DLAHSD: DYNAMIC LABEL ADOPTED IN AUXILIARY HEAD FOR SAR DETECTION

Xiaoxiao Yin[†], *Shiyong Lan*[◇†*], *Weikang Huang*[†], *Yitong Ma*[◇], *Wenwu Wang*[‡], *Hongyu Yang*[◇†] *Yilin Zheng*[◇]

[◇]College of Computer Science, Sichuan University, China.
[†] National Key Laboratory of Fundamental Science on Synthetic Vision, China.
[‡]Center for Vision Speech and Signal Processing, University of Surrey, UK.

ABSTRACT

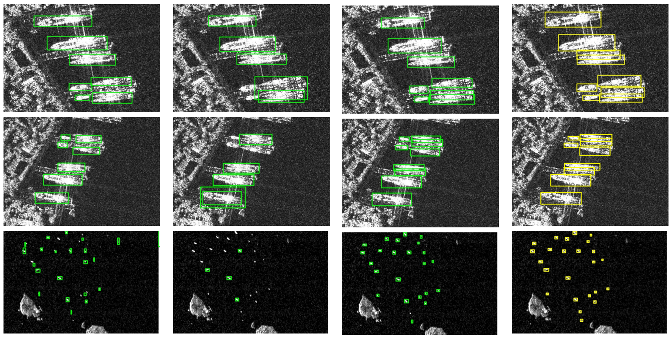
Ship detection in synthetic aperture radar (SAR) images is a major issue in maritime surveillance and port management. Existing challenges are mainly as follows: (1) Tiny ships are mixed with scattered noise spots on the sea. (2) Ships are present in extreme aspect-ratios and various scales. (3) The land background blurs the outline of coastal ships. To address these problems, we propose an efficient detection neural network (DLAHSD) that integrates the Multi-scale Feature Location Fusion (MFLF) module and the Auxiliary Detection Head (ADH) based CenterNet. In addition, we designed a Dynamic Elliptic Gaussian (DEG) module to label the heatmap of ships. Experimental results on the challenging SSDD dataset show that our model offers improved performance over the baseline methods. The codes will be available at https://github.com/SYLan2019/DLAHSD.

***Index Terms*—** Ship detection, Multi-scale Feature Location Fusion, Auxiliary Detection Head, Dynamic Ellipse Gaussian, SSDD dataset

1. INTRODUCTION

Synthetic Aperture Radar (SAR) is an active imaging device that can operate day and night and is not affected by weather conditions [1, 2]. Because of its ability to provide independent and all-weather monitoring of solar illumination, spaceborne SAR systems like Sentinel-1 [3] and TerraSAR-X [4] have produced a substantial number of high-resolution SAR images. These images are particularly useful for ship detection in marine traffic monitoring and port management due to the unique characteristics of the SAR images.

The objective of ship detection in SAR images is to locate ships in the scene. Despite significant efforts in this field, the issue has not been fully resolved due to the subtle SAR imaging mechanism. In SAR images, ships can be difficult to distinguish, as they may be mixed with the background. Many ship detection algorithms have been proposed, among which constant false alarm rate (CFAR) and its variations are widely adopted [5, 6]. These algorithms rely on manually designed geometric and texture features, which can be time-consuming to obtain, and tend to produce inaccurate predictions in complex scenarios.



(a) CenterNet (b) Atss (c) Ours (d) GT

Fig. 1. Ships are densely parked in parallel at the port, mixed with the boundary of the land background, which are rather difficult to be detected. In addition, the scattered noise on the sea surface is very similar to that of small ships, which inevitably leads to false detection. Our method addresses these challenges and achieves better results than the baselines.

In recent years, CNN-based deep learning techniques have been used for ship detection, where features are learned, thus enhancing the representation robustness of human-engineered features used in traditional algorithms. Furthermore, attention mechanisms are often used to enhance the learned features. For example, attention panels are connected with a feature pyramid network [7]. SAGAN [8] used an attention module in skip connection to capture additional local information. However, these attention mechanisms can not adequately handle the information redundancy between feature channels when they are used for ship detection from SAR images. As for detectors in the SAR images, the CenterNet [9], as a representative anchor-free detector, is superior in terms of computational efficiency and detection accuracy. TTFNet [10] encoded training samples by using elliptic Gaussian kernels for better accuracy and training speed. OSD-SSD [11] proposed some pretraining techniques to transfer the characteristics of ships in earth observations to

[*] Corresponding author. E-mail: lanshiyong@scu.edu.cn.

This work was funded by 2035 Innovation Pilot Program of Sichuan University, China.

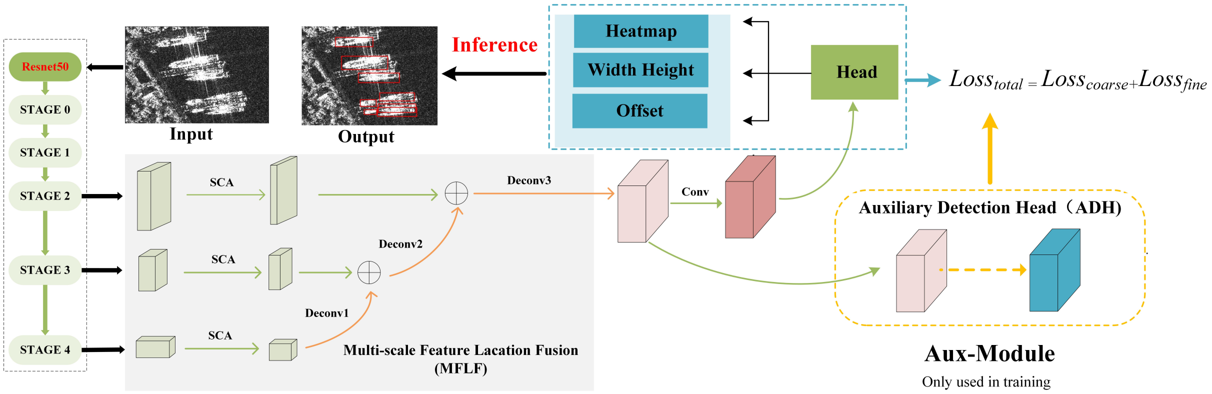


Fig. 2. Overview of the DLAHSD structure. DLAHSD has three major components. The MFLF enhances the region of interest first, then performs a multi-scale fusion. The ADH decodes the coarse feature map, and feeds the final refined feature map to the detection head. The DEG component is a dynamic elliptic Gaussian algorithm for heatmap labeling.

SAR images. EESD [12] combined grid-refined anchor boxes and multi-scale feature fusion to improve ship detection in SAR images with complex scenes. These methods have contributed greatly to SAR image detection, but still face many challenges, such as the confusion between scattered noise and small ships, or the tight adhesion between ships (as shown in Fig. 1).

To address the aforementioned challenges, we propose an efficient anchor-free network called DLAHSD. First, we introduce a novel Multi-scale Feature Location Fusion (MFLF) module that highlights the position of ships and fuses information from feature maps of different sizes to enable the detection of ships of various scales. Second, we design an Auxiliary Detection Head (ADH) that incorporates an additional coarse loss, further refining the final feature map and improving the overall performance. Finally, we propose the use of a Dynamic Elliptic Gaussian (DEG) kernel to label the heatmap, reducing the overlap between ships. We conduct ablation experiments on the challenging SSDD dataset [13] to demonstrate the advantages of DLAHSD over existing methods.

2. METHODOLOGY

Our network, DLAHSD, is comprised of three main components, as illustrated in Figure 2. The first component is the Multi-scale Feature Location Fusion (MFLF) module, which improves the region of interest for each stage feature map and performs coarse-grained fusion of the feature maps. The second component is the Auxiliary Detection Head (ADH), which decodes the shallow feature map and introduces an additional loss to further refine the final feature map for better performance. Lastly, the Dynamic Elliptical Gaussian (DEG) is utilized as a heatmap labeling algorithm to better fit the distribution of ships and reduce interference among ships.

2.1. Multi-scale Feature Location Fusion

The Multi-scale Feature Location Fusion (MFLF) module is proposed to address the limitations of ship detection at various scales. In CenterNet [9], only the feature map from the last stage of the backbone network is used, which may lack the representation of multi-scale targets features. To enhance the spatial region of interest and fuse the channel of feature maps, we designed the Squeeze Coordinate Attention (SCA) module, inspired by Coordinate Attention (CA) [14] but with some modifications to address its limitations in our module. On the basics of Coordinate Attention [14], we added a 3×3 convolution to squeeze channels of original feature map. Then multiply it with the channel attention mapping map of the two spatially directions of original feature map, to preserve certain essential information between channels. Figure 3 shows the implementation details of the SCA sub-module.

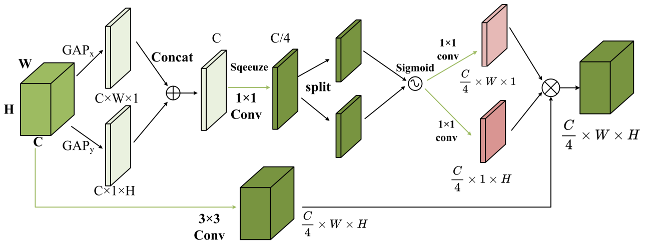

W
H
C
GAP_x
GAP_y
$C{\times}W{\times}1$
$C{\times}1{\times}H$
Concat
C
C/4
Sqeeuze
1×1
Conv
split
Sigmoid
1×1
conv
1×1
conv
$\frac{C}{4} \times W \times 1$
$\frac{C}{4} \times 1 \times H$
$\frac{C}{4} \times W \times H$
3×3
Conv
$\frac{C}{4} \times W \times H$

Fig. 3. Squeeze Coordinate Attention (SCA) sub-module. This mechanism enhances the presentation of the region of interest (ROI), via squeezing and fusing the information from the input channel.

We select feature maps C_1, C_2, C_3 from the last three stages of ResNet50 and apply the SCA operation to obtain K_1, K_2, K_3. We then upsample and merge them from the bottom to the top. The upsample process uses the Deconv module, which consists of a $3{\times}3$ deformable convolution [15] and a 4×4 transposed convolution. This approach enables the network to better focus on ships at different scales. The details of the MFLF are as follows:

$$\begin{aligned}&K_i = SCA(C_i) \quad (i = 1, 2, 3)\\&P1 = Deconv1(K_3) + K_2, P2 = Deconv2(P1) + K_1\\&Output = Deconv3(P2)\\&Deconv(X) = F'(Upsample(X))\end{aligned} \tag{1}$$

where $F'(\cdot)$ is a 3×3 deformable convolution. $Upsamle(\cdot)$ is a 4×4 transposed convolution. $P1$ and $P2$ is the intermediate feature map. $SCA(\cdot)$ is formulated as follows:

$$Z_c^h(h) = \frac{1}{W}\sum_{i=0}^{W} X_c(i,h), Z_c^w(w) = \frac{1}{H}\sum_{j=0}^{H} X_c(w,j) \tag{2}$$

$$f = \delta(F_1[Z_c^h, Z_c^w]), g^h = \sigma(F_h(f^h)), g^w = \sigma(F_w(f^w)) \tag{3}$$

$$\tilde{X} = BN(Conv_{3\times3}(X)) \tag{4}$$

$$\mathrm{Y_c(i,j) = \tilde{X}_c(i,j) \times g_c^h(i) \times g_c^w(j)} \tag{5}$$

where $X \in R^{C\times W\times H}$ is the input feature map. X_c is the spatial feature of the c_{th} channel. F_1 is a 1×1 convolution for channel reduction. δ is the non-linear activation function. σ is the Sigmoid function. F_h, F_w are also 1×1 convolution used to maintain current channel numbers. $Conv_{3\times3}(\cdot)$ reduces the number of channels to $\frac{C}{4}$.

2.2. Auxiliary Detection Head

The texture details in SAR images are often poor, making it difficult to distinguish the boundaries of ships, especially when they are closely aligned or berthed near the port and mixed in with the land. To address this issue, we drew inspiration from the auxiliary head introduced in YOLOv7 [16]. As shown in Figure 4, we propose an Auxiliary Detection Head (ADH) to carry out rough detection in advance and introduce an additional coarse loss (i.e., $Loss_{coarse}$). The coarse loss denotes the difference between the output of ADH and label, which can be used to guide the shallow network weights (i.e.m F_1) learning. Different from the YOLOv7, our auxiliary head is not used in the middle layers of the network, but the penultimate layer. Furthermore, our ADH does not need extra auxiliary label that is used in YOLOv7. So, we can get the total loss as follows:

$$Loss_{total} = Loss_{corase} + \lambda \cdot Loss_{fine} \tag{6}$$

$$\begin{aligned}Loss = \sum_{i=0}^{W}\sum_{j=0}^{H} GL(Y_{i,j}, \hat{Y_{i,j}}) + \frac{\lambda_1}{N_{pos}}\sum_{i=0}^{N_{pos}}\\ L_1(L_{wh}, \hat{L_{wh}}) + \frac{\lambda_2}{N_{pos}}\sum_{i=0}^{N_{pos}} L_1(s, \hat{s})\end{aligned} \tag{7}$$

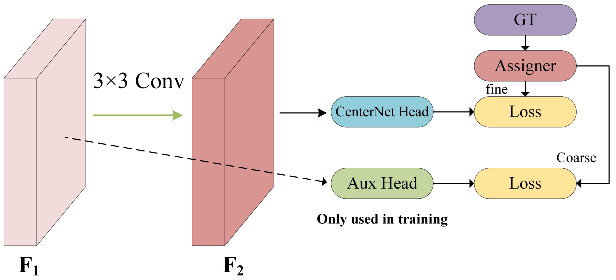


Fig. 4. The Auxiliary Detection Head module. The feature map F$_1$ on the left is used to coarsely calculate the loss $Loss_{coarse}$, while the feature map F$_2$ on the right is used to accurately calculate the loss $Loss_{fine}$.

$$GF(x, \hat{x}) = \begin{cases} -\lg \hat{x} \cdot (1-\hat{x})^{\alpha}, & if \quad x = 1\\ -\lg (1-\hat{x}) \cdot \hat{x}^{\alpha} \cdot (1-x)^{\gamma}, & otherwise\end{cases} \tag{8}$$

where we set λ to 0.3. the $Loss_{fine}$ is obtained from the CenterNet Head. $Loss_{fine}$ and $Loss_{coarse}$ are defined the same as $Loss$. GL is the Gaussian focal loss [17]. The $\hat{Y}$ is the predicted heatmap. The $\hat{L_{wh}}$ is the predicted width and height. $\hat{s}$ is the predicted offset. N_{pos} is the numbers of positive samples. λ_1 and λ_2 are set to 0.1. α is set to 2. γ is set to 4.

2.3. A novel heatmap labelling method

Ship detection is particularly challenging for ships with extreme aspect ratios. The CenterNet uses a Circle Gaussian to label the heatmap, but when ships are densely arranged, there is interference between each Gaussian distribution, resulting in multiple positive samples for a single ship, which can significantly impact detection accuracy. To address this issue, we propose the Dynamic Elliptic Gaussian (DEG) method, which takes into account the area and aspect ratio of each ground truth (GT) box. We design a Gaussian kernel dynamically based on each GT box's characteristics. Specifically, for each GT box with its center point $p \in (W, H)$, we generate a heatmap label $Y \in [0,1]^{1\times\frac{W}{4}\frac{H}{4}}$ and a Gaussian kernel $Y_{(x,y)}$ to calibrate the heatmap. The kernel centroid $\tilde{p} = \lfloor p/4 \rfloor$ is located in the heatmap label Y, and we use the Gaussian kernel to complete the calibration. Figure 5 below shows a comparison of DEG with other Gaussian kernels. The DEG details are as follows:

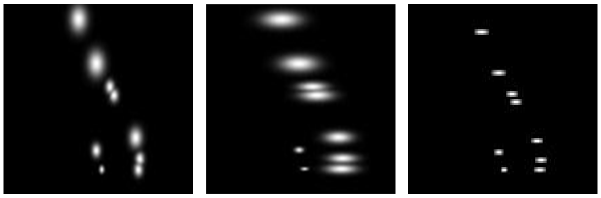
$$ratio_{wh}^{i} = \frac{max(W_i, H_i)}{min(W_i, H_i)} \tag{9}$$

$$\sigma_x = \alpha \cdot \frac{W_i}{\lg W_i H_i \sqrt[\beta]{ratio_{wh}^{i}}} \tag{10}$$

$$\sigma_y = \alpha \cdot \frac{H_i}{\lg W_i H_i \sqrt[\beta]{ratio_{wh}^{i}}} \tag{11}$$

$$Y_{(x,y)} = exp[-\frac{(x-\tilde{p}_x)}{2\sigma_x^2} - \frac{(y-\tilde{p}_y)}{2\sigma_y^2}] \tag{12}$$

where σ_x, σ_y is the dynamic standard deviation in both x and y dimensions for each GT box. $ratio_{wh}^i$ is the aspect-ratio of the i_{th} GT box. W_i, H_i are the width and height of the i_{th} GT box. α, β are hyperparameters.



(a) Circle Gaussian (b) Elliptic Gaussian (c) proposed DEG

Fig. 5. Comparasion with different heatmap labeling strategies. The Gaussian distribution of ships cannot be too large or too small. Therefore DEG took into account the area and aspect ratio of ship to design the Gaussian kernel dynamically. The Gaussian distribution is more fused. Obviously the interference between ship Gaussian distributions is weakened, which benefits for improving model robustness

3. EXPERIMENTS

3.1. Dataset and metrics

We implemented our approach using PyTorch and the MMdetection toolkit. Our approach was trained and evaluated on the SSDD dataset [13], which contains 1160 images and 2456 ships, with images of varying sizes. Prior to training, we preprocessed the image sizes to 512×512. The dataset was randomly divided into a training set, a validation set, and a test set in an 8:1:1 ratio.

3.2. Experimental Results and Analysis

Considering the balance between speed and precision, we have selected several single-stage baselines to compare with DLAHSD. These baselines include SSD512 [18], CenterNet [9], FCOS [19], YOLOv3 [20], ATSS [21], VFNet [22], YOLOX [23], OSD-SSD [11] and EESD [12]. The main metrics we used to compare these methods are AP, parameters, and FPS, which are shown in Table 1.

It is evident that our DLAHSD achieves the highest AP. In Table 2, we perform an ablation experiment on the SSDD dataset. With the addition of the MFLF module, our detector can now more accurately identify and focus on ships of different scales, further improving its capabilities. So, while the $AP@0.5$ has improved significantly, there hasn't been much improvement in the $AP@0.75$ metric. Next, we tried feeding the coarse feature map to ADH, which introduced additional coarse loss during training. So the ADH heightened the performance. Building upon these improvements, we utilized DEG for heatmap labelling. The results obtain a significant improvement in the $AP@0.75$ metric. Figure 6 shows the AP line chart for different α and β values selected by DEG in DLAHSD.

Table 1. Experimental results of DLAHSD and baselines

Method	backbone	AP(%)	AP@0.5(%)	AP@0.75(%)	FPS	Parameters(M)
SSD	VGG16	64.3	94.3	75.4	43.8	24.39
Yolov3	DarkNet-53	42.5	89.8	30.3	55.1	61.52
CenterNet	ResNet50	52.2	89.6	56.2	**78.1**	30.68
Fcos	ResNet50	54.5	88.2	63.0	47.3	31.84
VfNet	ResNet50	36.3	65.3	37.0	37.4	34.28
Atss	ResNet50	60.2	92.1	69.0	44.4	31.89
YoloX	CSPDarknet	64.5	95.2	**80.3**	71.5	**8.94**
OSD-SSD	VGG16	-	96.1	-	-	-
EESD	DarkNet-53	-	95.5	-	-	-
DLAHSD	ResNet50	**64.8**	**96.3**	79.1	52.3	39.41

Table 2. Ablation experiments on the SSDD dataset.

MFLF	ADH	DEG	AP(%)	AP@0.5(%)	AP@0.75(%)	FPS	Parameters(M)
✘	✘	✘	53.1	88.2	55.9	**78.1**	**30.68**
✔	✘	✘	55.1	95.2	58.7	53.7	38.32
✔	✔	✘	60.2	96.1	61.3	53.7	39.41
✔	✔	✔	**64.8**	**96.3**	**79.1**	52.3	39.41

MFLF, ADH and DEG modules all have performance improvements for the network. Figure 6 shows AP line chart for different α and β selected by DEG in DLAHSD.

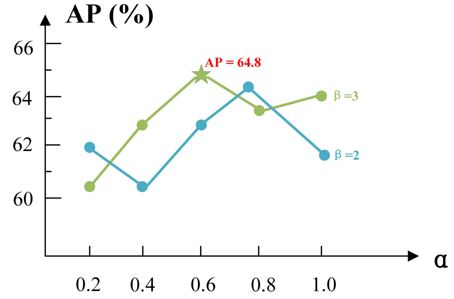


Fig. 6. The AP line chart of different α and β. We get the best AP when α=0.6 and β=3.

4. CONCLUSION

We have presented a novel and effective detector called DLAHSD. In the DLAHSD, the MFLF module can effectively focus on ships of various sizes under complex background, by accurately capturing the multi-scale dependence in the image. The ADH module calculates coarse losses from shallow features, which can refine network parameters learning by optimizing the total loss. The DEG module takes into account the area and aspect-ratio of the ship, and dynamically generates Gaussian kernel according to the characteristics of each ground truth box, thus obtaining it's accurate heatmap label. The experimental results of the DLAHSD on the SSDD dataset achieve a competitive performance compared to baselines.

5. REFERENCES

[1] A. Moreira, P. Prats-Iraola, M. Younis, G. Krieger, I. Hajnsek, and K. P. Papathanassiou, "A tutorial on synthetic aperture radar," *IEEE Geoscience and remote sensing magazine*, vol. 1, no. 1, pp. 6–43, 2013.

[2] X. Leng, K. Ji, S. Zhou, X. Xing, and H. Zou, "An adaptive ship detection scheme for spaceborne sar imagery," *Sensors*, vol. 16, no. 9, p. 1345, 2016.

[3] M. Stasolla and H. Greidanus, "The exploitation of sentinel-1 images for vessel size estimation," *Remote Sensing Letters*, vol. 7, no. 12, pp. 1219–1228, 2016.

[4] S. Brusch, S. Lehner, T. Fritz, M. Soccorsi, A. Soloviev, and B. van Schie, "Ship surveillance with terrasar-x," *IEEE transactions on geoscience and remote sensing*, vol. 49, no. 3, pp. 1092–1103, 2010.

[5] C. Wang, F. Bi, W. Zhang, and L. Chen, "An intensity-space domain cfar method for ship detection in hr sar images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 4, pp. 529–533, 2017.

[6] S.-I. Hwang and K. Ouchi, "On a novel approach using mlcc and cfar for the improvement of ship detection by synthetic aperture radar," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 2, pp. 391–395, 2010.

[7] Z. Cui, Q. Li, Z. Cao, and N. Liu, "Dense attention pyramid networks for multi-scale ship detection in sar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 8983–8997, 2019.

[8] G. Liu, S. Lan, T. Zhang, W. Huang, and W. Wang, "Sagan: skip-attention gan for anomaly detection," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 2468–2472.

[9] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[10] Z. Liu, T. Zheng, G. Xu, Z. Yang, H. Liu, and D. Cai, "Training-time-friendly network for real-time object detection," in *proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 11 685–11 692.

[11] W. Bao, M. Huang, Y. Zhang, Y. Xu, X. Liu, and X. Xiang, "Boosting ship detection in sar images with complementary pretraining techniques," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 8941–8954, 2021.

[12] Y. Chen, T. Duan, C. Wang, Y. Zhang, and M. Huang, "End-to-end ship detection in sar images for complex scenes based on deep cnns," *Journal of Sensors*, vol. 2021, pp. 1–19, 2021.

[13] T. Zhang, X. Zhang, J. Li, X. Xu, B. Wang, X. Zhan, Y. Xu, X. Ke, T. Zeng, H. Su *et al.*, "Sar ship detection dataset (ssdd): Official release and comprehensive data analysis," *Remote Sensing*, vol. 13, no. 18, p. 3690, 2021.

[14] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 713–13 722.

[15] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[16] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv preprint arXiv:2207.02696*, 2022.

[17] J. Wang, F. Li, and H. Bi, "Gaussian focal loss: Learning distribution polarized angle prediction for rotated object detection in aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[19] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.

[20] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[21] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9759–9768.

[22] H. Zhang, Y. Wang, F. Dayoub, and N. Sunderhauf, "Varifocalnet: An iou-aware dense object detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8514–8523.

[23] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.