Supplementary Material

5.1. Training Details

As there is no parameter shared between FSCN and the base detector, we can split the whole training process into two separate phases. In the first detector training phase, we train the Faster-RCNN detector $\mathcal{F}_d(\cdot)$ only on the base classes $\mathcal{C}_{bs}$ for the first 10 epoch, where the obtained detector is denoted as $\mathcal{F}_d^{bs}(\cdot)$. Then, we apply the proposed CGDP to create a clean and balanced training set $\mathcal{D}_{ft}$ which contains both base and novel classes. Next, the classification weights of $\mathcal{F}_d^{bs}(\cdot)$ are set for both novel classes and background through the above introduced imprinting strategy, the obtained $\mathcal{F}_d^{imp}(\cdot)$ is further fine-tuned on $\mathcal{D}_{ft}$ to obtain the final detector $\mathcal{F}_d^{*}(\cdot)$, while the backbone and RPN are kept to be fixed.

For the training of FSCN, firstly, it is pre-trained on a set of region proposals $\mathcal{X} = \{(\boldsymbol{I}_{pi}, c_i)\}$ sampled from $\mathcal{D}_{bs}$ by using the pre-trained base detector $\mathcal{F}_d^{bs}(\cdot)$, where $\boldsymbol{I}_{pi}$ is the image patch cropped from original input space in accordance of the proposal coordinates, and c_i denotes the corresponding one-hot target. During this pre-training stage, we aim to learn a good feature representation for FSCN $\mathcal{F}_r^{bs}(\cdot)$, which can be easily generalized to unseen categories in future. Upon the arrival of novel classes, similarly, the classifier of $\mathcal{F}_r^{bs}(\cdot)$ are extended for those novel classes and the background weights are also updated, denoted as $\mathcal{F}_r^{imp}(\cdot)$. Next, $\mathcal{F}_r^{imp}(\cdot)$ is further fine-tuned on the set of region proposals sampled from $\mathcal{D}_{bs} \cup \mathcal{D}_{nv}$ by using the final base detector $\mathcal{F}_d^{*}(\cdot)$. Particularly, for those hard backgrounds sampled from $\mathcal{D}_{bs}$ inside each training batch, they are first filtered by the proposed UOM strategy to get the potential unlabeled objects of novel classes. Therefore, the proposed distractor utilization loss is only applied on those obtained high-possibility backgrounds, while the conventional cross-entropy loss is used for the rests. In addition, we set the intersection ratio threshold for UOM as 0.4, i.e., the proposed distractor utilization loss is only applied to those base-set background proposals which have intersection ratio less than 0.4.

5.2. Dataset Settings

For the Pascal VOC dataset, following the common practice, our model is trained on the union of 07 and 12 train/validation set and is evaluated on 07 test set. Following the splitting rule of Meta-RCNN, we consider three different splits of base and novel classes, which are (novel: bird, bus, cow, bike, sofa / base: the others), (novel: aero, bottle, cow, horse, sofa / base: the others) and (boat, cat, bike , sheep , sofa/ the other). During training, model can only access k object instances from each novel class, where we set $k = \{1, 2, 3, 5, 10\}$ for Pascal VOC. For MS COCO dataset, we train our model on the union of the 80k train set and 35k trainval set, and evaluate on the 5k minival set, with $k = \{10, 30\}$. Considering the total 80 categories in COCO dataset, 20 categories included in Pascal VOC are used as novel classes, and the rest are used as base classes.

Table 3. Knowledge Retention on Base Classes

Methods	Base Split 1	Base Split 2	Base Split 3
FRCN-base	70.3	71.8	70.4
FRCN-ft	67.2	67.6	66.8
Meta-RCNN	67.9	-	-
FRCN-ft + FSCN	**73.1**	**73.6**	**72.7**

Table 4. Unfreeze representation learning for TFA

Methods	Novel AP	Base AP
TFA	9.8	27.7
TFA-finetuned	10.8	25.9
TFA+FSCN	**11.1**	**28.1**

5.2.1 Unfreeze representation learning for TFA

When incrementally adding novel classes, TFA freezes the feature representation and only finetunes the box classifier with novel data. This could be viewed as an approach for preserving the previously gained knowledge and countering catastrophic forgetting. However, without optimizing features for novel classes, the pre-trained embedding space could be biased towards features of base classes, which might lead to poor generalization and underfit performance on unseen classes, especially for those classes that have large feature discrepancy in comparison with base classes. To verify this, we finetune the RoI feature layer with the last two linear layers on a small balanced training set with early stopping. We denote it as TFA-finetuned and use it as an alternative baseline to show the superiority of our method. As shown in Table. 4, we notice that there is a performance gain 1.0 point on novel classes while 1.8 points significant performance degradation on base classes. This indicates that detector finetuning might bring more discriminative features to novel classes and alleviate category confusion. However, the worse results on overall performance prove that finetuning representation with small data is likely to suffer from overfitting. In contrast, we decouple the classifier finetuning and representation learning into two networks separately. By doing so, each branch can separately perform its own duty for knowledge retention and novel-class promotion. As a result, we not only enhance the feature representation for novel classes but also ensuring performance on base classes does not degrade.

Table 5. Ablation Study on Distractor Utilization Loss

Methods	Novel AP	Base AP
FSCN + CE	13.5	28.3
FSCN + DUL	15.6	26.8
FSCN + DUL + UOM	**15.1**	**27.6**

5.2.2 Knowledge Retention on Base Classes

To adapt a well-trained detector model to new detection tasks, previous works usually consider jointly fine-tuning, where base classes and novel classes are trained together on a small balanced training set. However, a consequent issue is that the obtained model are likely to be overfitted due to the small amount of training samples. As a result, it suffers from catastrophic forgetting on old categories, i.e., fails to preserve the previous gained knowledge and leads to a performance degradation on base classes. This issue is quite common but still under-explored yet [7].

To evaluate the effectiveness of our proposed FSCN in addressing catastrophic forgetting, we compare with three baseline methods. In detail, FRCN-base is only trained on the base set of Pascal VOC, where the base classes should have the highest accuracy as the model is only trained on the base classes. The second baseline FRCN-ft is fine-tuned from FRCN-base on a small balanced set that contains both base and novel classes. The third baseline Meta-RCNN also takes the same initialization point FRCN-base for the few-shot adaption, but with the difference of training another meta learner. As we are only interested in the performance of knowledge retention, CGDP is excluded from model evaluation. The results are shown as in Table. 3. After adapting to novel classes, both FRCN-ft and Meta-RCNN suffer from severely performance drop on base classes. In contrast, with the help of FSCN , our framework can easily overcome the catastrophic forgetting issue by taking the advantage of false positive suppression, and eventually outperforms the best performed baseline FRCN-base by a large margin.

5.2.3 Ablation Study on Distractor Utilization Loss

We compare results with three different strategy for dealing with distractors in the training of FSCN. Experiments are conducted on MS-COCO: (1) FSCN + CE: FSCN is trained with the default cross-entropy loss on all datasets. (2) FSCN + DUL: The proposed distractor utilization loss is applied on all background proposals sampled from the base set, but without further training sample selection. (3) FSCN + DUL +UOM: The proposed distractor utilization loss is only applied on the background proposals selected by UOM. Results are shown in Table 5. As we can see, FSCN + CE get lowest accuracy on novel classes, since it blindly making use of unlabeled objects as negative examples, the accumulated discouraging gradients will enforce a classification bias to predicting novel classes into background and lead to a catastrophic detection performance. In contrast, FSCN + DUL achieves the best accuracy on novel classes but with the worst performance on base classes. We conjecture that this is because the encouraging effect to novel classes is applied on all base-set background proposals, thus the learned classifier will be biased to predict background into novel classes and lead to significantly performance degradation on base classes. To this end, with the proposed adaptive sample selection, the encouraging effect is only applied to the potential unlabeled objects, thus FSCN + DUL + UOM has a much better base/novel accuracy trade-off over FSCN + DUL and FSCN + CE. Due to space limitation, more ablation studies are included in appendix.

Table 6. Ablation Study on FSCN Architecture

Backbone	Novel AP	Base AP
ResNet50	51.5	72.0
ResNet101	51.3	72.3
ResNet50-CGNL	56.7	73.4
ResNet101-CGNL	**56.9**	**73.8**

5.2.4 Ablation Study on FSCN Architecture

We also study the impact of different architecture design for the FSCN model. We compare four options, which are (1) ResNet50 (2) ResNet101 (3) ResNet50-CGNL(4) ResNet101-CGNL, to verify the importance of full receptive field as well as network depth. Experiments are conducted on Pascal VOC dataset, and CGDP is excluded from model evaluation. The results are shown in Table 6. Although stacking more layers does provide sightly better accuracy on base classes, performance on the few-shot classes is insensitive to network depth. Regarding this, ResNet50 has a better speed/accuracy trade-off over ResNet101 and it is also more suited for real-time applications.

We also compare the influence of adding fully receptive field into FSCN, e.g ResNet50 v.s. ResNet50-CGNL. Since CGNL module is simply a fully connected layer, these two networks have approximately the same number of parameters. However, we only observe a marginal gain of 2.4 point with ResNet50 while our ResNet50-CGNL has a much larger gain of 7.6 point. To this end, we conclude that the performance gain mainly come from the global receptive field rather than enabling more parameters.