Supplementary Material for : Applying Mutual Guidance to Anchor-free detectors

Heng ZHANG[1,3][0000−0001−6093−1729], Elisa FROMON[1,4][0000−0003−0133−3491], Sébastien LEFEVRE[2][0000−0002−2384−8202], and Bruno AVIGNON[3]

[1] Univ Rennes, IRISA, France
[2] Univ Bretagne Sud, IRISA, France
[3] ATERMES, France
[4] IUF, Inria, France

Abstract. The principle of *Mutual Guidance* in object detection is to assign labels of one task according to the prediction on the other task, and vice versa. Apparently, it is not limited to anchor-based methods, but applicable for any object detector that performs localization and classification tasks. Here we introduce the application of *Mutual Guidance* in anchor-free methods. Experiments conducted on PASCAL VOC dataset demonstrate the consistent precision improvements brought by our method on this category of detectors.

1 Summary of Mutual Guidance

In order to realize a content/context-sensitive label assignment and avoid the task-misalignment problem, we propose the *Mutual Guidance* mechanism in object detection, which can be summarised as follows:

- When assigning labels for the classification task:
  - If the prediction of localization is good, the label is assigned as positive (i.e., "object") to encourage good predictions on both tasks;
  - If the prediction of localization is poor, the label is assigned as negative (i.e., "background") to avoid misaligned false positive detection;
- When assigning labels for the localization task:
  - If the prediction of classification is good, the label is assigned as positive (i.e., we optimize this sample) to encourage good predictions on both tasks;
  - If the prediction of classification is poor, the label is assigned as negative (i.e., we do not optimize this sample) to avoid unnecessary optimizations.

2 Applying Mutual Guidance to FCOS

Fully Convolutional One-Stage Object Detection (FCOS) [1] is one of the most representative anchor-free methods, which solves object detection in a per-pixel

Model	Matching strategy	*AP*	*AP*50	*AP*75
FCOS with	original strategy	49.2%	73.7%	52.1%
ResNet-18 backbone	*Mutual Guidance*	**51.1%**	**74.4%**	**54.3%**
FCOS with	original strategy	53.9%	78.4%	57.4%
VGG-16 backbone	*Mutual Guidance*	**55.9%**	**79.4%**	**60.2%**

Table 1. Comparison of different label assignment strategies (the original one and *Mutual Guidance*) for FCOS. Experiments are conducted on the PASCAL VOC dataset. The best score for each architecture is in bold.

prediction fashion. On each pixel of feature maps, it classifies the category of this sample point and regresses the four distances to the target bounding box borders. When assigning training labels in FCOS, firstly the corresponding detection layer is selected according to the scale of the object to detect, then positive samples (for both localization and classification tasks) are assigned to all the points inside the ground truth box. Based on that, [2] proposes to only sample the points in the central region of the ground truth box.

When applying *Mutual Guidance* to FCOS, identical to the implementation in anchor-based methods, two strategies are applied: *Localize to Classify* and *Classify to Localize*. Moreover, the same dynamic thresholding strategy is applied. Specifically, for *Localize to Classify*, we firstly note the number of positive samples assigned by the original strategy (N_p), then label the N_p highest $IoU_{regressed}$ points as positive; for *Classify to Localize*, we amplify each point's centerness score according to its classification prediction, and label the N_p highest amplified centerness score points as positive.

Experiments are performed on the PASCAL VOC dataset. ResNet-18 [3] and VGG-16 [4] are adopted as backbone networks in our experiments. Unless specified, all other implementation details are the same as in our paper. Experimental results are listed in Table 1. Same as with anchor-based methods, our *Mutual Guidance* strategy significantly boosts the detection precision for FCOS, especially on the AP75 metric.

We then conduct qualitative analysis on the label assignment difference and detection result difference between the original FCOS strategy and *Mutual Guidance*. As shown in Figure 1, the original strategy (red points) assigns positive to points in the central region of the object box regardless of their content/context, whereas the *Localize to Classify* (yellow points) and *Classify to Localize* (green points) strategies adaptively assign positive to points with representative semantic information, and assign negative to points on background or nearby objects. Since the labels assigned by the original FCOS strategy are always the same for localization and classification tasks, the task-misalignment problem exists in FCOS as well. Several misaligned false positive detections are observed in Figure 2, however, the proposed *Mutual Guidance* provides more accurate detection.

Fig. 1. Visualization of the difference in the label assignment during training phase (images are resized to 320×320 pixels). The ground truth object in each image is marked by white dotted-line box. Red, yellow and green points are positive samples assigned by original FCOS strategy, *Localize to Classify* and *Classify to Localize* respectively. Zoom in to see details.

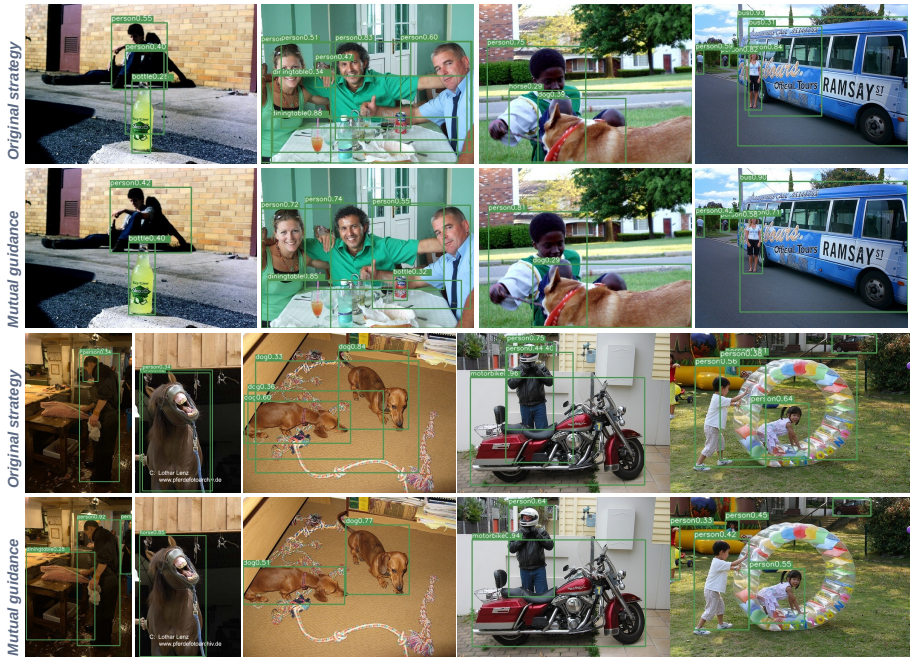

Fig. 2. Examples of detection results using the original FCOS label assignment strategy (odd lines) and our proposed *Mutual Guidance* one (even lines). The results are given for all images after applying a Non-Maximum Suppression process with a IoU threshold of 50%. Zoom in to see details.

References

1. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: fully convolutional one-stage object detection. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE (2019) 9626–9635
2. Yao, Y.: FCOS_PLUS. (https://github.com/yqyao/FCOS_PLUS)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society (2016) 770–778
4. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In Bengio, Y., LeCun, Y., eds.: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. (2015)