Adaptive Multi-task Learning for Few-shot Object Detection

Supplementary Material

This supplementary material contains additional information that could not be included in the main paper due to the space constraint. It includes more experimental analysis and visualization details.

1 More Detailed COCO and VOC Results

More detailed experimental results of the proposed method in terms of mean and standard deviation values are given in Tab. S1 and Tab. S4 for COCO and PASCAL VOC datasets, respectively. These tables highlight the stability of the proposed method across multiple runs. In addition, it is evident that the standard deviation values for the VOC dataset appear to be higher than those for the COCO dataset. The COCO dataset has a broader range of object categories and more object instances compared to the VOC dataset. Increased diversity of object classes and instances for training could potentially enhance the stability of performance for few-shot object detection.

2 Frozen or Not

During the fine-tuning process of the proposed method, the Res5_{cls} block is kept frozen. The results presented in Tab. S3 indicate that by freezing only this block, the proposed method performs the best in terms of 10-shot object detection on the COCO dataset. The same trend is observed for the baseline method [32]. When Res5_{cls} is frozen but Res5_{reg} is unfrozen, an imbalanced learning ability between classification and localization tasks becomes evident, underscoring the importance of addressing task conflicts. Tab. S3 also indicates that in both the

Table S1: A detailed analysis of the detection performance of the proposed method on the COCO dataset is presented in the COCO style, taking into account mAP, mAP_{50}, and mAP_{75} of the novel classes to examine the performance. The mean and standard deviation scores are computed across multiple evaluations.

COCO	mAP	mAP_{50}	mAP_{75}
1-shot	12.8 ± 0.0	21.9 ± 0.1	13.7 ± 0.2
2-shot	16.9 ± 0.1	29.5 ± 0.1	17.4 ± 0.2
3-shot	17.5 ± 0.1	31.2± 0.1	17.4 ± 0.2
5-shot	19.5 ± 0.1	35.0 ± 0.1	18.8 ± 0.1
10-shot	22.7 ± 0.1	40.0 ± 0.1	22.3 ± 0.2
30-shot	25.2 ± 0.1	43.3 ± 0.2	25.3 ± 0.2

Table S2: A detailed analysis of the detection performance of the proposed method on the PASCAL VOC dataset is presented in the VOC style, including the mean and standard deviation values for mAP, mAP_{50}, and mAP_{75} of the novel classes. The mean and standard deviation scores are computed across multiple evaluations.

VOC	mAP	mAP_{50}	mAP_{75}
Split1			
1-shot	39.7 ± 0.9	68.9 ± 0.9	41.1 ± 0.8
2-shot	42.3 ± 1.9	71.5 ± 2.5	45.3 ± 2.9
3-shot	42.7 ± 0.7	72.1 ± 0.8	45.2 ± 1.8
5-shot	44.8 ± 0.2	74.5 ± 0.2	48.0 ± 0.7
10-shot	45.9 ± 0.3	72.2 ± 0.0	50.2 ± 0.6
Split2			
1-shot	35.7 ± 0.9	65.5 ± 1.5	32.7 ± 1.6
2-shot	40.0 ± 1.5	69.8 ± 2.6	41.2 ± 2.0
3-shot	43.9 ± 0.3	73.5 ± 0.2	46.1 ± 0.9
5-shot	46.0 ± 0.4	74.4 ± 0.5	50.1 ± 0.8
10-shot	46.5 ± 0.8	73.1 ± 0.6	51.0 ± 0.5
Split3			
1-shot	42.4 ± 0.4	68.8 ± 0.4	45.6 ± 0.8
2-shot	43.2 ± 0.2	69.8 ± 0.9	47.6 ± 0.6
3-shot	43.0 ± 0.1	70.0 ± 0.2	46.2 ± 0.2
5-shot	44.8 ± 0.1	71.6 ± 0.3	49.5 ± 0.2
10-shot	45.7 ± 0.3	71.9 ± 0.2	51.8 ± 0.7

Table S3: The performance analysis of freezing some blocks in the architectures of the baseline method and the proposed method. The experiments are conducted on the COCO dataset in 10-shot with a batch size of 8. It should be noted that no KDSR scheme is used in this table.

method	backbone	Res5		mAP
baseline [32]	-	-		14.3
	-	frozen		**16.9**
	frozen	frozen		15.8
method	backbone	$Res5_{cls}$	$Res5_{reg}$	mAP
Ours	-	-	-	13.6
	-	-	frozen	12.8
	-	frozen	-	**18.6**
	-	frozen	frozen	16.1
	frozen	frozen	-	17.3

baseline and the proposed method, an unfrozen backbone can facilitate the acquisition of more distinctive features, leading to improved detection performance.

3 The Simulation Network

Additional insights into the construction of the simulation network $\tilde{\Psi}$ are provided in Fig. S1. In Fig. S1 (a), the data episodes (L_{cls}, L_{reg}, AP) sampled through fine-tuning the proposed model on the novel support dataset D_{novel} are visualized. The additional details on collecting the data episodes can be found in

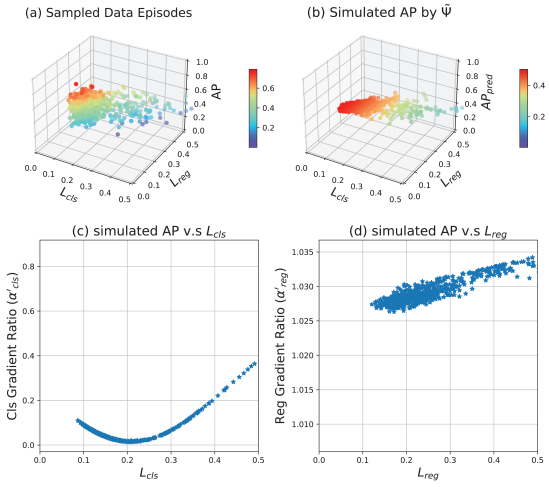

Fig. S1: (a) The illustration of the collected data episodes on the COCO dataset (b) The simulated AP results based on the simulation network. (c)-(d) The gradient ratios based on the simulated AP results.

Table S4: Ablation study on VOC-1shot dataset (Split2).

Double-head RCNN	Precision-driven Grandient Balancer	KDSR	$mAP50_{voc}$
-	-	-	32.1
✓	-	-	39.7
✓	✓	-	60.4
✓	✓	✓	65.5

Sec.4.3. Fig. S1 (b) presents the simulated AP (AP_{pred}) results generated by the simulation network $\tilde{\Psi}$. The gradient ratios for classification α'_{cls} and regression α'_{reg} (Eq. 7), derived by $\tilde{\Psi}$ based on the simulated AP results, are depicted in Fig. S1 (c) and (d) respectively. It is clear that the fine-tuning process demands distinct gradient ratios for classification and regression tasks. By prioritizing AP-maximization as the objective, the gradient balancer proposed, which relies on the simulation network, can proficiently alleviate task conflicts and enhance the FSOD performance. Additionally, we present another ablation study on VOC-1shot in Tab. S4, which shows the significant performance boost (**88.2%**) by our gradient balancer, particularly in fewer shot senarios.

4 Parameter Analysis on η

The parameter analysis of the trade-off parameter η in Eq. (9) is given in Tab. S5. For scenarios with fewer shots, such as 1-shot or 2-shot, η is increased to maximize enhancement from the adapted image-text features. Conversely, in cases with relatively higher shots, like 10-shot or 30-shot, η is decreased to preserve

Table S5: The parameter analysis of the trade-off η for CLIP-based score refinement on COCO dataset.

η	0	1	2	3	4	5	6	7	8	9
1-shot	11.1	11.4	11.5	11.8	12.4	12.7	**12.8**	12.8	12.8	12.8
10-shot	22.4	**22.6**	22.6	22.5	22.4	22.3	22.2	22.1	22.0	22.0

the knowledge acquired during the fine-tuning process. When compared to the COCO dataset, the VOC dataset is less complex. Therefore, we maintain η at 1 for all VOC shots. In addition, the attention function is defined as Eq. (S1) [54], c represents a trade-off parameter empirically set to 11 for all experiments.

$$\varrho\left(x_{\mathcal{T}} \cdot f_{sup}^{T}\right) = \exp\left(-c \cdot \left(1 - x_{\mathcal{T}} \cdot f_{sup}^{T}\right)\right) \tag{S1}$$

5 Visualization of Negative Results

The visualization of the negative results of the proposed method is given in Fig. S2. The figures can be enlarged for a clearer view. Due to the limited availability of training data, FSOD techniques struggle to classify objects belonging to novel classes with similar appearance (Fig. S2). For example, a horse with white markings on its brown body might be incorrectly classified as a cow. Therefore, there is still room for further improvement in the FSOD performance.

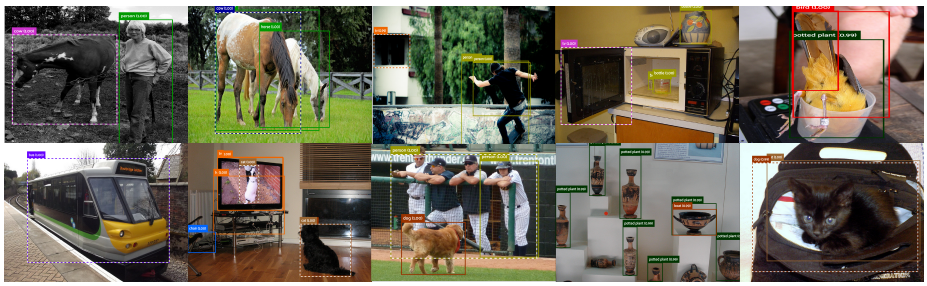

Fig. S2: Visualization of the proposed method's negative results on the COCO dataset under the 10-shot setting. The detection scores have been adjusted using the proposed detection score refinement scheme (SR).