Supplementary Material

1. Analytical Proof of CFA

In this section, we mathematically derive the update rules for the proposed CFA method. We formulate our objective function as follows:

$$\begin{aligned} \text{minimize}_{\tilde{g}_n, \tilde{g}_b} \quad & \frac{1}{2}||g_n - \tilde{g}_n||_2^2 + \frac{1}{2}||g_b - \tilde{g}_b||_2^2 \\ \text{subject to} \quad & \tilde{g}_n^\top g_b \geq 0, \\ & \tilde{g}_b^\top g_n \geq 0, \end{aligned} \tag{1}$$

where g_n and g_b represents the proposed gradient update for the novel and base task, respectively. $\tilde{g}_n$ and $\tilde{g}_b$ denotes the projected gradient update for the novel and base task, respectively. If both constraints are satisfied, the update rule will be the average of g_n and g_b. Otherwise, we solve the constrained optimization problem using the method of Lagrange multipliers.

First, we reformulate the problem in the standard form as follows:

$$\begin{aligned} \text{minimize}_{z_n, z_b} \quad & \frac{1}{2}z_n^\top z_n - g_n^\top z_n + \frac{1}{2}z_b^\top z_b - g_b^\top z_b \\ \text{subject to} \quad & -z_n^\top g_b \leq 0, \\ & -z_b^\top g_n \leq 0, \end{aligned} \tag{2}$$

where $\tilde{g}_b$ and $\tilde{g}_n$ are denoted as z_b and z_n, respectively. We ignore the constant terms $g_n^\top g_n$ and $g_b^\top g_b$. In addition, the sign of the inequality constraints is changed. Then, the Lagrangian can be formulated as:

$$\begin{aligned} \mathcal{L}(z_n, z_b, \alpha_1, \alpha_2) =& \frac{1}{2}z_n^\top z_n - g_n^\top z_n - \alpha_1 z_n^\top g_b \\ &+ \frac{1}{2}z_b^\top z_b - g_b^\top z_b - \alpha_2 z_b^\top g_n, \end{aligned} \tag{3}$$

where α_1 and α_2 are the dual variables. To find the solution of the primal variables z_n^* and z_b^*, we need to find the lower bound solution of the primal problem by computing the solution of the dual problem:

$$\theta_{\mathcal{D}}(\alpha_1, \alpha_2) = \min_{z_n, z_b} \mathcal{L}(z_n, z_b, \alpha_1, \alpha_2). \tag{4}$$

We find z_n^* and z_b^* as a function of dual variables α_1 and α_1, respectively, by minimizing the Lagrangian $\mathcal{L}(z_n, z_b, \alpha_1, \alpha_2)$. This is achieved by setting its derivatives w.r.t z_n and z_b to zero,

$$\begin{aligned} \nabla_{z_n}\mathcal{L}(z_n, z_b, \alpha_1, \alpha_2) = 0, \\ z_n^* = g_n + \alpha_1 g_b, \end{aligned} \tag{5}$$

$$\begin{aligned} \nabla_{z_b}\mathcal{L}(z_n, z_b, \alpha_1, \alpha_2) = 0, \\ z_b^* = g_b + \alpha_2 g_n. \end{aligned} \tag{6}$$

Next, we can find the solution of the primal variables by solving the dual problem. We substitute Eq. (5) and Eq. (6) in Eq. (4). Now, the dual problem can be rewritten as:

$$\begin{aligned} \theta_{\mathcal{D}}(\alpha_1, \alpha_2) =& \frac{1}{2}(g_n^\top g_n + 2\alpha_1 g_n^\top g_b + \alpha_1^2 g_b^\top g_b) \\ &- g_n^\top g_n - 2\alpha_1 g_n^\top g_b - \alpha_1^2 g_b^\top g_b \\ &+ \frac{1}{2}(g_b^\top g_b + 2\alpha_2 g_b^\top g_n + \alpha_2^2 g_n^\top g_n) \\ &- g_b^\top g_b - 2\alpha_2 g_b^\top g_n - \alpha_2^2 g_n^\top g_n \\ =& -\frac{1}{2}g_n^\top g_n - \alpha_1 g_n^\top g_b - \frac{1}{2}\alpha_1^2 g_b^\top g_b \\ &- \frac{1}{2}g_b^\top g_b - \alpha_2 g_b^\top g_n - \frac{1}{2}\alpha_2^2 g_n^\top g_n. \end{aligned}$$

Next, we find the solution α_1^* and α_2^* of dual problem as follows:

$$\begin{aligned} \nabla_{\alpha_1}\theta_{\mathcal{D}}(\alpha_1, \alpha_2) = 0, \\ \alpha_1^* = -\frac{g_n^\top g_b}{g_b^\top g_b}, \end{aligned} \tag{7}$$

$$\begin{aligned} \nabla_{\alpha_2}\theta_{\mathcal{D}}(\alpha_1, \alpha_2) = 0, \\ \alpha_2^* = -\frac{g_b^\top g_n}{g_n^\top g_n}. \end{aligned} \tag{8}$$

Given the solutions of the dual problem, we can find closed form solutions of $\tilde{g}_n$ and $\tilde{g}_b$ by substituting the dual solutions α_1^* Eq. (7) and α_2^* Eq. (8) in Eq. (5) and Eq. (6), respectively:

$$z_n^* = g_n - \frac{g_n^\top g_b}{g_b^\top g_b} g_b = \tilde{g}_n, \tag{9}$$

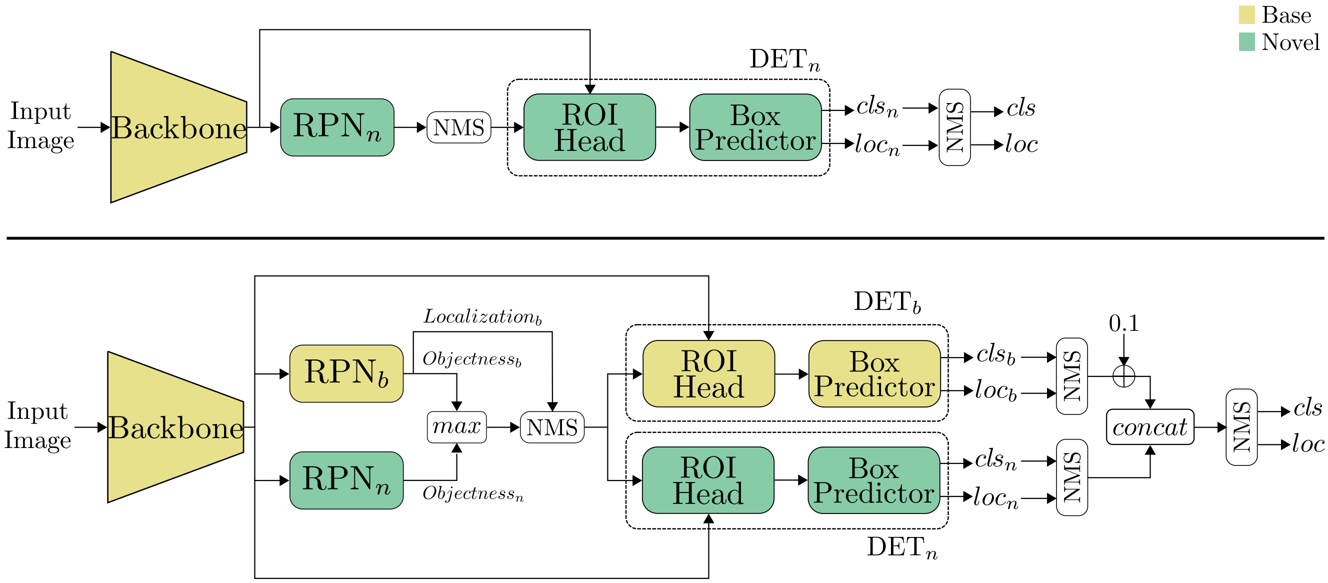

Figure 1. **Top:** Illustration of the single model inference. **Bottom:** A detailed overview of the ensemble model evaluation protocol proposed by Retentive R-CNN [1].

Methods / Shots	5 shot AP	5 shot AP50	5 shot AP75	10 shot AP	10 shot AP50	10 shot AP75	30 shot AP	30 shot AP50	30 shot AP75
FRCN-ft-full [2] ‡ §	4.6	8.7	4.4	5.5	10.0	5.5	7.4	13.1	7.4
Meta-YOLO [3]	-	-	-	5.6	12.3	4.6	9.1	19.0	7.6
Meta R-CNN [4]	-	-	-	8.7	19.1	6.6	12.4	25.3	10.8
TFA w/ cos [5] ‡ §	7.0	13.3	6.5	9.1	17.1	8.8	12.1	22.0	12.0
Meta Det [6]	-	-	-	7.1	14.6	6.1	11.3	21.7	8.1
FSOD [7]	-	-	-	12.0	22.4	11.8	-	-	-
FsDetView [8] §	10.7	24.5	6.7	12.5	27.3	9.8	14.7	30.6	12.2
MPSR [9] ‡	7.4	12.3	7.7	9.8	17.9	9.7	14.1	25.4	14.2
FSCE [10] ‡ §	-	-	-	11.1	-	9.8	15.3	-	14.2
CME [11] ‡	-	-	-	15.1	24.6	16.4	16.9	28.0	17.8
Deformable-DETR-ft-full [12] §	-	-	-	11.7	19.6	12.1	16.3	27.2	16.7
DeFRCN [13]	15.5	**29.4**	14.2	18.3	33.7	17.4	22.6	39.8	22.8
CFA-DeFRCN (Ours)	**15.6**	29.1	**15.2**	**19.1**	**34.8**	**18.7**	**23.0**	**40.5**	**23.0**

Table 1. Few-shot detection performance on MS-COCO for the novel categories. ‡ indicates methods using multi-scale features. § indicates results averaged on multiple runs.

$$z_b^* = g_b - \frac{g_b^\top g_n}{g_n^\top g_n} g_n = \tilde{g}_b. \tag{10}$$

After finding the closed form solution, a single update rule can be realized as:

$$\tilde{g} = \frac{\tilde{g}_n + \tilde{g}_b}{2}. \tag{11}$$

2. Evaluation Protocols

In Fig. 1, the utilized evaluation protocols are presented. The single model inference comprises the RPN$_n$ and DET$_n$, finetuned with a few-shots from the novel data, while the backbone is kept frozen. The evaluation is conducted as follows: (1) the image is fed to the backbone (2) the RPN$_n$ generates proposals (3) the proposals with IoU scores lower than a predefined threshold are omitted via a non-maximum suppression (NMS) (4) the DET$_n$ outputs both the classification logits cls_n and bounding boxes loc_n, respectively (5) finally, the final predictions are filtered via a NMS.

On the other hand, the ensemble inference model further employs the RPN$_b$ and DET$_b$ from the base model. The inference is done as follows: (1) the image is fed to the backbone (2) the image features are fed to both the RPN$_b$ and RPN$_n$ to compute the objectness logits O_b and O_n, respec-

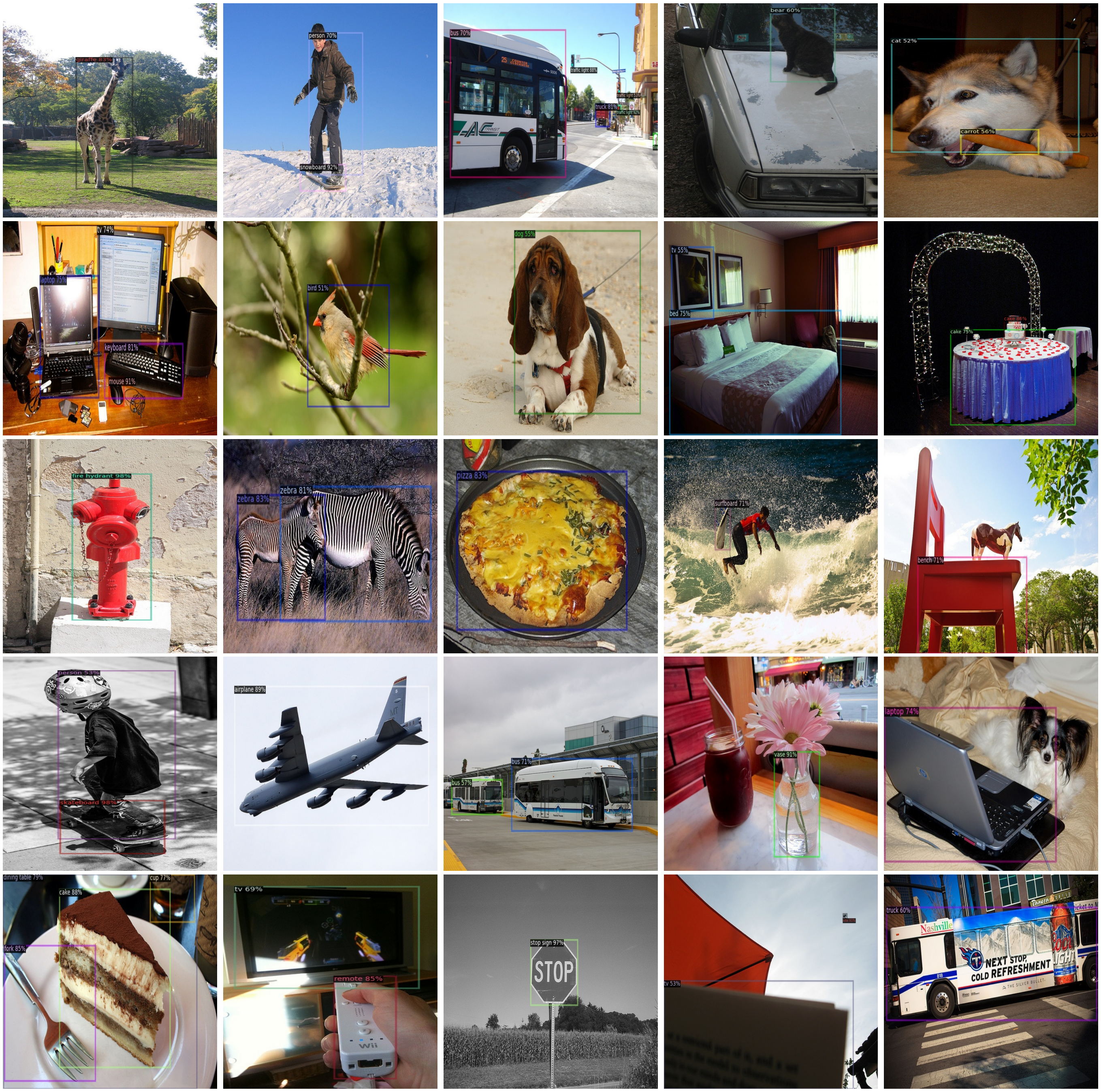

Figure 2. Qualitative analysis of the proposed CFA method on the MS-COCO dataset. The shown results are based on CFA w/cos finetuned under 30-shot setting. The first three columns show success scenarios while the last two columns present the failure scenarios.

tively (3) the maximum between O_b and O_n is fed to NMS along with the bounding boxes from RPN$_b$ (4) the filtered proposals are then fed to both DET$_b$ and DET$_n$ to output the classification logits and bounding boxes (5) after the detectors' predictions are fed separately to a NMS, a bonus of 0.1 are added to cls_b (6) finally the output from both detectors are concatenated and fed to a NMS to output the final predictions. We emphasize that we did not use the ensemble models during finetuning (as in Retentive-RCNN [1]), but rather we finetuned a single model and used both the base and finetuned models during inference.

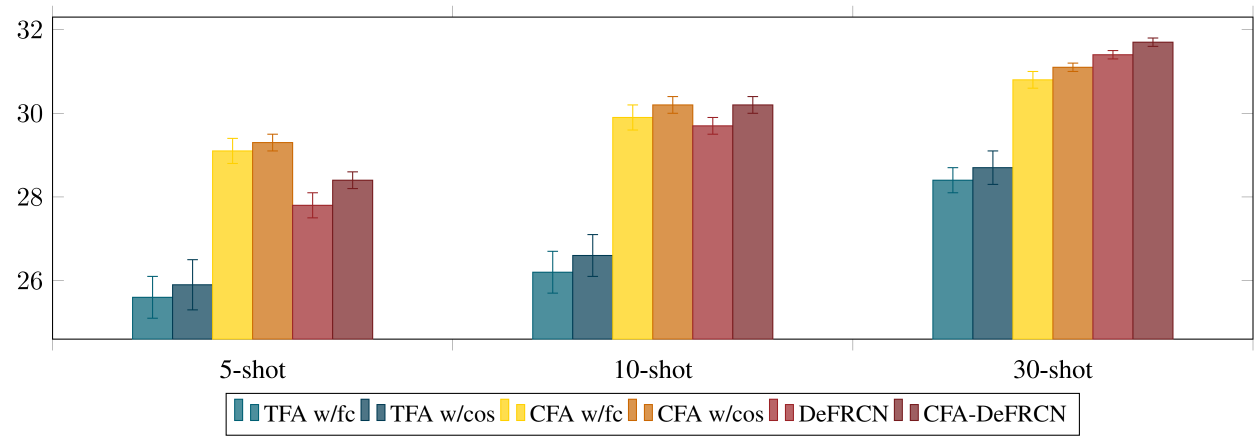

Figure 3. Results over 10 random runs on MS-COCO under $K = 5, 10, 30$-shot setting. The mean and 95% confidence interval are reported.

Methods / Shots	5 shot			10 shot			30 shot		
	AP	**bAP**	**nAP**	**AP**	**bAP**	**nAP**	**AP**	**bAP**	**nAP**
TFA w/ fc [5]	25.6±0.5	31.8±0.5	6.9±0.7	26.2±0.5	32.0±0.5	9.1±0.5	28.4±0.3	33.8±0.3	12.0±0.4
TFA w/ cos [5]	25.9±0.6	32.3±0.6	7.0±0.7	26.6±0.5	32.4±0.6	9.1±0.5	28.7±0.4	34.2±0.4	12.1±0.4
CFA w/ fc	29.1±0.3	36.2±0.3	7.7±0.6	29.9±0.3	36.7±0.2	9.6±0.6	30.8±0.2	36.6±0.2	13.6±0.3
CFA w/ cos	29.3±0.2	36.0±0.2	9.2±0.5	30.2±0.2	36.6±0.1	11.2±0.5	31.1±0.1	36.6±0.1	14.8±0.2
DeFRCN [13]	27.8±0.3	32.6±0.3	13.6±0.7	29.7±0.2	34.0±0.2	16.8±0.6	31.4±0.1	34.8±0.1	21.2±0.4
CFA-DeFRCN	28.4±0.2	32.8±0.2	15.2±0.5	30.2±0.2	34.0±0.2	18.8±0.4	31.7±0.1	34.6±0.1	23.0±0.3

Table 2. G-FSOD experimental results for 5,10,30-shot settings on MS-COCO. We report AP, bAP, nAP for all, base, and novel classes, respectively.

3. Qualitative Results

In Fig. 2, we present qualitative results on CFA w/cos finetuned with 30-shot setting. The first three columns show various success scenarios while the last two columns show different failure cases. Compared to base classes, the model is less confident with novel categories. This can be attributed to learning indiscriminative features, hence resulting in false positives and false negatives.

4. Additional Experiments

Comparison against FSOD baselines. To further investigate the impact of CFA on the novel classes, we compare the performace of CFA-finetuned models (CFA w/fc, CFA w/cos and CFA-DeFRCN) with FSOD models on the challenging MS-COCO benchmark. CFA-DeFRCN outperforms existing approaches on the novel AP metric, although it was trained in a G-FSOD setting which generally leads to lower performance on the novel classes. The results are shown in Tab. 1.

Multiple runs. We run the CFA-finetuned models (CFA w/fc, CFA w/cos, and CFA-DeFRCN) using 10 different seeds on MS-COCO and compare with the baselines (TFA [5] and DeFRCN [1]). The results are shown in Tab. 2 and Fig. 3. We use the same random seeds as TFA [5] and DeFRCN [13]. CFA consistently improves the overall AP while displaying a narrower confidence interval.

Method	w/E	Inference Time (ms)	Model Capacity
TFA w/ fc	✗	85	60.6M
TFA w/ cos	✗	87	60.6M
CFA w/ fc	✗	85	60.6M
CFA w/ cos	✗	86	60.6M
CFA-DeFRCN	✗	147	52.7M
CFA w/ fc	✓	211	75.4M
CFA w/ cos	✓	211	75.4M
CFA-DeFRCN	✓	376	105.3M

Table 3. Inference time and model capacity for different evaluation protocols. Ensemble methods have a significant overhead compared to single model. w/E denotes whether the ensemble method is employed.

5. Further Ablation Experiments

Inference time and model capacity. We study the impact of the two evaluation methods (single model vs ensemble model) on the inference time and number of parameters during inference. Although ensemble model evaluation achieves less forgetting, the inference time increases in av-

Model	Backone	RPN	RoI Head	AP	bAP	nAP
TFA w/ fc				27.9	33.9	10.0
		✓		29.9	37.2	7.9
			✓	28.9	35.4	9.6
		✓	✓	28.9	35.1	10.2
	✓	✓	✓	24.1	29.0	9.1
CFA w/ fc				29.6	36.0	10.4
		✓		30.3	37.4	9.3
			✓	**30.8**	**37.8**	9.6
		✓	✓	**30.8**	37.6	**10.5**
	✓	✓	✓	23.9	28.6	10.1
TFA w/ cos				28.7	35.0	10.0
		✓		28.9	35.8	8.3
			✓	29.0	35.3	10.3
		✓	✓	29.2	35.2	11.2
	✓	✓	✓	24.1	28.5	10.9
CFA w/ cos				29.4	35.9	9.8
		✓		28.7	35.3	8.9
			✓	30.2	**36.8**	10.6
		✓	✓	**30.3**	36.6	**11.3**
	✓	✓	✓	23.6	27.9	10.9

Table 4. Effect of unfreezing different components of our detection model in comparison to TFA [5]. ✓ denotes unfreezing a component. The results are reported for MS-COCO under 10-shots.

Model	Base-Shots	AP	bAP	nAP
TFA w/ fc	1-Shots	22.2	26.3	9.8
	2-Shots	24.8	29.8	9.9
	3-Shots	26.1	31.5	10.1
	5-Shots	27.0	32.6	10.2
	10-Shots	27.9	33.9	10.0
CFA w/ fc	1-Shots	28.8	34.9	**10.5**
	2-Shots	30.0	36.5	**10.5**
	3-Shots	30.3	37.0	10.3
	5-Shots	30.5	37.2	10.4
	10-Shots	**30.8**	**37.6**	**10.5**
TFA w/ cos	1-Shots	24.2	28.9	10.0
	2-Shots	26.5	32.0	10.2
	3-Shots	27.2	32.9	10.3
	5-Shots	27.8	33.6	10.3
	10-Shots	28.7	35.0	10.0
CFA w/ cos	1-Shots	28.6	34.3	**11.5**
	2-Shots	29.8	35.9	11.3
	3-Shots	30.0	36.2	11.3
	5-Shots	30.2	36.4	11.3
	10-Shots	**30.3**	**36.6**	11.3

Table 5. Impact of variable number of base shots on the catastrophic forgetting of base classes. We compare CFA against TFA [5]. The experiments are conducted on MS-COCO dataset given 10-shots of the novel categories.

erage by 52%. On the other hand, the number of parameters increases by 50% in CFA w/fc (and w/cos) and by 102% in CFA-DeFRCN, since DeFRCN unfreezes the backbone during finetuning.

Extended ablation study. We extend the ablation studies conducted on TFA w/fc and CFA w/fc to include the TFA w/cos and CFA w/cos. The results are presented in Tab. 4 and Tab. 5.

References

[1] Zhibo Fan, Yuchen Ma, Zeming Li, and Jian Sun. Generalized few-shot object detection without forgetting. In IEEE Conference on Computer Vision and Pattern Recognition, pages 4527–4536, 2021. 2, 3, 4

[2] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems, pages 91–99, 2015. 2

[3] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In IEEE International Conference on Computer Vision, pages 8419–8428, 2018. 2

[4] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta R-CNN: Towards general solver for instance-level low-shot learning. In IEEE International Conference on Computer Vision, pages 9577–9586, 2019. 2

[5] Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E. Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. In International Conference on Machine Learning, pages 9919–9928, 2020. 2, 4, 5

[6] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Meta-learning to detect rare objects. In IEEE International Conference on Computer Vision, pages 9924–9933, 2019. 2

[7] Qi Fan, Wei Zhuo, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In IEEE Conference on Computer Vision and Pattern Recognition, pages 4012–4021, 2020. 2

[8] Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. In European Conference on Computer Vision, pages 192–210, 2020. 2

[9] 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, 2020. 2

[10] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. FSCE: few-shot object detection via contrastive proposal encoding. In IEEE Conference on Computer Vision and Pattern Recognition, pages 7352–7362, 2021. 2

[11] Bohao Li, Boyu Yang, Chang Liu, Feng Liu, Rongrong Ji, and Qixiang Ye. Beyond Max-Margin: Class margin equilibrium for few-shot object detection. In IEEE Conference on Computer Vision and Pattern Recognition, pages 7359–7368, 2021. 2

[12] 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, 2020. 2

[13] Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. DeFRCN: Decoupled faster R-CNN for few-shot object detection. In IEEE International Conference on Computer Vision, 2021. 2, 4