A Supplementary

A.1 Training and Testing Time

One concern about our multi-step transfer framework is the elongated training time over a one-step transfer. However, we want to note that the refinements do not necessarily require a full training schedule. In the K-th refinement, the model is initialized from the $(K-1)$-th step, and trained with only a fractional number of steps of the initial schedule for both the one-class detector and the MIL classifier, as described in Sec. 4.

We list the actual time in Table 3. To reach 58% mAP with 1 initial training and 2 refinements, the total time is 190+23+(22+5+50+8)*2=383m, which is only 1.8 times of that of the initial model (190+23m) instead of 3 times. To reach the best 59.7% mAP, 1 initial training and 5 refinements took 638m, which is 3 rather than 6 times. As progressive knowledge transfer is an offline process, we hope the higher detection mAP can pay off the longer offline cost.

Table 3. Time of different training stages of the COCO-60-to-VOC experiment on a 4 Nvidia V100 GPU server.

Stage	Training steps	Wall clock time
OCUD initial ($K = 0$)	17500	190 min
OCUD refine ($K > 0$)	5000	50 min
MIL initial ($K = 0$)	5000	23 min
MIL refine ($K > 0$)	2000	8 min
Mine pseudo GT on COCO-60	(22K images)	22 min
Mine pseudo GT on VOC trainval	(5K images)	5 min

As for the testing time, we provided optional distillation post-processing step, reported as "Ours(distill)" in the main text. The distillation retrains an ordinary Faster RCNN with the mined pseudo GT for 10000 steps in 121min, whose testing time is comparable to a usual Faster RCNN (about 2min on 4952 VOC test images). Without distillation, the combined OCUD and classifier takes roughly 5min (multi-scale).

A.2 ILSVRC Transfer Setting

Some previous works [10,29,30,35] have studied the transfer learning based weakly supervised object detection problem with the ILSVRC 2013 dataset [4]. We test the effectiveness of our algorithm under this setting with minimal modification from the COCO-to-VOC experiment.

Source and Target Datasets. The ILSVRC detection dataset contains 200 object categories [4]. We construct the source and target data following the protocol in [10,29,30,35]. The category names are sorted in alphabetic order.

The first half of 100 categories are the source domain categories, and the rest half are the target domain categories.

The val1 and val2 validation set splits as in [7] are adopted. The val1 set originally contains 9886 images. As the source dataset, we keep 4487 images (6881 box annotations) of the first 100 categories from val1, and augment it with images sampled from the ILSVRC training set. A maximum of 1000 images per category are sampled, resulting in 78486 images (108591 box annotations). As for the target dataset, we similarly augment the val1 images of the latter 100 categories (3609 images and 7103 annotations) with maximum 1000 ILSVRC training images (76847 images and 124991 annotations) as the target training set, while keeping only image-level labels (the box annotations are ignored). The val2 dataset with annotations of the latter 100 categories (9917 images and 14079 boxes) is used as the target test set.

Evaluation Metric. We report the same mAP (calculated by the VOC 07 method) at IoU threshold 0.5 as in Sec. 4 of the main text on the val2 set.

Network. ResNet-50 is used as the backbone for both the OCUD and the MIL classifier. The output channels are modified to 101 to account for the different number of classes.

Training and Inference. Due to the increased size of the dataset, the iteration 0 OCUD was trained for 70000 steps (with a batch size of 8 images on 4 GPUs). The later refinements of OCUD took 20000 gradient steps. The MIL classifier was initially trained for 40000 steps and fine-tuned for 10000 steps during refinements. The learning rate schedule is the same as the COCO-to-VOC experiments. The hyper-parameters were also comparable $\beta = 5, \lambda = 0.2, \tau = 0.8, o = 0.1$, imagesize $= 640$, except for the larger score interpolation coefficient η: $\eta = 0.85$ and the smaller prediction score threshold 0.001. We ran 2 refinement iterations in total. During inference, we did single-scale testing as it delivers roughly the same performance as the multi-scale counter-part.

Results. We compare results from the ILSVRC transfer experiment with prior methods in Table 4. Our initial iteration detector without refinement (Ours(K=0)) achieved 33.50% mAP. After 2 refinements, the mAP was boosted to 37.00% (Ours(K=2, ResNet50, flip)). Our result is on-par with the best previously-known result under the same protocol [30] and better than most existing works. However, there was a performance gap between our initial detector and the one in [30], although these two should be comparable due to the method similarity. We suspect that the reason behind this is the use of a weaker backbone in our experiment (ResNet-50) compared to the Inception ResNet. We are currently investigating deeper into this. Nevertheless, the improved performance after refinements again demonstrates the effectiveness of our progressive knowledge transfer method.

A.3 More Visualizations

We visualize more images to support the arguments made in the main text.

Table 4. mAP (%) on the ILSVRC2013 val2 set (weak 100 categories)

Method	$mAP_{.5}$
LSDA [10] (AlexNet)	18.08
LSDA reproduced by [29] (VGG16)	18.86
LSDA reproduced by [29] (VGG19)	21.02
Tang et al. [29] (RCNN, VGG16)	24.91
Tang et al. [29] (RCNN, ResNet50)	28.30
Tang et al. [29] (Fast RCNN, VGG16)	26.22
Tang et al. [29] (Fast RCNN, ResNet50)	29.71
MSD [35] (AlexNet)	22.28
MSD [35] (VGG16)	25.26
Uijlings et al.[30] (AlexNet)	23.3
Uijlings et al.[30] (Inception-ResNet)	36.9
Ours(K=0, ResNet50, no augmentation)	32.97
Ours(K=0, ResNet50, flip)	33.50
Ours(K=2, ResNet50, no augmentation)	36.90
Ours(K=2, ResNet50, flip)	37.00

COCO-60-full to VOC Transfer. In the main text, we visualize the mined pseudo ground truth after 2 refinements in Fig. 6. More successful examples are shown in Fig. 7 and 8. We further observed that more refinement iterations could sometimes lead to higher quality pseudo ground truths by, for example, discovering objects missed by previous iterations or localizing objects more accurately.

ILSVRC-179 to VOC Transfer. Compared with COCO-60 and COCO-60-full, the ILSVRC-179 has more missing annotations and less consistency with VOC. This can be seen from the basic statistics of the datasets and the visualizations (Fig. 9). The ILSVRC-179 has an average of $\frac{345854+55502}{395909+20121} \approx 0.96$ boxes per image while COCO-60 has $\frac{70549+2914}{21987+921} \approx 3.21$, and VOC test has $\frac{12032}{4952} \approx 2.43$.

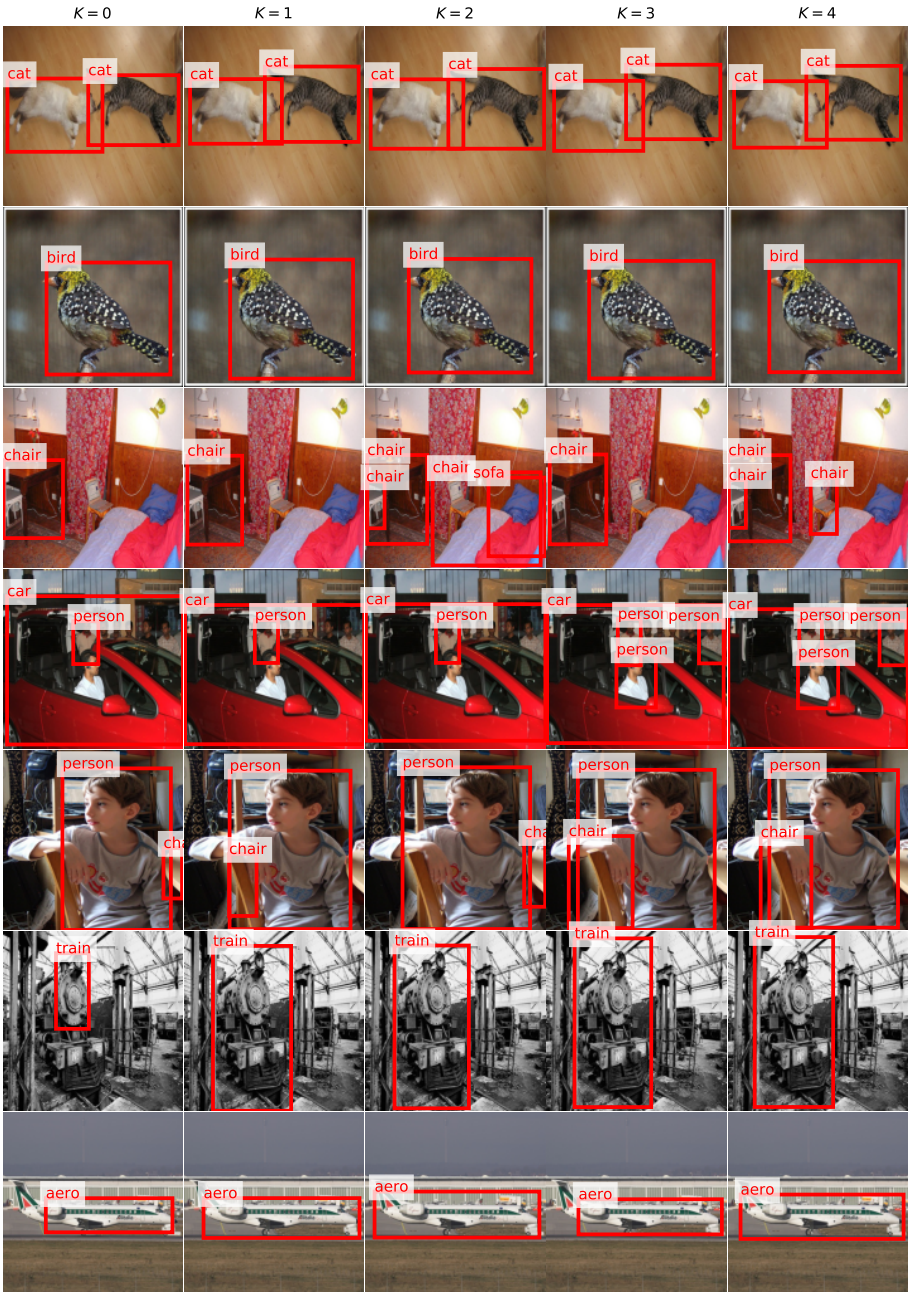

Fig. 7. Successfully mined pseudo ground truths (in red) in VOC trainval at number of refinements $K = 0, 1, 2, 3, 4$ of the COCO-60-full to VOC experiment.

Fig. 8. Successfully mined pseudo ground truths (in red) in COCO-60-full at number of refinements $K = 0, 1, 2, 3, 4$ of the COCO-60-full to VOC experiment.

Fig. 9. Successfully mined pseudo ground truths (in red) in ILSVRC-179 at different iterations of the ILSVRC-179 to VOC experiment.