Cross-Level Semantic Segmentation Guided Feature Space Decoupling And Augmentation for Fine-Grained Ship Detection

Zhengning Zhang
Tsinghua University and Space Star Technology Co.Ltd.,
Beijing, China
23880666@qq.com

Lin Zhang
Tsinghua Shenzhen International Graduate School
Shenzhen, China
linzhang@sz.tsinghua.edu.cn

Yue Wang
Department of Electronic Engineering Tsinghua University
Beijing, China
wangyue@tsinghua.edu.cn

Pengming Feng
State Key Laboratory of Space-Ground Integrated Information Technology
Beijing, China
p.feng.cn@outlook.com

Shaobo Liu
Space Star Technology Co.Ltd.,
Beijing, China
liu_shaobo@163.com

Jian Wang
Space Star Technology Co.Ltd.,
Beijing, China
wangjian1@spacestar.com.cn

***Abstract*—Fine-grained ship detection in optical remote sensing images is a challenging problem due to its long-tailed distributed dataset, which is often coupled with the multi-scale of ship and complex environment. In this paper, a novel average instance area imbalance ratio (AIAIR) is firstly used for quantitatively evaluating long-tailed distribution and multi-scale coupled problem. Based on which, we propose the idea of feature space decoupling and augmentation guided by cross-Level semantic segmentation, where features on different classwise-balance level are scheduled. On this basis, a Siamese Semantic Segmentation Guided Ship Detection Network (SGSDet) is proposed to effectively facilitate fine-grained ship detection performance. Our proposed method can be easily plugged into existing object detection models. Numerical experiments show that the proposed method outperforms the baseline by 2.32% mAP on the ShipRSImageNet dataset without extra annotations.**

***Index Terms*—Ship detection, Remote sensing, Object detection, Long-tailed distribution problem**

I. INTRODUCTION

Ship detection in optical satellite images has been widely applied in military and civilian fields, such as maritime situation assessment, monitoring of important ports and targets, maritime rescue, illegal fishing, etc. Although deep learning-based ship detection methods have achieved the state-of-the-art performance in optical remote sensing images, fine-grained ship detection has always been a challenging problem, which has been aroused extensive attention in recent years. One of the most essential problem in fine-grained detection is its long-tailed class distribution of samples, where the imbalance distributed of classes makes the training of deep learning based models moves to over-fitting.

Many studies have been conducted in recent years to address this problem, Yifan Zhang et al. group these methods into three categories, e.g., class re-balancing (SimCal [1], focal loss [2]), information augmentation and module improvement [3]. Besides, transfer Learning, as the mainstream paradigm approach of information augmentation methods, has also been employed to address the long-tailed problem, where knowledge learned from the head classes are transferred to underrepresented tail classes [3], such as Feature Cloud [4]. Xue Yang et al. proposed a instance-level feature denoising method to enhance the detection accuracy of small and cluttered objects [5].

Imbalance Ratio (IR) is the most popular measure of the imbalanced data numbers across classes, as formulated in (3). For classification and object detection tasks in general image, the larger the IR, the larger the imbalance extent, which leads to the lower of detection accuracy.

$$IR = \frac{N_{majority}}{N_{minority}} \tag{1}$$

where $N_{majority}$ is the sample size of the largest majority class, and $N_{minority}$ is the sample size of the smallest minority class.

However, we observe from detection results on many object detection tasks in remote sensing that performance of tail-class is often better than head-class. Compared with general images, there are two main difficulties in fine-grained ship detection in remote sensing images. Firstly, the long-tailed distribution problem often coupled with multi-scale of target, specifically, targets with larger size can often achieve better performance even with fewer training samples. Secondly, imaging environment in remote sensing images is complex, which causes the various features of targets, and leads to low classification accuracy. For above reasons, the class re-balancing methods are not suitable for long-tailed distributions and multi-scale coupled problem in fine-grained ship detection task in remote sensing image.

In this paper, we present Average Instance Area Imbalance Ratio (AIAIR) for quantitatively evaluating long-tailed distributions and multi-scale coupling problems. Furthermore, the AIAIR is proposed to schedule the classification level, where semantic segmentation information is utilized for decoupling and augmentation of the assigned feature space. In general, a Siamese Semantic Segmentation Guided Ship Detection Network (SGSDet) is designed for fine-grained ship detection. Finally, we extensively evaluate our SGSDet method on the ShipRSImageNet datasets [6], and demonstrate that it can significantly increase the ship detection accuracy.

II. Proposed Method

A. Pipeline

Inspired by the idea proposed by Peng Chu et al. [7], the feature space of each ship class could be decoupled to class-generic, class-specific feature, and background feature. However, if the assigned classification level is extremely unbalanced and coupled with the multi-scale problem, it is difficult to decouple the feature space directly. Therefore, in this work, the features are decoupled according to a more balanced level, where the class-generic features of each class are extracted firstly, then they are transferred to the assigned level, as shown in Fig.1.

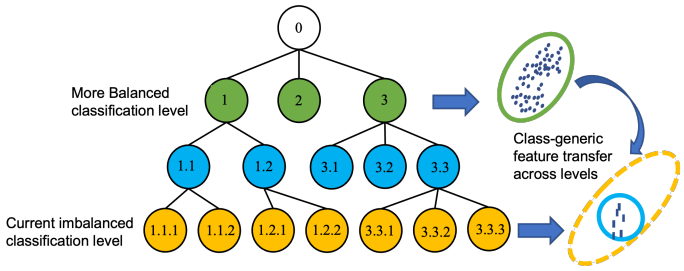

Fig. 1. A hierarchically object detection dataset and class-generic feature transfer.

The pipeline of the proposed method consists of seven steps.

1) Calculating and schedule the AIAIR for each classification levels.
2) Choosing a more balanced classification level for feature space decoupling.
3) Mapping the labels of the assigned classification level to a more balanced level.
4) Training a Siamese Semantic Segmentation Network with mapped labels and annotations supervision.
5) Generating Decoupling Matrix with features from the Siamese Semantic Segmentation Network.
6) Decoupling and augmentation in features space.
7) Original features and augmented feature fusion.

B. Average Instance Area Imbalance Ratio

Firstly, We define the average instance area (AIA) to represent the multi-scale of ship instances as:

$$AIA_{C_j} = Average(Area^i_{C_j}) \tag{2}$$

where AIA_{C_j} is average instance area of the C_j class, and $Area^i_{C_j}$ is the area of the C_j with class instances index i.

Secondly, based on IR, AIAIR is calculated to quantitatively describe the imbalance extent of long-tailed distribution coupled with multi-scale problem of dataset:

$$AIAIR = \frac{AIA_{max}}{AIA_{min}} \tag{3}$$

where AIA_{max} and AIA_{min} are the largest and smallest AIA of the dataset, respectively.

C. Feature Space Decoupling and Augmentation Guided by Cross-Level Semantic Segmentation

After calculating AIAIR of the dataset, features across different class level are decoupled and argumented by following processes, as shown in Fig.2.

- Decoupling Phase: Decoupling the features space according to a lower but more balanced level, obtaining a decoupling matrix as $D = G \cup S \cup B$, where D denotes the decoupling matrix, G denotes the class-generic features, S denotes the class-specific features, and B denotes the background feature.
- Augmentation Phase: Merging the decoupling matrix with assigned feature to achieve feature transfer and augment class-generic features by Hadamard product, as $\hat{F} = F \odot D = F \odot (G \cup S \cup B)$, where F denotes the features in assigned level, and $\hat{F}$ denotes the features after augmentation.
- Fusion Phase: The assigned feature map and the augmented feature map are fused to complement class-specific features as $\hat{F}' = \hat{F} \oplus F$, where $\hat{F}'$ denotes the fused features

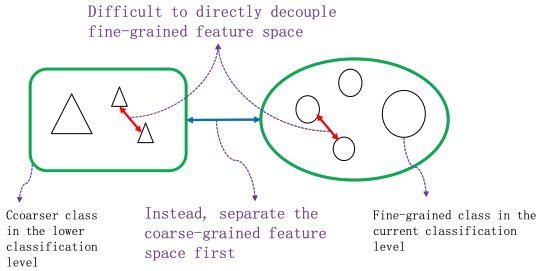

Fig. 2. Cross level feature space decoupling.

Inspired by Yang [5], a Siamese semantic segmentation network is used in this work to obtain a more approximate feature decoupling matrix, which is trained using cross-level approximate semantic segmentation annotated by rotated horizontal boxes. The overall feature space decoupling and

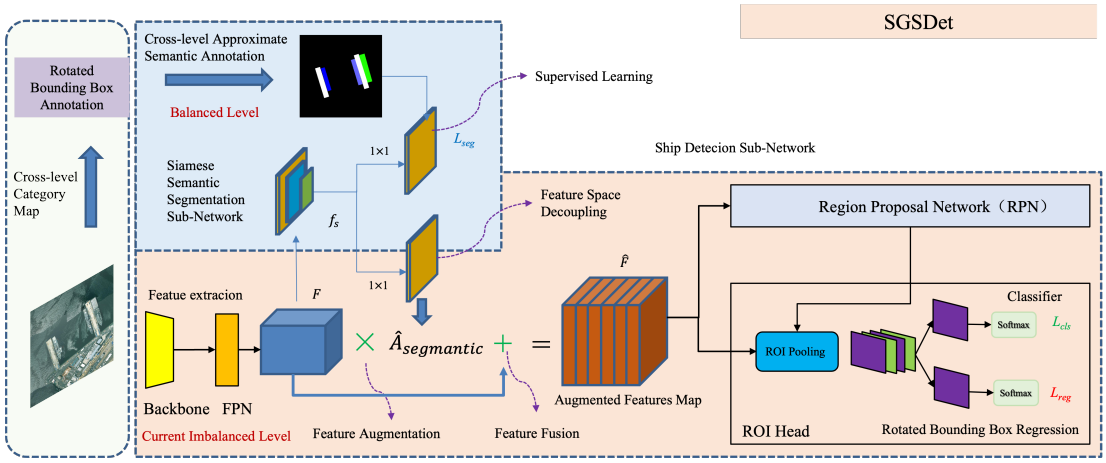

Fig. 3. The structure of our SGSDet.

augmentation processing can be formulated as:

$$\begin{aligned}\hat{F} &= F \odot \hat{A}_{\text{semantic}} \\ &= \bigcup_{i=1}^{M+1} F^i \odot \hat{A}^i_{\text{semantic}} \\ &= \bigcup_{i=1}^{M+1} F^i \odot (\hat{G}^i \cup \hat{S}^i \cup \hat{B}) \\ &= \bigcup_{i=1}^{M+1} F^i \odot \hat{G}^i \cup \bigcup_{i=1}^{M+1} F^i \odot \hat{S}^i \cup F^i \odot \hat{B})\end{aligned} \tag{4}$$

where F and $\hat{F} \in R^{C\times H\times W}$ denote the input feature map and output feature map after augmentation, i is class index. M denotes the number of ship classes at the another classification level, $\hat{A}_{\text{semantic}} \in R^{C\times H\times W}$ denotes the cross-level approximate decoupling matrix, $\hat{G}$ denotes the approximate class-generic features, $\hat{S}$ denotes the approximate class-specific features, and $\hat{D}$ denotes the approximate background feature.

D. Overview of Framework

Fig.3 illustrates the pipeline of the proposed Siamese Semantic Segmentation Guided Ship Detection Network (SGSDet), when merging with Faster R-CNN Network, it consists of two sub-networks.

- Ship Detection Sub-Network: used for feature extraction, feature fusion, classification and rotated bounding box regression.
- Siamese Semantic Segmentation Sub-Network: a simple segmentation network take feature maps from FPN of Ship Detection Network for generating approximate feature decoupling matrix.

E. Ship Detection Sub-Network

Ship Detection Sub-Network consists of five modules.

1) Backbone: For feature extraction, same to Faster R-CNN.
2) Neck: Using FPN for fusing multi-scale features.
3) Feature Fusion Module: For feature fusion of the assigned feature map and the augmented feature map.
4) Region Proposal Network (RPN): For generating region proposals, same to Faster R-CNN.
5) ROI Head: Converts generated proposals from RPN to a fixed size, then run a classifier and regress a rotated bounding box.

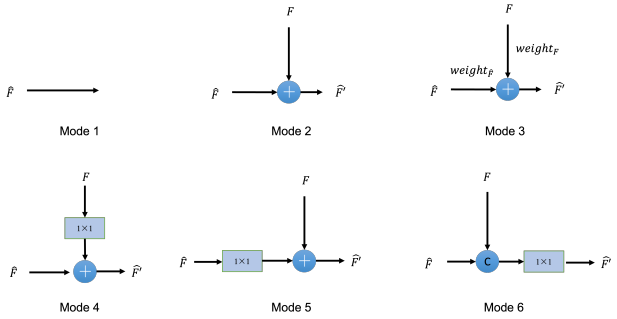

Fig. 4. Features map fusion modes.

As shown in Fig.4, six different modes are proposed for feature fusion. After feature fusion, RPN with ROI Head are used to classify the fused features.

F. Siamese Semantic Segmentation Sub-Network

Since the input feature map of Siamese Semantic Segmentation Sub-Network (SSSNet) is multi-branched, the decoupling matrix should keep the same dimension as the input. Based on above analysis, the proposed SSSNet is designed as shown in Fig. 5. SSSNet consists of multi-branch cascade convolutions for generating multi-scale features. The output of these concatenated convolutions will be concatenated together for semantic segmentation after a 1×1 convolution and interpolated to the same size, and then output in parallel after a 1×1 convolution and concatenated as a decoupling matrix.

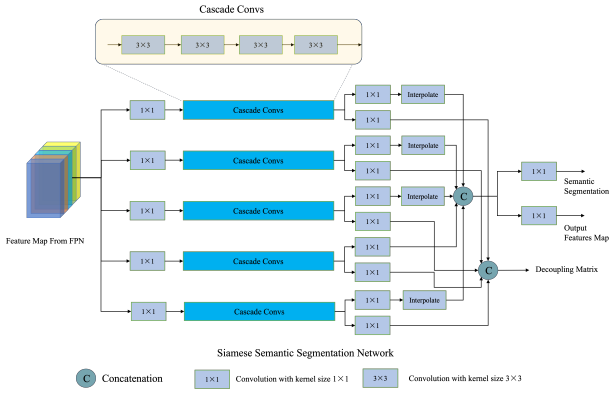

Fig. 5. The structure of our SSSNet.

G. Loss Function Design

The overall loss of the SGSDet can be formulated as:

$$\mathcal{L} = \mathcal{L}_{RPN} + \mathcal{L}_{reg} + \mathcal{L}_{cls} + \mathcal{L}_{seg} \tag{5}$$

where $\mathcal{L}_{RPN}$ is the RPN loss in two-stage detector, $\mathcal{L}_{reg}$ denotes the rotated bounding box regression loss, $\mathcal{L}_{cls}$ denotes the classification loss, and $\mathcal{L}_{seg}$ denotes the Siamese semantic segmentation loss. $\mathcal{L}_{RPN}$ and $\mathcal{L}_{cls}$ all use standard cross-entropy loss. The Smooth L1 loss is used for rotated bounding box regression, and Focal loss is used as the semantic segmentation loss.

III. EXPERIMENTAL RESULTS

A. Dataset, Evaluation Metrics, and Experiments Setting

A variety of datasets have been established for long-tailed image classification and object detection in recent years. For fine-grained ship detection, there are three benchmark datasets: HRSC2016 [8], and ShipRSImageNet [6], up to now, ShipRSImageNet is the largest fine-grained remote sensing dataset for ship detection. It contains over 3435 images with 17573 ship instances in 50 categories, providing precise horizontal and orientated bounding boxes annotation. Due to the variety of ship types, ShipRSImageNet is a long-tailed ship detection dataset. We evaluate the performance of the proposed SGSDet architecture on the ShipRSImageNet dataset [6]. Following common practice, we use total average precision mAP, and mAP for different AIAIR ($AIAIR < 5$, $5 < AIAIR < 10$, $AIAIR > 10$) as metrics for performance evaluation.

All the evaluation are implemented with the PyTorch framework [9] and MMdetection toolkit [10] in default settings. For all experiments, we use 2 NVIDIA RTX 3090 GPUs each with 24GB memory, batch size is set as 4, and the SGD optimizer is utilized to train the model. Models are trained 100 epochs; the initial learning rate is set as 0.01 and decreases by a ratio of 0.001 in the $80th$ and $90th$ epoch.

B. AAAIR of the ShipRSImageNet dataset

The sorted instance number of each class and corresponding average mAP of different object detection methods in ShipRSImageNet are shown in Fig.6. We can deduce from the Fig.6 that the ship detection performs better when the number of training samples decreases. The sorted AIA of each class is positively correlated with the corresponding average mAP of different object detection methods, as shown in Fig.7.

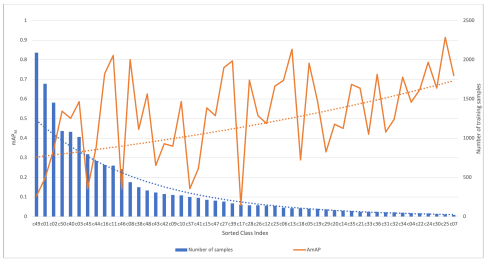

Fig. 6. Class distribution and total average mAP of ship detection task based on ShipRSImageNet dataset.

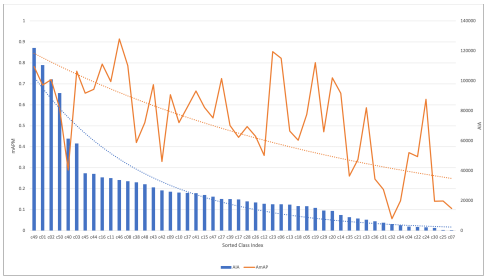

Fig. 7. Average instance area distribution and average mAP of ship detection task based on ShipRSImageNet dataset.

Ships in ShipRSImageNet dataset [6] are hierarchically classified into four levels. The IR and AIAIR are also different because the number of classes is different for different classification levels, as summarized in Table I. The larger the AIAIR, the larger the imbalance extent of the long-tailed distributions and multi-scale coupled problem.

C. Evaluation Results on ShipRSImageNet

1) Qualitative Evaluation: We qualitatively compare the differences between SGSDet (with cross-level segmentation from level0) and Faster R-CNN (Update for oriented bounding boxes detection, OBB) on ShipRSImageNet Level3 ship detection task [6]. The sample detection results are illustrated in Fig. 8. For a fair comparison, we employ ResNet50 as backbone network and compare the detection performance of the same images with score-threshold 0.7. Fig.8 shows that our SGSDet effectively captures small objects and performs better on complex scenes by exploring more discriminating features.

TABLE I
IR AND AIAIR OF THE SHIPRSIMAGENET DATASET.

Classification Level	# classes	IR	AIAIR
Level 0	2	15.08	1.18
Level 1	4	6.83	4.23
Level 2	25	14.32	3223.04
Level 3	50	90.91	259.74

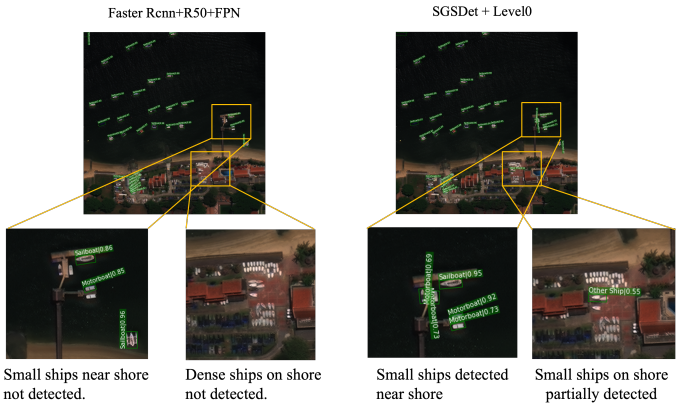

Fig. 8. Sample detection results from Faster R-CNN and our SGSDet method on ShipRSImageNet.

TABLE II
SHIP DETECTION RESULTS ON THE SHIPRSIMAGENET LEVEL3 SHIP DETECTION TASK WITH DIFFERENT FUSION MODES

Fusion Mode	mAP	Gain
Baseline	62.23	-
Mode 1	62.64	0.41
Mode 2	63.74	1.51
Mode 3	62.54	0.31
Mode 4	64.55	2.32
Mode 5	62.82	0.59
Mode 6	62.96	0.73

2) Quantitative Evaluation: We compare the detection performance of proposed SGSDet using Faster R-CNN (OBB) as baseline. Table II reports the experimental results on the ShipRSImageNet Level3 ship detection task using Level0 as the cross-level semantic segmentation but with different fusion modes. As Shown in Table II, the mAP gain obtained with fusion mode 4 is significantly better than the other modes. As shown in Table III, SGSDet improves the mAP accuracy by a clear margin guided by different cross-level semantic segmentation. When using the semantic information of level3 and with the fusion mode 0, the method is equivalent to Instance-level Feature Map Denoising proposed by Yang [5]. Using a more balanced classification level, such as the semantic information of level 0 and level 1, as the guiding information for feature space decoupling, is the most effective. Hence can deduce the proposed feature space augmentation method can significantly improve long-tailed classification accuracy.

IV. CONCLUSION

This paper studied an feature space decoupling and augmentation method guided by cross-level semantic segmentation for fine-grained ship detection. The proposed scheme consisted of three major steps: 1) Generate decouple matrix by using the Siamese Semantic Segmentation sub-Network, 2) Feature augmentation features map using the decouple matrix, 3) feature fusion. In the experiments, it was shown that the proposed SGSDet can effectively decouple and enhance feature space guided by a more balanced classification level semantic segmentation, hence greatly facilitated detection performance.

TABLE III
QUANTITATIVE COMPARISON OF EXPERIMENTAL RESULTS OF SGSDET ON THE LEVEL 3 SHIP DETECTION TASK OF SHIPRSIMAGENET DATASET

[3]

#Level	Fusion Mode	mAP			
		Total	*AIAIR<5*	5<AIAIR<10	*AIAIR>10*
-	-	62.23	77.42	65.31	37.03
0	mode 4	64.55(+2.32)	79.20(+1.78)	67.91(+2.61)	39.79(+2.76)
0	mode 0	62.54(+0.31)	78.89(+1.47)	67.32(+2.01)	36.47(-0.57)
1	mode 4	63.97(+1.33)	78.83(+1.41)	67.14(+1.82)	39.12(+2.08)
1	mode 0	63.21(+0.98)	77.30(-0.11)	66.54(+1.23)	39.26(+2.23)
2	mode 4	62.39(+0.16)	76.62(-0.80)	65.56(+0.33)	38.34(+1.31)
2	mode 0	62.53(+0.30)	74.47(-2.95)	67.88(+2.58)	39.36(+2.33)
3	mode 4	62.51(+0.28)	76.59(-0.83)	66.21(+0.89)	38.34(+1.15)
3	mode 0	62.62(+0.39)	76.64(-0.78)	66.97(+1.66)	37.62(+0.58)

[1] All experiments based on Faster R-CNN framework.
[2] All experiments use ResNet50 as the backbone, and use the standard FPN after backbone.
[3] # Level indicates which classification level should be used for cross-level semantic segmentation guided feature space decoupling and augmentation.

REFERENCES

[1] T. Wang, Y. Li, B. Kang, J. Li, J. Liew, S. Tang, S. Hoi, and J. Feng, "The devil is in classification: A simple framework for long-tail instance segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 728–744.

[2] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[3] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," *arXiv preprint arXiv:2110.04596*, 2021.

[4] J. Liu, Y. Sun, C. Han, Z. Dou, and W. Li, "Deep representation learning on long-tailed data: A learnable embedding augmentation perspective," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2970–2979.

[5] X. Yang, J. Yan, X. Yang, J. Tang, W. Liao, and T. He, "SCRDet++: Detecting Small, Cluttered and Rotated Objects via Instance-Level Feature Denoising and Rotation Loss Smoothing," *arXiv:2004.13316 [cs, eess]*, Apr. 2020.

[6] Z. Zhang, L. Zhang, Y. Wang, P. Feng, and R. He, "ShipRSImageNet: A large-scale fine-grained dataset for ship detection in high-resolution optical remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 8458–8472, 2021.

[7] P. Chu, X. Bian, S. Liu, and H. Ling, "Feature space augmentation for long-tailed data," in *European Conference on Computer Vision*. Springer, 2020, pp. 694–710.

[8] Z. Liu, J. Hu, L. Weng, and Y. Yang, "Rotated region based CNN for ship detection," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 900–904.

[9] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019, pp. 8026–8037.

[10] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open MMLab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.