1. The training results of neural network are influenced by the richness of the data. In order to enhance the learning ability of the model for various scale targets, a mosaic enhancement method is used in this paper. In order to enrich the sample at different scales and make the images closer to the real scenes, this paper adopts mosaic enhancement method in which mirroring, brightness enhancement, contrast enhancement, and linear blurring are utilized. The data enhancement method used in this paper is shown in Figure 11.

4.2. Ablation Study. To demonstrate the performance of the proposed coordinate attention and the double-layer cascade module, a series of ablation experiments are conducted on the surface target dataset. The corresponding results are all listed in Table 2. We compare the baseline with the model containing SE attention (SE), CBAM attention (CBAM), the horizontal attention (HA), the vertical attention (VA), coordinate attention (CA), the double-layer cascade module (DC), and the combination of multiscale module and coordinate attention (CA+DC). r is set to 24 when attention module is used. The average precision of small targets (AP_S), medium targets (AP_M), large targets (AP_L), mean average precision (mAP), and frame per second (FPS) is recorded. We determine whether the prediction is correct by whether the interaction ratio between the predicted box and the real box is bigger than 0.5.

As can be seen in Table 2, the model has significantly improved its recognition effect of small and large targets with the addition of coordinate attention and the double-layer cascade module. The indicator of mAP also performs best in this case. The results show that the addition of SE and CBAM affects the speed of the model and is not suitable for tasks in this paper. In comparison, the addition of coordinate attention and the double-layer cascade module can improve the accuracy of surface target recognition while ensuring the real time of the algorithm as much as possible. We also record the curves of the various parameters during the training process in Figure 12. The results show that the parameters are fitted quickly, and the model has a superior performance. The model basically converges when the epoch reaches 50.

4.3. Comparison with Other Methods. In this section, we evaluate the proposed model on the COCO dataset and the VOC dataset and compare it with other state-of-the-art methods. On the COCO dataset, we test AP, AP_S, AP_M, and AP_L, respectively, with the previous algorithms to verify the recognition performance on various scale targets. The FPS is not compared as the input is different. The results are shown in Table 3.

From the above results, it can be concluded that our model performs best on small-scale targets and only slightly behind Sniper on medium- and large-scale targets. We also show the test results of our model with other advanced methods on the VOC2012 dataset in Table 4, aiming to verify the recognition accuracy on specified targets. We select the target types like aero, boat, and bus. The reason that we select those targets is that they are similar to surface targets in appearance.

Our model achieves the highest recognition precision on three types of objects, aero, boat, bus, and performs worse on two types of objects, bike and car. The mean average precision is the highest among the listed methods. In addition, the model is tested on our self-built dataset and other public images. As can be seen in Figure 13, the classification and location of the target can be precisely returned regardless of the change in background and scale.

5. Summary and Prospect

To address the problems of diverse target types, easy confusion, large-scale span, and high requirements for the real time, this paper proposes a surface target recognition algorithm based on coordinate attention and double-layer cascade. The coordinate attention, which aggregates features along two spatial directions by feature encodings, retains the location information while learning spatial dependencies without additional network overhead. The double-layer cascade module aggregates parameters at different scales to further improve the feature extraction capability of the model for multiscale targets. Experimental results on the COCO dataset, the VOC dataset, and our self-built surface target dataset show that the proposed method is suitable for surface targets and has outstanding performance under evaluation metrics such as AP and FPS.

The surface target recognition algorithm is different from common target detection algorithm and recognition algorithm. It is more like a combination of both, which needs to give both the bounding box of the target and accurately identify the classification of the target. In addition, surface target recognition is usually applied to mobile platforms such as USVs, which puts higher requirements on the real-time nature of the algorithm. This paper still stays in the recognition of images in fixed scenes. In the future, more attention should be paid to the continuous tracking of targets in moving scenes and the collaborative perception of multiplatform and multiperspective under weak observation conditions.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

We declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Natural Science Foundation of Hunan Province Youth Fund (Project No. 2020JJ5672).

References

[1] J. He, Y. Guo, and H. Yuan, “Ship target automatic detection based on hypercomplex flourier transform saliency model in high spatial resolution remote-sensing images,” *Sensors*, vol. 20, no. 9, p. 2536, 2020.

[2] E. T. Steimle and M. L. Hall, “Unmanned surface vehicles as environmental monitoring and assessment tools,” in *Oceans*, IEEE, 2006.

[3] M. Kristan, V. S. Kenk, S. Kovacic, and J. Perš, “Fast image-based obstacle detection from unmanned surface vehicles,” *IEEE Transactions on Cybernetics*, vol. 46, no. 3, pp. 641–654, 2016.

[4] Y. Zhang, H. Wang, and X. Fang, “Object detection and recognition of intelligent service robot based on deep learning,” in *2017 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)*, pp. 171–176, Ningbo, China, 2018.

[5] R. Guo, S. Su, Z. Zuo, and B. Sun, “A multi-scale surface target recognition algorithm based on attention fusion mechanism,” *2021 IEEE International Conference on Computer Science,*

Artificial Intelligence and Electronic Engineering, 2021, SC, USA, 2021.

[6] W. Zhang, F. Jiang, C. F. Yang, Z. P. Wang, and T. J. Zhao, "Research on unmanned surface vehicles environment perception based on the fusion of vision and lidar," *IEEE Access*, vol. 9, pp. 63107–63121, 2021.

[7] T. R. Gadekallu, D. S. Rajput, M. Reddy et al., "A novel PCA–whale optimization-based deep neural network model for classification of tomato plant diseases using GPU," *Journal of Real-Time Image Processing*, vol. 18, no. 4, pp. 1383–1396, 2021.

[8] M. Sun, Y. Yuan, F. Zhou, and E. Ding, *Multi-Attention Multi-Class Constraint for Fine-Grained Image Recognition*, Springer, Cham, 2018.

[9] T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2117–2125, Hawaii, America, 2017.

[10] G. Wang, S. Xie, and C. Xiumin, "Image recognition method of ships in front of unmanned surface vessel based on deep learning," *Ship Engineering*, vol. 40, pp. 19–22, 2018.

[11] R. Girshick, J. Donahue, and T. Darrell, "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, Columbus, America., 2014.

[12] P. Purkait, C. Zhao, and C. Zach, "SPP-Net: deep absolute pose regression with synthetic views," (2017), https://arxiv.org/abs/1712.03452.

[13] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, Santiago, Chile., 2015.

[14] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: object detection via region-based fully convolutional networks," in *Advances in neural information processing systems*, Curran Associates Inc, 2016.

[15] W. Liu, D. Anguelov, D. Erhan et al., *SSD: Single Shot MultiBox Detector [C]//European Conference on Computer Vision*, Springer International Publishing, 2016.

[16] J. Redmon, S. Divvala, and R. Girshick, "You only look once: unified, Real-Time Object Detection," *Computer Vision & Pattern Recognition*, pp. 779–788, 2016.

[17] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *IEEE Conference on Computer Vision & Pattern Recognition*, pp. 6517–6525, Hawaii, America., 2017.

[18] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement," 2018, https://arxiv.org/abs/1804.02767.

[19] H. Jie, S. Li, and S. Gang, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, 2017.

[20] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, *CBAM: Convolutional Block Attention Module*, Springer, Cham, 2018.

[21] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7794–7803, Salt Lake City, USA., 2018.

[22] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: non-local networks meet squeeze-excitation networks and beyond," *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 1971–1980, 2019.

[23] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "A^2$-Nets: double attention networks," *Advances in Neural Information Processing Systems*, vol. 11301, 2018.

[24] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: exploiting feature context in convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[25] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV*, pp. 3286–3295, Seoul, Korea, 2020.

[26] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: convolutional triplet attention module," 2020, http://arxiv.org/abs/2010.03045.

[27] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection-SNIP," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[28] B. Singh, M. Najibi, and L. S. Davis, "SNIPER: efficient multi-scale training," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[29] Z. Cai, Q. Fan, R. S. Fe Ris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9908 of Lecture Notes in Computer Science, Springer, Cham, 2016.

[30] A. Shrivastava, R. Sukthankar, J. Malik, and A. Gupta, "Beyond skip connections: top-down modulation for object detection," 2016, https://arxiv.org/abs/1612.06851.

[31] S. Liu, L. Qi, and H. Qin, "Path aggregation network for instance segmentation," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[32] R. G. Thippa, K. Praveen, and K. Lakshmanna, "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 99, pp. 1–1, 2020.

[33] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.