Supplementary Material

Figure 1: Randomly selected detection results of a pretrained base class detector. Ground truths for novel classes are bounded with black boxes, while detected instances of the base detector are bounded with green boxes and labeled with predicted classes. The base detector's ability to reject novel class objects, even with great salience to human, is apparent and phenomenal.

1. Implementation Details

Retentive R-CNN. As a transfer learning based method, Retentive R-CNN is trained in two steps: the first step is trained on $\mathcal{D}^b$, which follows the same hyperparameters and learning schedule as in TFA[2]; the second step is trained on a balanced dataset of both $\mathcal{C}^b \cup \mathcal{C}^n$. During the finetuning stage, we set learning rate to 0.05, coefficient for consistency loss to 0.1 across all settings, and only the finetuned RPN objectness is used. Note that the model is trained until full convergence; thus, the learning schedule for finetuning may vary from different datasplits. During inference, the base detector's classification logits are padded with 0 on the novel class entries; then, softmax operation is conducted on the padded logits to produce classification scores. As all activation in the network is ReLU and the base detector utilizes an fc classifier, logits with zero value can make good prior probabilities for novel classes, thus balance the scale of scores as the number of class entries are less than the novel detector. This improves base class AP and overall AP, *e.g.*, overall AP increases from 32.0 to 32.1 under

(a)

(b)

Figure 2: (a) Visualization of Retentive R-CNN and TFA w/cos[2] under Pascal VOC split1 2-shot settings. (b) Typical failure cases of Retentive R-CNN.

MS-COCO 10-shot setting. The novel detector also predicts base class probabilities, so we include these predictions for the non-maximum suppression procedure as well. Though consistency loss enhances the similarity between the prediction of the base detector and novel detector on base classes, the novel detector's base class predictions show ensembling effect to a certain extent, improving 0.05-0.1 base class AP upon base class AP of the pretrained model.

Meta R-CNN[3] & FsDetView[1]. These two meta-learning methods are originally finetuned on randomly selected samples; we fix the samples to be the same as ours for fintuning for a fair comparison. Note that in both works, they finetune with base class samples as much as three times more than novel class samples to maintain base class performance, while we use the same number of samples to make a fair comparison. As Meta R-CNN[3] does not provide code for training on MS-COCO in the published implementation, we train Meta R-CNN[3] with identical hyperparameters and settings as FsDetView[1] on MS-COCO, which is implemented on the top of Meta R-CNN[3].

2. Examples for the Base Detector Rejecting Novel Class Instances

Here we show more detection results from the pretrained base model in Figure1 to better demonstrate the property that the pretrained detector can reject novel class instances even if they are of great saliency to humans. The images are randomly selected from the first 100 images ordered by image id of MS-COCO 2014 minival set without cherry-picking. We bound the unrecognized novel class instances with black boxes and the detected objects with green boxes and their corresponding predicted category. Obviously, the base detector has a strong ability to ignore novel classes, thus false positives seldom occur from the base detector when encountering unseen classes. This property is utilized in Retentive R-CNN to maintain base class performance.

3. More Detection Results & Failure Case Analysis

In this section, we show some extra detection results for further demonstration of the effectiveness of our method and a qualitative failure case analysis. Figure2(a) shows representative results for comparing our method and TFA w/cos[2] under Pascal VOC split1 2-shot setting. The conclusion is consistent with the qualitative comparison shown in the main paper that our method typically performs better on base classes due to the non-forgetting property and reduces object confusion on novel class instances, successfully detecting many of the ignored objects by TFA w/cos[2]. Some extra detection visualization of our method is shown in Figure3.

Nevertheless, both our method and previous works have a vast metrics gap between few-shot classes and classes trained from abundant data, indicating that few-shot object

Figure 3: Extra visualization of the detection results from Retentive R-CNN. The first row shows results under MS-COCO 10-shot settings while the second row shows results under Pascal VOC split1 2-shot settings.

Methods	5 shot			10 shot			30 shot		
	AP	**bAP**	**nAP**	**AP**	**bAP**	**nAP**	**AP**	**bAP**	**nAP**
Retentive R-CNN	**31.4**	**39.3**	**7.7**	**31.8**	**39.2**	**9.5**	**32.6**	**39.3**	**12.4**
FRCN-ft-full[2]	14.4	17.6	4.6	13.4	16.1	5.5	13.5	15.6	7.4
TFA w/ fc[2]	25.6	31.8	6.9	26.2	32.0	9.1	28.4	33.8	12.0
TFA w/ cos[2]	25.9	32.3	7.0	26.6	32.4	9.1	28.7	34.2	12.1

Table 1: Results over **10 random runs** on COCO dataset under 5, 10, 30-shot settings. Note that we use the same samples as TFA[2] so that the metrics are directly comparable. We obtain better performance in terms of all metrics.

detection is still hard by nature. We analyze several typical failure patterns in Figure2(b). The first four columns show false positive cases, mainly due to: 1) Though not common, the base detector sometimes produces false positives on unseen objects, producing overlapped predicted boxes of both base and novel categories on the same instance; 2) features are not discriminative enough for few-shot categories, thus confusion among classes like misclassification among few-shot classes and domination of base classes over novel classes. The fifth column shows a typical case for transfer learning based methods where novel class objects are hard to be detected due to deactivation in the backbone, showing that such bias caused by pretraining is hard to be alleviated. The sixth column shows another failure pattern caused by box regression, probably because accurate localization for categories with complex shapes is also challenging to learn under low-shot scenarios.

4. Results over Multiple Runs

To show the effectiveness of our method without random effect, we ran our model over 10 sets of random samples under 5, 10, 30-shot settings on COCO dataset, using exactly the same samples as TFA[2]. The results are shown in Table1. We obtain better performance in terms of all metrics (AP, bAP, nAP) under each of these settings.

References

[1] 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, 2019. 2

[2] Xin Wang, Thomas Huang, Joseph Gonzalez, Trevor Darrell, and Fisher Yu. Frustratingly simple few-shot object detection. In *International Conference on Machine Learning*, 2020. 1, 2, 3

[3] 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 9576–9585, 2019. 2