PONDERING ABOUT TASK SPATIAL MISALIGNMENT: CLASSIFICATION-LOCALIZATION EQUILIBRATED OBJECT DETECTION

Yudong Zhang[4,5], *Wei Lu*[2,5], *Xu Wang*[4,5], *Pengkun Wang*[4,5,*], *Yang Wang*[1,2,3,4,5,*]

[1] Key Laboratory of Precision and Intelligent Chemistry, University of Science and Technology of China (USTC), Hefei, China
[2] School of Software Engineering, USTC, Hefei, China
[3] School of Computer Science and Technology, USTC, Hefei, China
[4] School of Data Science, USTC, Hefei, China
[5] Suzhou Institute for Advanced Research, USTC, Suzhou, China

ABSTRACT

Object detection is a fundamental task in computer vision, consisting of both classification and localization tasks. Previous works mostly perform classification and localization with shared feature extractor like Convolution Neural Network. However, the tasks of classification and localization exhibit different sensitivities with regard to the same feature, hence the "task spatial misalignment" issue. This issue can result in a hedge issue between the performances of localizer and classifier. To address these issues, we first propose a novel Dynamic Coefficient Loss to simultaneously consider and balance the performances of classification and localization tasks. To well address anchor label misjudgment issue in irregular-shaped object detection, we define a new classification-aware IoU metric to assign anchors intelligently. Finally, we further introduce the localization factor into NMS by proposing a Classification-Localization balanced NMS. Extensive experiments on MS COCO and PASCAL VOC demonstrate that our proposals can improve RetinaNet by around 1.5% AP with various backbones.

***Index Terms*—** Object detection, task spatial misalignment, Non-Maximum Suppression

1. INTRODUCTION

Object detection is one of the most important and fundamental tasks in the field of computer vision. In recent years, the performance of object detection has been greatly boosted by Convolutional Neural Network (CNN) based models [1–3] which have been widely used in the field of computer vision. To date, extensive efforts have been made to address the issues of object detection with deep learning technologies, and existing solutions can be divided into two categories, anchor-based [4–11] and anchor-free [12–16] models. Nevertheless, both anchor-based and anchor-free models employ CNN as an feature extractor, then feed extracted features into diverse localizer and classifier to respectively solve localization and classification problems. As a consequence, for an individual detector, its localizer and classifier share a set of identical features that are extracted by a same CNN. On the other hand, previous works [17, 18] have already pointed out that the sensitivities of classifier and localizer to same features are different. For instance, the existence of the phenomenon so called "task spatial misalignment" can be obviously observed in Figure 1. Specifically, from this figure, we discover that localization task is more interested in the marginal details of an object, while classification task focuses more on the specific robust features of an object which are mostly non-existent in marginal areas. In conclusion, this kind of phenomenon impacts the performance of object detection in three ways: i) there exists an obvious hedge within the performances of localizer and classifier, i.e., the performance of localizer is excellent while the performance of classifier is rather poor, or vice versa. ii) For anchor-based methods, the labels of anchors are determined by the maximum overlapping between the anchors and the Ground Truth boxes. Regarding the

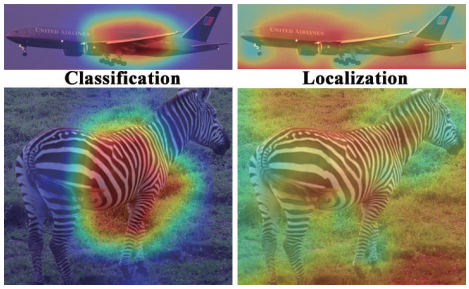

Fig. 1: **Illustration of the phenomenon so called "task spatial misalignment".** The two images in the first column are the sensitive locations for classifier, and images in the second column correspond to the sensitive locations for localizer.

* Prof. Yang Wang is the corresponding author, and Dr. Pengkun Wang is the joint corresponding author. Email: angyan@ustc.edu.cn.

task of irregular-shaped object detection, the aforementioned hedge within classifier and localizer may directly lead to the misjudgment of the anchors which are with poor localization performance but contain abundant classification information. iii) Non-Maximum Suppression (NMS) algorithms are usually employed to suppress a portion of bounding boxes after predicting the bounding boxes with an object detector. However, most traditional NMS algorithms determine which part of bounding boxes should be reserved only by taking the single metric of classification score in account, and the phenomenon of "task spatial misalignment" indicates that this kind of operation is irrational.

In light of the above findings, we propose a systematical proposal to enhance the performance of object detection by energizing existing solutions from three different aspects to neutralize the negative impacts of spatial misalignment. In particular, by integrating a dynamic coefficient to evaluate the localization error of the model, we first propose a novel loss function that targets the trade-off between the performance of localization and classification. Further, regarding the anchor label misjudgment issue of anchor-based models in irregular shaped object detection, we define a novel classification-score-considered label criterion for selecting functional anchors. This novel criterion enhances both the quality and number of positive samples simultaneously. Finally, regarding the irrational bounding box suppression issue in NMS algorithms, we add the localization scores of bounding boxes into the criterion.

2. PROPOSED METHOD

To systematically and respectively eliminate the three aspects of negative influences of the "task spatial misalignment" phenomenon, we first propose a dynamic coefficient loss, then introduce a novel score function to modify anchor assignment strategy, and finally devise a classification-localization-balanced NMS algorithm. In this section, we will introduce the detailed implementation of each individual component.

2.1. Dynamic coefficient loss

Almost for all object detection models, the loss function can be concluded as,

$$L = \sum_{i=1}^{N+M} L_{cls}\left(\widehat{p_i}, p_i\right) + \sum_{i=1}^{N} L_{reg}\left(\widehat{x_i}, x_i\right) \tag{1}$$

where L_{cls} and L_{reg} correspond to the losses of classification and localization respectively, N and M correspond to the numbers of positive and negative samples, $\widehat{p_i}$ and p_i are the predicted classification scores and classification label correspondingly, $\widehat{x_i}$ and x_i are the coordinates of predicted boxes and ground truth boxes. As discussed above, these two loss functions drive CNN to update model parameters in two different directions, and hence extracting different features. Our target here is to equilibrate these two independent loss functions. Specifically, if the localization error is relatively large, we hope the classification gradient of the overall loss function can be relatively small, so that the extracted features are more beneficial to localization, otherwise the classification gradient may drive model parameters to update more in the direction of classification. Similarly, the classification gradient should be relatively large in case that the localization error is relatively small. To achieve this goal, we define a dynamic coefficient, which is the reciprocal of localization error, as follows,

$$\lambda\left(x_i, \widehat{x_i}\right) = \frac{1}{\left\{ \begin{array}{l} \text{Sigmoid}\left[\left(c_i - \widehat{c_i}\right)^2\right] + \\ \text{Tanh}\left(\left|\frac{w_i}{h_i} - \frac{\widehat{w_i}}{\widehat{h_i}}\right|\right) + \varepsilon \end{array} \right\}} \tag{2}$$

where $\widehat{c_i}$ and c_i indicate the center coordinates of the i-th predicted bounding boxes and its corresponding ground truth respectively, $\widehat{w_i}$ and $\widehat{h_i}$ correspond to the width and height of the i-th predicted bounding boxes, and w_i and h_i are the width and height of the ground truth. The additional item ε here is to ensure that the denominator cannot be 0. To make sure that the value range of the overall loss function remains invariant, we then normalize the coefficient λ by employing the classification loss function, i.e.,

$$\lambda'\left(x_i, \widehat{x_i}\right) = \lambda\left(x_i, \widehat{x_i}\right) \frac{\sum_{j=1}^{N} L_{cls}(p_j, \widehat{p_j})}{\sum_{j=1}^{N} \lambda\left(x_j, \widehat{x_j}\right) L_{cls}(p_j, \widehat{p_j})} \tag{3}$$

Then we integrate the dynamic coefficient λ' into the overall loss function of object detection as,

$$\mathcal{L} = \left\{ \begin{array}{l} \lambda'\left(x_i, \widehat{x_i}\right) \sum_{i=1}^{N} L_{cls}(p_i, \widehat{p_i}) + \\ \sum_{i=1}^{M} L_{cls}(p_i, \widehat{p_i}) + \sum_{i=1}^{N} L_{reg}(x_i, \widehat{x_i}) \end{array} \right\} \tag{4}$$

By employing the dynamic coefficient, the loss function will possess some interesting characteristics. The dynamic coefficient decreases with the increasing localization error, hence reducing the classification gradient of weight updating and forcing CNN to focus more on the features of localization. Inversely, in case that the localization is relatively small, the dynamic coefficient is also relatively large and the classification gradient may be comparatively large, therefore the updated weights may drive CNN to pay more attention to the features of classification.

2.2. IoU-classification-aware anchor assignment

Regarding traditional anchor-based models, the label γ_i of the i-th anchor is determined by,

$$\gamma_i = \left\{ \begin{array}{cc} 1 & in\ case\ IoU_i \geq fg\_threshold \\ 0 & in\ case\ IoU_i \leq bg\_threshold \\ -1 & otherwise \end{array} \right. \tag{5}$$

where IoU_i is the IoU between the i-th anchor and its corresponding ground truth, $fg\_threshold$ and $bg\_threshold$ are the corresponding thresholds for judging the positive and negative samples. Notice that $\gamma_i = -1$ means that the i-th anchor is ignored. Intuitively, traditional approaches only consider the IoU between an anchor and its corresponding ground truth in judging the label of this anchor. We define a novel IoU-classification-aware score for judging the labels of anchors more accurately. This new score comprehensively considers both localization performance and classification information, and can be used to recall some anchors with low IoUs but extensive object information. For the i-th anchor, the IoU-classification-aware score can be calculated by,

$$s_i^{iou} = \alpha * \widehat{p_i} + (1 - \alpha) * \frac{area_i^{intersect}}{area_i^{min}} \tag{6}$$

where $area_i^{intersect}$ is the intersection area between the i-th anchor and its corresponding ground truth, $area_i^{min}$ corresponds to the minimum area between the i-th anchor and its corresponding ground truth, parameter α is an adjustable weight. The new protocol for judging the label Γ_i of the i-th anchor can be written as,

$$\Gamma_i = \begin{cases} 1 & in\ case\ IoU_i \geq fg\_threshold \\ & or\ s_i^{iou} \geq score\_threshold \\ 0 & in\ case\ IoU_i \leq bg\_threshold \\ -1 & otherwise \end{cases} \tag{7}$$

where $score\_threshold$ indicates the new threshold for judging the positive and negative samples in terms of the new IoU-classification-aware score. This novel IoU-classification-aware score not only imports the classification score of anchor, but also updates the metric for evaluating the localization performance of anchors.

2.3. Classification-localization-balanced NMS

We modify traditional NMS algorithms by adding the localization scores of candidate bounding boxes into the selection criterion of the NMS algorithms, and name the modified NMS algorithm as Classification-Localization-Balanced NMS (CL-Balanced NMS). Different from [19–21], we propose an automatic and unsupervised solution without any additional branch. Regarding an object o, we first check the classification scores of the bounding boxes $\mathcal{B} = \{b_1, \cdots, b_m\}$ of all anchors, select the box with the highest classification score within the score set $\mathcal{S} = \{\int_1, \cdots, \int_m\}$ as the reference box, and then calculate the IoU between each bounding box and the reference one. Regarding a specific bounding box, if the IoU between this box and the reference one is bigger than a pre-defined threshold $nms\_threshold$, we consider this bounding box may detect the same object that the reference box detects, and add this bounding box into a set. Regarding an object o, we then have the set of bounding boxes $\mathbb{B} = \{b_1^o, \cdots, b_n^o\}$, and the corresponding classification scores of these bounding boxes can be denoted as $\mathbb{S} = \{\int_1^o, \cdots, \int_n^o\}$. Based on the idea of bagging, we calculate the weighted average center coordinates of all the boxes within $\mathbb{B}$ by using their classification scores as weights, and take the weighted average center as the approximate center $\widehat{c^0}$ of the ground truth o, i.e.,

$$\widehat{c^o} = \sum_{i=1}^{n} \frac{\exp(\int_i^o)}{\sum_{j=1}^{n} \exp(\int_j^o)} * c_i^o \tag{8}$$

where c_i^o is the center coordinates of the bounding box b_i^o. After generating the approximate center coordinate of the ground truth box, we define the normalized Euclidean distance between the center coordinate of each bounding box within $\mathbb{B}$ and c_i^o as the localization score of the corresponding box, and integrate the new localization score into the selection criterion of NMS algorithm. Regarding a specific bounding box $b_i^o (1 \leq i \leq n)$, the new selection score of NMS algorithm can be written as,

$$s\_nms_i^o = p_i^o + \mathrm{dis}(c_i^o, \widehat{c^o}) \tag{9}$$

where p_i^o is the classification score of bounding box b_i^o, and dis indicates the function for calculating the normalized Euclidean distance. After calculating the new score, CL-Balanced NMS employs this new score as the criterion to suppress bounding boxes.

Method	Backbone	AP	AP^{50}	AP^{75}	AP^s	AP^m	AP^l
YOLOv3 [8]	DarkNet-53	33.0	57.9	34.4	18.3	35.4	41.9
Faster R-CNN [4]	ResNet-101	36.2	59.1	39.0	18.2	39.0	48.2
Mask R-CNN [22]	ResNet-101	38.2	60.3	41.7	20.1	41.1	50.2
Deformable R-FCN [23]	Aligned-Inception-ResNet	37.5	58.0	40.8	19.4	40.1	52.5
RetinaNet [9]	ResNet-101	39.1	59.1	42.3	21.8	42.7	50.2
IoU-Net [21]	ResNet-101	40.0	59.0	-	-	-	-
Ours	**ResNet-101**	**40.6**	**60.2**	**43.5**	**23.9**	**43.9**	**51.1**
GHM [24]	ResNeXt-101	41.6	62.8	44.2	22.3	45.1	55.3
Faster R-CNN [4]	ResNeXt-101	40.3	62.7	44.0	24.4	43.7	49.8
Mask R-CNN [22]	ResNext-101	41.4	63.4	45.2	24.5	44.9	51.8
FCOS [12]	ResNeXt-101	42.1	62.1	45.2	25.6	44.9	52.0
CornerNet [14]	Hourglass-104	40.5	56.5	43.1	19.4	42.7	53.9
RetinaNet [9]	ResNeXt-101	40.8	61.1	44.1	24.1	44.2	51.2
Ours	**ResNeXt-101**	**42.2**	**62.5**	**45.1**	**25.5**	**44.9**	**52.3**
ATSS [19]	ResNeXt-101-64x4d-DCN	47.7	66.5	51.9	29.7	50.8	59.4
GFL [25]	ResNeXt-101-32x4d-DCN	48.2	67.4	52.6	29.2	51.7	60.2
Ours	**ResNeXt-101-32x4d-DCN**	**48.9**	**67.8**	**53.2**	**30.0**	**52.2**	**61.7**

Table 1: **performance of alternative detectors with different backbones on MS-COCO test-dev.**

3. EXPERIMENTS

3.1. Experiment settings

Datasets. We conduct extensive experiments on MS COCO 2017 dataset [26] and PASCAL VOC dataset [27]. MS COCO has 80 object categories. We train models on train2017 and report results on val2017 and test-dev. PASCAL VOC provides 20 object categories. We train models on the VOC 2007 and VOC 2012 trainval sets, and evaluate them on the VOC 2007 test set.

IAC -score	DC Loss	CL-B NMS	VOC 2007 test mAP	MS COCO 2017 AP	AP^{50}	AP^{75}
			78.2	35.9	56.1	38.6
✓			78.6	36.4	56.5	39.1
✓	✓		79.2	37.1	56.3	39.7
✓	✓	✓	79.7	37.5	56.9	40.0

Table 2: **Contributions of individual components on COCO val and PASCAL VOC 2007 test.** The baseline is ResNet-50 RetinaNet.

backbone	NMS	Soft	CL-B	AP	AP^{50}	AP^{75}
ResNet-50	✓			36.4	55.5	39.0
		✓		36.5	55.7	39.0
			✓	37.0	56.2	39.3
ResNet-101	✓			38.5	57.6	41.0
		✓		38.6	57.5	41.4
			✓	39.0	58.0	41.6

Table 3: **Impacts of NMS algorithms on COCO val2017.** Comparison of CL-Balanced NMS with other NMS using RetinaNet with ResNet-50 and ResNet-101 as backbone.

Implementation Details. We exploit ResNet-50, ResNet-101, ResNeXt-101 [28] and deformable ResNeXt-101 as backbones, The models are trained on 8 Tesla V100 GPUs with the batch size of 16 (2 images per GPU). Stochastic Gradient Descent(SGD) with the weight decay of 0.0001 and the momentum of 0.9 is adopted for optimizing these backbone networks. Regarding ablation studies, we use ResNet-50 as the backbone, and execute 12K training iterations for VOC 2007 test and 90K training iterations for COCO val2017 respectively. While comparing the performance between our model and some other state-of-the-art solutions on the COCO test-dev, we extend our proposed methods to several above-mentioned backbones and train them with 180K iterations. The initial learning rate is 0.01. For COCO dataset, learning rate is decreased by a factor of 10 after 67.5K and 75K iterations respectively while the setting is 90K, and is decreased with the same factor after 135K and 150K iterations respectively in case of the setting of 180K. Further, for VOC dataset, learning rate is also decreased by a factor of 10 after 9K iterations. A linear warmup strategy is adopted in the first 500 iterations. We set α as 0.75 and $score\_threshold$ as 0.5 in Equation 6 according to experiments in Supplementary Document. For $fg\_threshold$ and $bg\_threshold$ in Equation 7, we inherit the optimized settings ($fg\_threshold$ is typically set to 0.5 and $bg\_threshold$ is set to 0.4) in [9].

3.2. Main results

To investigate the performance of our proposed methods, we use them to modify RetinaNet and compare the modified network with other state-of-the-art detectors by conducting extensive experiments with different backbonesThe results are shown in Table 1. In case of using RetinaNet with backbone ResNet-101 and ResNext-101, our methods achieve 1.5% and 1.4% AP improvements respectively, verifying the superiority of our proposed methods. Worth noting that we use DC Loss and CL-Balanced NMS to modify FCOS with Generalized Focal Loss [25] and achieve an amazing high AP of 48.9%.

3.3. Ablation Studies

To demonstrate the effectiveness of the three individual components of our methods, we use these components to modify RetinaNet. The performance of these variations is evaluated with both COCO val2017 and VOC 2007 test. The results are respectively demonstrated in Table 2. We first update traditional anchor assignment mechanism with IoU-classification-aware score, and the results are given in the second row of these two tables. As shown, the performance is boosted from 35.9% to 36.4% on COCO val2017 and from 78.2% to 78.6% on VOC 2007 test respectively. Based on this modification, we further add DC Loss for extending RetinaNet, and the results are shown in the third row of these two tables. As can be observed, the employment of DC Loss can improve AP by 0.7% and mAP by 0.6% on COCO val2017 and VOC 2007 test respectively. Finally, we replace conventional NMS in RetinaNet with CL-Balanced NMS, and this replacement operation can bring 0.4% AP improvement on COCO val2017 and 0.5% mAP improvement on VOC 2007 test.

Besides, to further investigate the superiority of CL-Balanced NMS, we conduct extensive experiments on the detector of RetinaNet with different backbones and different NMS algorithms. Table 3 reports the results on COCO val2017. As shown, CL-Balanced NMS outperforms all other alternative NMS algorithms, indicating the validity of simultaneously considering classification and localization in NMS.

4. CONCLUSION

In this paper, we propose a systematical method to enhance the performance of object detection by energizing existing solutions from three different aspects to neutralize the negative impacts of task spatial misalignment. Specifically, a novel DC Loss is proposed to address the "task spatial misalignment" in object detection. Further, IoU-classification-aware score is devised to involve classification scores during assigning labels of anchors. Finally, CL-Balanced NMS is designed to address the misalignment between classification and localization via adding the localization score of candidate bounding boxes into conventional NMS algorithm.

Acknowledgement. This work was partially supported by the National Natural Science Foundation of China (No.62072427, No.12227901), the Project of Stable Support for Youth Team in Basic Research Field, CAS (No.YSBR-005), and the Academic Leaders Cultivation Program, USTC.

5. REFERENCES

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," 2016.

[2] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017.

[3] Qian Xie, Yu Kun Lai, Jing Wu, Zhoutao Wang, Yiming Zhang, Kai Xu, and Jun Wang, "Mlcvnet: Multi-level context votenet for 3d object detection," 2020.

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Sun Jian, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 6, 2015.

[5] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin, "Libra r-cnn: Towards balanced learning for object detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[6] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *Computer Vision & Pattern Recognition*, 2016.

[7] Joseph Redmon and Ali Farhadi, "Yolo9000: Better, faster, stronger," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2017.

[8] Joseph Redmon and Ali Farhadi, "Yolov3: An incremental improvement," *arXiv e-prints*, 2018.

[9] Tsung Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr DollÂ¨Â¢r, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. PP, no. 99, pp. 2999–3007, 2017.

[10] Zhaowei Cai and Nuno Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," 2017.

[11] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, and Alexander C. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*, 2016.

[12] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He, "Fcos: Fully convolutional one-stage object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2020.

[13] Xingyi Zhou, Dequan Wang, and Philipp KrhenbÂ¨Â¹hl, "Objects as points," 2019.

[14] Hei Law and Jia Deng, "Cornernet: Detecting objects as paired keypoints," *International Journal of Computer Vision*, 2018.

[15] Xingyi Zhou, Jiacheng Zhuo, and Philipp KrhenbÂ¨Â¹hl, "Bottom-up object detection by grouping extreme and center points," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[16] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin, "Reppoints: Point set representation for object detection," 2019.

[17] Rui Zhu, Shifeng Zhang, Xiaobo Wang, Longyin Wen, and Tao Mei, "Scratchdet: Training single-shot object detectors from scratch," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[18] Guanglu Song, Yu Liu, and Xiaogang Wang, "Revisiting the sibling head in object detector," 2020.

[19] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang, "Bounding box regression with uncertainty for accurate object detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[20] Kang Kim and Hee Seok Lee, "Probabilistic anchor assignment with iou prediction for object detection," 2020.

[21] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang, "Acquisition of localization confidence for accurate object detection," 2018.

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[23] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun, "R-fcn: Object detection via region-based fully convolutional networks," 2016.

[24] Buyu Li, Yu Liu, and Xiaogang Wang, "Gradient harmonized single-stage detector," 2019.

[25] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," *arXiv*, 2020.

[26] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, and C. Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014.

[27] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[28] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He, "Aggregated residual transformations for deep neural networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.