Aligning Localization and Classification for Anchor-Free Object Detection in Aerial Imagery

Cong Zhang, Yakun Ju, Jun Xiao, Yuting Yang, and Kin-Man Lam

***Abstract*—Aerial object detection involves two subtasks, localization and classification. Existing anchor-free detectors ignore the spatial misalignment caused by inconsistent optimization between the two subtasks, significantly degrading detection performance. To address this issue, this paper proposes a novel anchor-free detector, which explicitly aligns the multi-task predictions of localization and classification, by an aligned head and an aligned sample assignment metric. Experimental results demonstrate the superiority of the proposed method for aerial object detection.**

I. Introduction

As a fundamental yet challenging task of intelligent unmanned systems, aerial object detection aims to localize and recognize objects of interest in aerial images. It is typically formulated as a multi-task learning problem [1], [2], by jointly optimizing two subtasks, *i.e.*, object localization and classification, based on two separately parallel branches in the detection heads. Recently, anchor-free detectors enjoy lightweight architectures and high computational efficiency, thereby garnering increasing attention. Compared to anchor-based detectors with predefined enclosed anchor boxes [2], anchor-free methods, such as FCOS [3], perform multi-task predictions only on the individual center of each object candidate, referred to as "anchor points". However, such heuristic design makes anchor-free detectors more susceptible to inconsistent spatial distributions of the learned representations for the two subtasks. Specifically, they usually suffer from two deficiencies. (1) Existing separately dual-branch head leads to independence or isolation between localization and classification, degrading detection accuracy. (2) Most anchor-free detectors simply rely on a geometry-based sample assignment scheme that ignores the misalignment of anchor points required for the two different subtasks. To address the issues, this paper proposes a novel anchor-free detector, namely localization-classification-aligned (LCA) detector (LCA-Det), which aligns the two subtasks for more accurate and efficient aerial object detection.

II. Methodology

Overview. The proposed LCA-Det follows a concise single-stage detection pipeline similar to previous anchor-free detectors [3], while distinguishing itself by its two core collaborative components, LCA head (LCA-H) and LCA sample assignment (LCA-SA). As illustrated in Fig. 1, LCA-Det explicitly aligns localization and classification. Based on the initial multi-task predictions of LCA-H, *i.e.*, classification probabilities and localization precision, LCA-SA first measures the task alignment as a sample assignment metric. Then, LCA-H refines its final predictions according to this metric.

C. Zhang, J. Xiao, and K.-M. Lam are with Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong. Y. Ju is with Nanyang Technological University, Singapore. Y. Yang is with Shandong University of Science and Technology, China.

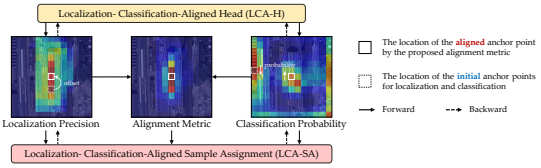

Fig. 1. Overall alignment mechanism of the proposed LCA-Det.

TABLE I
Performance comparison of different aerial object detectors

Methods	mAP_{50} (Δ)	mAP_{75}	$mAP_{50:95}$	Params	FLOPs
RetinaNet [2]	67.2	48.8	45.3	36.5M	133.1G
FCOS [3]	66.0	42.6	41.3	31.9M	123.6G
LCA-Det (Baseline)	68.1	47.0	44.4	**31.2M**	117.1G
LCA-Det (Only LCA-H)	70.7 (+2.6)	52.0	48.1	31.8M	**113.4G**
LCA-Det (Only LCA-SA)	69.8 (+1.7)	48.3	45.7	**31.2M**	117.1G
LCA-Det (LCA-H + LCA-SA)	**72.0 (+3.9)**	**52.9**	**49.3**	31.8M	**113.4G**

LCA-H aims to perform preliminary alignment between the two subtasks by enhancing their feature interaction. To generate such subtask-interactive representations, instead of utilizing two separate branches, LCA-H examines a single-branch structure, which, however, unavoidably introduces learning conflicts due to the inconsistent optimization objectives. Thus, LCA-H further adjusts the spatial distribution of the two predictions, *i.e.*, dense classification scores $\mathcal{S}$ and bounding boxes $\mathcal{B}$, through two auxiliary components, as follows:

$$\mathcal{S}^* = \sqrt{\mathcal{S} \times \mathcal{P}}, \quad \mathcal{B}^* = \mathcal{A}(\mathcal{B}, \mathcal{O}), \tag{1}$$

where $\mathcal{S}^*$ and $\mathcal{B}^*$ denote the aligned final predictions. $\mathcal{P}$ and $\mathcal{O}$ represent the learned spatial probability map and offset map, respectively, both of which are computed from the task-interactive features. $\mathcal{A}(\cdot)$ is a pixel-wise alignment function.

LCA-SA aims to make further alignment by guiding the optimization during training. It mainly comprises an advanced anchor-point alignment metric, formulated as follows:

$$m = s^{\alpha} \times b^{\beta}, \tag{2}$$

where s and b represent the classification score and IoU value of each candidate anchor point, respectively, and α and β are the balancing hyperparameters. This metric quantifies the level of subtask alignment through a high-order combination of both localization and classification measures.

III. Experiments and Conclusion

We have preliminarily evaluated the proposed method on the DIOR dataset [4], as shown in Table I, where the performance of both LCA-H and LCA-SA are clearly demonstrated. Moreover, LCA-Det significantly outperforms its competitors, including anchor-based RetinaNet [2] and anchor-free FCOS [3], in terms of both detection accuracy and efficiency.

References

[1] S. Ren, *et al.*, "Faster R-CNN: Towards real-time object detection with region proposal networks," NeurIPS, 2015, pp. 91–99.
[2] T. Lin, *et al.*, "Focal loss for dense object detection," ICCV, 2017, pp. 2980–2988.
[3] Z. Tian, *et al.*, "FCOS: Fully convolutional one-stage object detection," ICCV, 2019, pp. 9627–9636.
[4] K. Li, *et al.*, "Object detection in optical remote sensing images: A survey and a new benchmark," ISPRS P&RS, pp. 29-307, 2020.