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:

\[
\begin{align*}
\text{minimize}_{\tilde{g}_n, \tilde{g}_b} & \quad \frac{1}{2} \| g_n - \tilde{g}_n \|^2 + \frac{1}{2} \| g_b - \tilde{g}_b \|^2 \\
\text{subject to} & \quad \tilde{g}_n^\top g_n \geq 0, \\
& \quad \tilde{g}_b^\top g_n \geq 0,
\end{align*}
\]

(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 is followed. 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:

\[
\begin{align*}
\text{minimize}_{z_n, z_b} & \quad \frac{1}{2} z_n^\top z_n - g_n^\top z_n + \frac{1}{2} z_b^\top z_b - g_b^\top z_b \\
\text{subject to} & \quad -z_n^\top g_n \leq 0, \\
& \quad -z_b^\top g_n \leq 0,
\end{align*}
\]

(2)

where \( \tilde{g}_n \) and \( \tilde{g}_b \) are denoted as \( z_n \) and \( z_b \), 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:

\[
\begin{align*}
\mathcal{L}(z_n, z_b, \alpha_1, \alpha_2) &= \frac{1}{2} z_n^\top z_n - g_n^\top z_n - \alpha_1 z_n^\top g_b \\
& \quad + \frac{1}{2} z_b^\top z_b - g_b^\top z_b - \alpha_2 z_b^\top g_n,
\end{align*}
\]

(3)

where \( \alpha_1 \) and \( \alpha_2 \) 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:

\[
\theta_D(\alpha_1, \alpha_2) = \min_{z_n, z_b} \mathcal{L}(z_n, z_b, \alpha_1, \alpha_2).
\]

(4)

We find \( z_n^* \) and \( z_b^* \) as a function of dual variables \( \alpha_1 \) and \( \alpha_1 \), 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,

\[
\nabla_{z_n} \mathcal{L}(z_n, z_b, \alpha_1, \alpha_2) = 0,
\]

\[
z_n^* = g_n + \alpha_1 g_b,
\]

(5)

\[
\nabla_{z_b} \mathcal{L}(z_n, z_b, \alpha_1, \alpha_2) = 0,
\]

\[
z_b^* = g_b + \alpha_2 g_n.
\]

(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:

\[
\theta_D(\alpha_1, \alpha_2) = \frac{1}{2} (g_n^\top g_n + 2 \alpha_1 g_n^\top g_b + \alpha_2^2 g_b^\top g_b) \\
- g_n^\top g_n - 2 \alpha_1 g_n^\top g_b - \alpha_2^2 g_b^\top g_b \\
+ \frac{1}{2} (g_b^\top g_b + 2 \alpha_2 g_b^\top g_n + \alpha_2^2 g_n^\top g_n) \\
- g_b^\top g_b - 2 \alpha_2 g_b^\top g_n - \alpha_2^2 g_n^\top g_n \\
= - \frac{1}{2} g_n^\top g_n - \alpha_1 g_n^\top g_b - \frac{1}{2} \alpha_1^2 g_b^\top g_b \\
- \frac{1}{2} g_b^\top g_b - \alpha_2 g_b^\top g_n - \frac{1}{2} \alpha_2^2 g_n^\top g_n.
\]

Next, we find the solution \( \alpha_1^* \) and \( \alpha_2^* \) of dual problem as follows:

\[
\nabla_{\alpha_1} \theta_D(\alpha_1, \alpha_2) = 0,
\]

\[
\alpha_1^* = \frac{g_n^\top g_b}{g_b^\top g_b},
\]

(7)

\[
\nabla_{\alpha_2} \theta_D(\alpha_1, \alpha_2) = 0,
\]

\[
\alpha_2^* = \frac{g_b^\top g_n}{g_n^\top g_n},
\]

(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 \( \alpha_1^* \) Eq. (7) and \( \alpha_2^* \) Eq. (8) in Eq. (5) and Eq. (6), respectively:

\[
z_n^* = g_n - \frac{g_n^\top g_b}{g_b^\top g_b} g_b = \tilde{g}_n,
\]

(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].

### Table 1. Few-shot detection performance on MS-COCO for the novel categories.

<table>
<thead>
<tr>
<th>Methods / Shots</th>
<th>5 shot</th>
<th>10 shot</th>
<th>30 shot</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AP</td>
<td>AP50</td>
<td>AP75</td>
</tr>
<tr>
<td>FRCN-ft-full [2] ‡§</td>
<td>4.6</td>
<td>8.7</td>
<td>4.4</td>
</tr>
<tr>
<td>Meta-YOLO [3]</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Meta R-CNN [4]</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TFA w/ cos [5] ‡§</td>
<td>7.0</td>
<td>13.3</td>
<td>6.5</td>
</tr>
<tr>
<td>Meta Det [6]</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FSOD [7]</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FsDetView [8] ‡§</td>
<td>10.7</td>
<td>24.5</td>
<td>6.7</td>
</tr>
<tr>
<td>MPSR [9] ‡</td>
<td>7.4</td>
<td>12.3</td>
<td>7.7</td>
</tr>
<tr>
<td>FSCE [10] ‡§</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CME [11] ‡</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Deformable-DETR-ft-full [12] §</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DeFRCN [13]</td>
<td>15.5</td>
<td>29.4</td>
<td>14.2</td>
</tr>
<tr>
<td>CFA-DeFRCN (Ours)</td>
<td>15.6</td>
<td>29.1</td>
<td>15.2</td>
</tr>
</tbody>
</table>

‡ indicates methods using multi-scale features. § indicates results averaged on multiple runs.

\[
z_b^* = g_b - \frac{g_n^T g_n}{g_n^T g_n} g_n = \tilde{g}_b .\tag{10}
\]

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.

1. \(O_b\) and \(O_n\) are fed to NMS along with the bounding boxes from RPN\(_b\).
2. The filtered proposals are then fed to both DET\(_b\) and DET\(_n\) to output the classification logits and bounding boxes.
3. After the detectors’ predictions are fed separately to a NMS, a bonus of 0.1 are added to \(cls_b\).
4. 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.

<table>
<thead>
<tr>
<th>Methods / Shots</th>
<th>AP bAP nAP</th>
<th>AP bAP nAP</th>
<th>AP bAP nAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>TFA w/ fc [5]</td>
<td>25.6±0.5</td>
<td>31.8±0.5</td>
<td>6.9±0.7</td>
</tr>
<tr>
<td>TFA w/ cos [5]</td>
<td>25.9±0.6</td>
<td>32.3±0.6</td>
<td>7.0±0.7</td>
</tr>
<tr>
<td>CFA w/ fc</td>
<td>29.1±0.3</td>
<td>36.2±0.3</td>
<td>7.7±0.6</td>
</tr>
<tr>
<td>CFA w/ cos</td>
<td>29.3±0.2</td>
<td>36.0±0.2</td>
<td>9.2±0.5</td>
</tr>
<tr>
<td>DeFRCN [13]</td>
<td>27.8±0.3</td>
<td>32.6±0.3</td>
<td>13.6±0.7</td>
</tr>
<tr>
<td>CFA-DeFRCN</td>
<td>28.4±0.2</td>
<td>32.8±0.2</td>
<td>15.2±0.5</td>
</tr>
</tbody>
</table>

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 performance 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.

5. Further Ablation Experiments

<table>
<thead>
<tr>
<th>Method</th>
<th>w/E</th>
<th>Inference Time (ms)</th>
<th>Model Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>TFA w/ fc</td>
<td>✗</td>
<td>85</td>
<td>60.6M</td>
</tr>
<tr>
<td>TFA w/ cos</td>
<td>✗</td>
<td>87</td>
<td>60.6M</td>
</tr>
<tr>
<td>CFA w/ fc</td>
<td>✗</td>
<td>85</td>
<td>60.6M</td>
</tr>
<tr>
<td>CFA w/ cos</td>
<td>✗</td>
<td>86</td>
<td>60.6M</td>
</tr>
<tr>
<td>CFA-DeFRCN</td>
<td>✗</td>
<td>147</td>
<td>52.7M</td>
</tr>
<tr>
<td>CFA w/ fc</td>
<td>✓</td>
<td>211</td>
<td>75.4M</td>
</tr>
<tr>
<td>CFA w/ cos</td>
<td>✓</td>
<td>211</td>
<td>75.4M</td>
</tr>
<tr>
<td>CFA-DeFRCN</td>
<td>✓</td>
<td>376</td>
<td>105.3M</td>
</tr>
</tbody>
</table>

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.
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.

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.

References

12. Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-