Semantic-aligned Fusion Transformer for One-shot Object Detection Supplementary Material

Anonymous CVPR submission

Paper ID 2246

A. Implementation Details

Since SaFT is a general fusion neck, it may employ any detector-specific loss. For the FCOS-based detector as in our main paper, we simply combine losses of FCOS [6] and DETR [1]. Further implementations on two-stage detectors can try correspondingly related losses as well as metric-learning ones.

Given network predictions for classification c, regression r, center-ness t as defined in [6] and their corresponding targets c^*, r^*, t^* respectively, we present our loss function as follows

$$\begin{aligned}\mathcal{L} = &\frac{\lambda_{cls}}{N}\sum_{x,y}\mathcal{L}_{cls}(c_{x,y}, c^*_{x,y}) \\ &+ \frac{\lambda_{reg}}{N_{pos}}\sum_{x,y}\mathbb{1}_{\{c^*_{x,y}>0\}}\mathcal{L}_{reg}(r_{x,y}, r^*_{x,y}) \\ &+ \frac{\lambda_{ctn}}{N_{pos}}\sum_{x,y}\mathbb{1}_{\{c^*_{x,y}>0\}}t^*_{x,y}\mathcal{L}_{ctn}(t_{x,y}, t^*_{x,y})\end{aligned} \tag{1}$$

where $\mathcal{L}_{cls}$ is focal loss [3], $\mathcal{L}_{reg}$ is a joint of L1 loss and GIoU loss [5] as in [1], and $\mathcal{L}_{ctn}$ is binary cross entropy (BCE) loss for center-ness as in [6]. λ_{cls}, λ_{reg} and λ_{ctn} are balance weights, being 20, 2 and 0.5 so as to keep three terms in the same scale. N and N_{pos} denote the number of all locations and that of positive samples.

B. Ablation Study

For further insights into SaFT, we use the same setup as in our main paper to perform more ablation studies. Experiments are carried out on VOC with a mini-batch size of 4. Tab. 1 adopts $N = 4$ HA blocks by default, and Tab. 3 discusses the effect of HA block numbers.

B.1. Feature Levels Used in Fusion

In Tab. 1, we explore the effect of feature fusion with different query-support feature levels. For corresponding-scale fusion in rows 1-3, we can see that more levels of features utilized do not necessarily mean better performances.

Cross-sample	Query 4	Query 5	Query 6	Support 4	Support 5	Support 6	Base	Novel
Corresponding-scale	✓			✓			75.8	61.4
	✓	✓		✓	✓		79.2	70.0
	✓	✓	✓	✓	✓	✓	77.5	69.9
One-to-all-scale	✓				✓		76.0	62.5
		✓			✓		76.3	60.7
			✓		✓		74.1	57.7
	✓	✓			✓		79.5	67.9
		✓	✓		✓		79.4	65.8
One-to-all-scale	✓	✓	✓	✓			78.7	68.1
	✓	✓	✓		✓		79.5	71.7
	✓	✓	✓			✓	78.2	69.7

Table 1. Ablation study for feature levels used in fusion on VOC. One-to-all-scale means associating a single level of support features with all available query features, whereas corresponding-scale is limited to corresponding levels. All experiments use VFM for cross-scale fusion and HFM for cross-sample fusion.

Concretely, adding level 5 to feature fusion provides a huge improvement (8.6%), while further adding level 6 leads to a slight drop (0.1%). This is probably because the semantic misalignment in feature fusion at level 6 distracts the detector. Next, rows 4-8 show results of fusion with different levels of query features. From these, we observe a coarser query feature generally benefits the performance, with the one-to-all-scale corresponding strategy. This result is intuitive since coarser feature maps provide stronger positioning priors. In addition, more levels of query features also help, which is different from the corresponding-scale scheme. Then we include all three levels of query features and investigate how support features make a difference. Results in the last three lines show that the level 5 support feature obtains the best performance. We consider this in two folds. On one hand, it is likely due to a preference for this data distribution. On the other hand, as the intermediate one among 4, 5, and 6, this level is relatively more comparable with the whole feature pyramid.

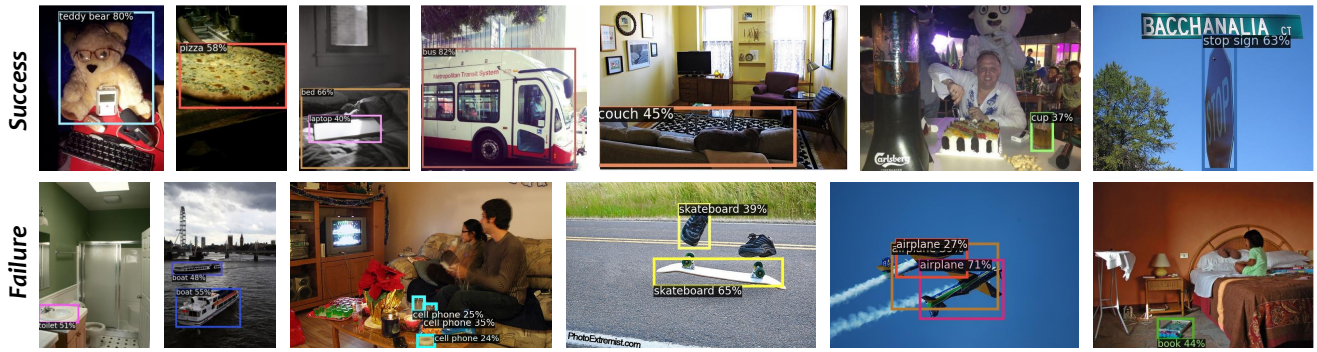

Figure 1. Success and failure cases of SaFT on COCO novel classes. We visualize bounding boxes with scores over 0.2.

Base / Set	VOC Base																	VOC Novel				
	Plant	Sofa	TV	Car	Bottle	Boat	Chair	Person	Bus	Train	Horse	Bike	Dog	Bird	Mbike	Table	Avg.	Cow	Sheep	Cat	Aero	Avg.
VOC	59.7	81.3	82.4	86.9	73.0	72.0	62.3	83.7	85.9	88.1	86.7	87.7	87.7	83.5	86.1	75.1	80.1	88.1	77.0	84.3	48.5	74.5
COCO	11.1	46.1	53.6	67.2	16.8	24.0	22.9	6.9	45.0	45.1	50.2	10.0	46.7	42.3	8.9	23.3	32.5	86.8	76.4	40.5	34.7	59.6

Table 2. Cross-domain comparison results on the VOC 2007 test set in terms of AP50 (%). The first column indicates the base training set, where VOC is trained on 16 VOC base classes and COCO on 60 COCO base classes non-overlapping with all VOC categories. Note that the first 16 categories (columns 2-17) are base classes for the VOC base training set while novel classes for COCO.

Cross-sample	# HA Blocks	# Parameters	Base	Novel
Reweighting	0	55.1M	72.8	64.2
Correlation	0	55.1M	77.7	64.3
HFM	2	61.6M	78.5	69.5
HFM	4	67.9M	79.5	71.7
HFM	6	74.2M	79.8	72.8

Table 3. Ablation study for the number of iterative HA blocks on VOC. All experiments employ VFM for cross-scale fusion and the one-to-all-scale scheme for cross-sample fusion.

B.2. Number of Iterative Fusion Blocks

From the perspective of performance and complexity, we compare our SaFT with different numbers of HA blocks in Tab. 3. Reweighting and correlation are presented in the first two rows as baselines with the same number of parameters, since their only difference is non-parameter pooling. Out of their 55.1M parameters, 53.5M are in the backbone, with the same below. We notice that while correlation beats reweighting on base classes by a large margin, their results on novel classes are very close. This indicates an overfitting tendency for convolution-based methods with large kernels. Comparing these baselines with no HA blocks to HFM with 2 HA blocks, base and novel AP50 improve by 0.8% $\sim$ 5.7% and 5.2% $\sim$ 5.3% respectively. These improvements demonstrate the effectiveness of attention-based HFM, with 6.5M more parameters. Also, from rows 3-5, performances on base classes grow by 1.3% and that on novel classes by 3.3%. This suggests more HA blocks provide more sufficient fusion, which leads to better results. But as the complexity increases linearly, performance growth gradually slows down.

C. Comparison Results

We evaluate cross-domain OSD performances following [2, 4, 7], which selects 60 categories in COCO14 non-overlapping with VOC as base classes and all 20 categories in VOC as novel classes. Other experimental settings are the same as in our main experiments. In the bottom line of Tab. 2, performances on different categories vary greatly. For instance, the model produces extremely low results for people and motorbikes. We attribute this to the feature extractor, lack of ability to highlight never-before-seen foregrounds. Although most classes experience a downswing compared with same-domain results, we notice that performances on cows and sheep are basically unchanged. Considering there are several base categories of animals, this suggests our offline SaFT is easier to adapt to a novel data distribution similar to base classes.

D. Qualitative Analysis

We provide extra qualitative visualizations of detected novel objects on COCO in Fig. 1. Success cases are presented in the upper row and failure ones in the lower row. The latter include false positives, e.g., the toilet, missing cases, e.g., the boat, and repeat detections, e.g., the airplane.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1

[2] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *ICCV*, 2019. 2

[3] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1

[4] Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. Defrcn: Decoupled faster r-cnn for few-shot object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8681–8690, 2021. 2

[5] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019. 1

[6] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *Proc. Int. Conf. Computer Vision (ICCV)*, 2019. 1

[7] Jiaxi Wu, Songtao Liu, Di Huang, and Yunhong Wang. Multi-scale positive sample refinement for few-shot object detection. In *European Conference on Computer Vision*, pages 456–472. Springer, 2020. 2