Supplementary Material for Multi-Faceted Distillation of Base-Novel Commonality for Few-shot Object Detection

Shuang Wu[2], Wenjie Pei[2,*], Dianwen Mei[2], Fanglin Chen[2], Jiandong Tian[3], and Guangming Lu[1,2,*]

[1] Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
[2] Harbin Institute of Technology, Shenzhen, China
[3] Chinese Academy of Sciences, Shenyang Institute of Automation
{wushuang9811, wenjiecoder}@outlook.com, {178mdw, linwers}@gmail.com, luguangm@hit.edu.cn, tianjd@sia.cn

1 More Implementation Details

Unlike previous works [2,3,4] that fine-tune the detector on a small balanced training set with K novel instances and K randomly sampled base instances, we utilize more base instances used in the first stage for fine-tuning, in that our approach requires abundant base instances stored in the memory bank to calculate the prototypes and distributions. Specifically, the training data for fine-tuning consists of two sets: base set with abundant instances and novel set with K instances per class. During each iteration, a batch is composed of two equally sized parts, one from the base set and another from the novel set. Then we update the memory bank by enqueuing the RoI features of instances in current batch to the corresponding class queue. The dimension of RoI features stored in the memory bank is 2048 for DeFRCN [2] baseline and 1024 for other three baselines (TFA [4], Retentive R-CNN [1], FSCE [3]). All other training settings (batch size, training iterations, learning rate, etc) are the same as that in corresponding baselines.

2 Performance for Base Classes

The proposed commonality distillation from base classes to novel classes allows leveraging the samples of base classes to train the object detector on the novel classes. Such commonality distillation pushes the model to fit the base samples to other classes (novel classes) with semantic similarities instead of their own groundtruth classes (base classes). Thus, it would not lead to the overfitting on the base classes. Table 1 shows that while the commonality distillation improves the performance of our model on novel classes substantially, it does not degrade the performance on base classes.

[*] Corresponding authors.

Table 1. Performance for base classes (bAP50) and novel classes (nAP50) on PASCAL VOC Novel Split 1.

Method / Shots	bAP50			nAP50		
	1	2	3	1	2	3
w/o distillation	78.4	**78.1**	76.8	57.0	58.6	64.3
w/ distillation	**78.5**	78.0	**78.3**	**63.4**	**66.3**	**67.7**

3 Additional Ablation Studies

Effect of varying the maximum queue size L in the memory bank. Fig 1 shows the performance as a function of maximum queue size L in the memory bank for different number of training shots. It can be seen that the performance improves initially as L increases because larger size of queue leads to more accurate estimation of the class prototypes. The performance reaches a plateau at $L = 2048$, which is selected for our method in other experiments.

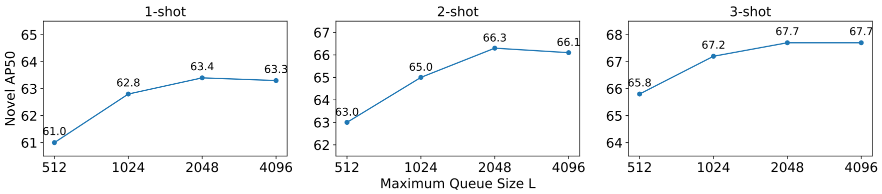

Fig. 1. Effect of varying the maximum queue size L in the memory bank.

Effect of varying the scaling factor. We explore the effect of different scaling factors α for computing similarity distribution and report the results in Table 2. It can be observed $\alpha = 5$ outperforms the other scaling factors for both recognition-related similarity and localization-related similarity. Therefore, we adopt $\alpha = 5$ in all of our experiments.

Table 2. Effect of varying the scaling factor for computing recognition-related and localization-related similarity on PASCAL VOC Novel Split 1.

(a) For recognition-related similarity.

α	nAP50		
	1-shot	2-shot	3-shot
1	58.4	60.9	62.2
3	60.8	62.5	65.9
5	**62.3**	**64.8**	**67.3**
10	61.3	64.5	66.9

(b) For localization-related similarity.

α	nAP50		
	1-shot	2-shot	3-shot
1	59.4	62.3	65.4
3	59.6	63.6	65.4
5	**59.9**	**64.1**	**65.7**
10	58.9	63.6	**65.7**

Effect of varying the loss weights. We conduct experiments to evaluate the effect of the hyper-parameters λ_c, λ_l and λ_d, which control the weight of each distillation loss. As shown in Table 3, we obtain the best results with $\lambda_c = 0.1$, $\lambda_l = 1.0$ and $\lambda_d = 0.1$, which are used for all other experiments.

Table 3. Effect of varying the weight of each distillation loss on PASCAL VOC Novel Split 1.

(a) Parameter: λ_c.

λ_c	nAP50		
	1-shot	2-shot	3-shot
0.001	59.2	63.0	64.9
0.01	59.3	62.8	65.9
0.1	**62.3**	**64.8**	**67.3**
1.0	60.7	62.2	63.3

(b) Parameter: λ_l.

λ_l	nAP50		
	1-shot	2-shot	3-shot
0.01	58.1	62.0	64.7
0.1	57.6	63.9	65.0
1.0	**59.9**	**64.1**	**65.7**
2.0	59.7	62.6	64.7

(c) Parameter: λ_d.

λ_d	nAP50		
	1-shot	2-shot	3-shot
0.001	57.9	63.2	65.9
0.01	60.5	63.4	65.9
0.1	**62.6**	**65.1**	**66.2**
1.0	58.0	58.7	63.2

Hyper-parameters for distribution commonalities. We study the hyper-parameters, i.e., k and $|\mathbb{S}_c|$ adopted in distribution commonalities. k is the number of the closest base classes to novel class c for transferring the variance. $|\mathbb{S}_c|$ is the number of instances sampled from the calibrated distribution for novel class c during each iteration. As shown in Table 4, these two hyper-parameters have a mild impact on the performance, and we observe that $k = 2$ and $|\mathbb{S}_c| = 10$ work best for nAP50.

Table 4. Ablation study for distribution commonalities. Results (nAP50) are reported on 1-shot of PASCAL VOC Novel Split 1.

Number of the Closest Base Classes	$\|\mathbb{S}_c\|$			
	1	5	10	20
$k = 1$	61.3	61.1	61.6	62.4
$k = 2$	61.4	61.8	**62.6**	61.6
$k = 3$	62.3	62.3	61.8	60.0

'Variance' vs 'mean' & 'variance' for distribution commonality. In contrast to Distribution Calibration [5] transferring both the mean and variance from base classes to novel classes, our method only distills the variance as the distribution commonalities to avoid the distributional overlapping between base and novel classes. We conduct experiments to compare such two mechanisms. The results in Table 5 show that transferring both the mean and variance degrades the performance by a large margin than transferring only variance, and performs even worse than the baseline without commonality distillation.

Table 5. Effect of transfer 'mean' for distribution commonality.

Dist	nAP50		
	1-shot	2-shot	3-shot
Baseline	58.5	62.6	65.4
Mean & variance	57.5	62.4	64.1
Variance	**62.6**	**65.1**	**66.2**

4 Results over Multiple Runs

We report the few-shot object detection results (nAP50) over 10 random runs on PASCAL VOC Novel Split 1 in Table 6. It can be observed that our method outperforms the baseline (DeFRCN) under all settings, which shows the effectiveness of our method.

Table 6. Results (nAP50) over 10 random runs on VOC Novel Split 1.

Method / Shots	nAP50				
	1	2	3	5	10
DeFRCN	43.8	57.5	61.4	65.3	67.0
Ours	**53.7**	**64.3**	**66.6**	**69.6**	**70.4**

5 More Qualitative Visualizations

In this section, we provide more qualitative visualizations on PASCAL VOC and MS COCO datasets. As shown in Figure 2, our approach could rescue various error cases, including missing detections, misclassifications and imprecise localizations.

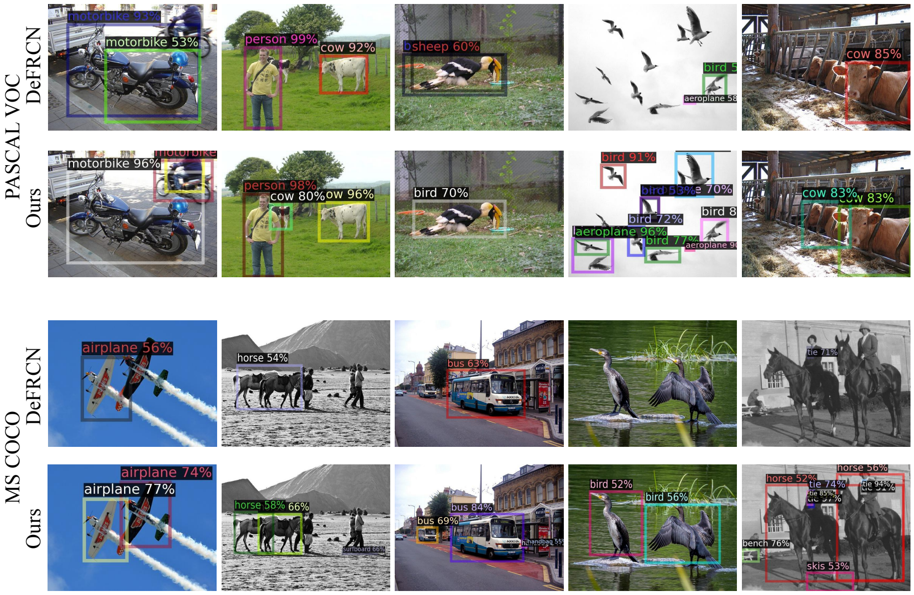

Fig. 2. The visualization results of DeFRCN and our approach under 1-shot setting of PASCAL VOC Novel Split 1, and under 1-shot setting of MS COCO.

References

1. Fan, Z., Ma, Y., Li, Z., Sun, J.: Generalized few-shot object detection without forgetting. In: CVPR (2021)
2. Qiao, L., Zhao, Y., Li, Z., Qiu, X., Wu, J., Zhang, C.: Defrcn: Decoupled faster r-cnn for few-shot object detection. In: ICCV (2021)
3. Sun, B., Li, B., Cai, S., Yuan, Y., Zhang, C.: Fsce: few-shot object detection via contrastive proposal encoding. In: CVPR (2021)
4. Wang, X., Huang, T., Gonzalez, J., Darrell, T., Yu, F.: Frustratingly simple few-shot object detection. In: ICML (2020)
5. Yang, S., Liu, L., Xu, M.: Free lunch for few-shot learning: Distribution calibration. In: ICLR (2020)