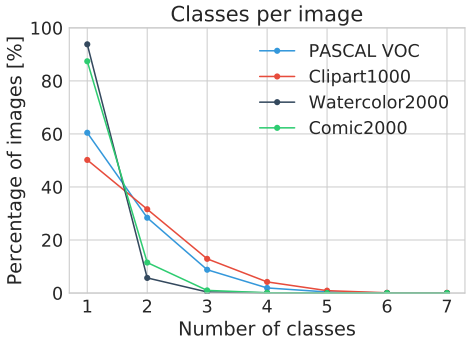
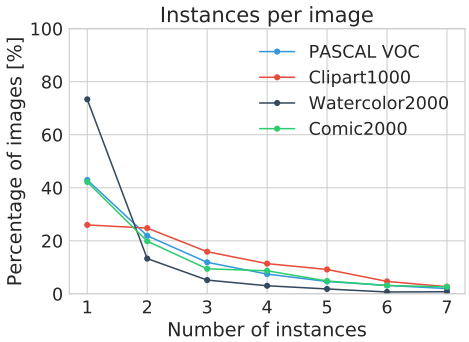
Supplementary Material: Cross-Domain Weakly-Supervised Object Detection through Progressive Domain Adaptation

Naoto Inoue Ryosuke Furuta Toshihiko Yamasaki Kiyoharu Aizawa
The University of Tokyo, Japan
{inoue, furuta, yamasaki, aizawa}@hal.t.u-tokyo.ac.jp



(a) The number of classes per image in our datasets.



(b) The number of instances per image in our datasets.

Figure 1: Number of classes and instances in our datasets. For PASCAL VOC, we used all the annotations including `difficult` boxes. Note that there are twenty object classes in PASCAL VOC and Clipart1k, and six object classes in Watercolor2k and Comic2k.

1. Statistics of Our Datasets

An important characteristic of our datasets is that they contain a sufficient number of objects. The number of classes and instances per image is shown in Fig. 1. For comparison, the figure contains the statistics of PASCAL VOC [1], which is designed for detecting objects of twenty classes in natural images. Clipart1k contains 1.7 classes and 3.2 instances per image. Clipart1k contains almost the same number of classes and instances per image as PASCAL VOC. The average number of classes and instances in Clipart1k is almost the same as that in PASCAL VOC, which ensures the difficulty for the process of object detection. Watercolor2k contains 1.1 classes and 1.7 instances per image. Comic2k contains 1.1 classes and 3.2 instances per image. Note that Watercolor2k and Comic2k are for detecting the six classes.

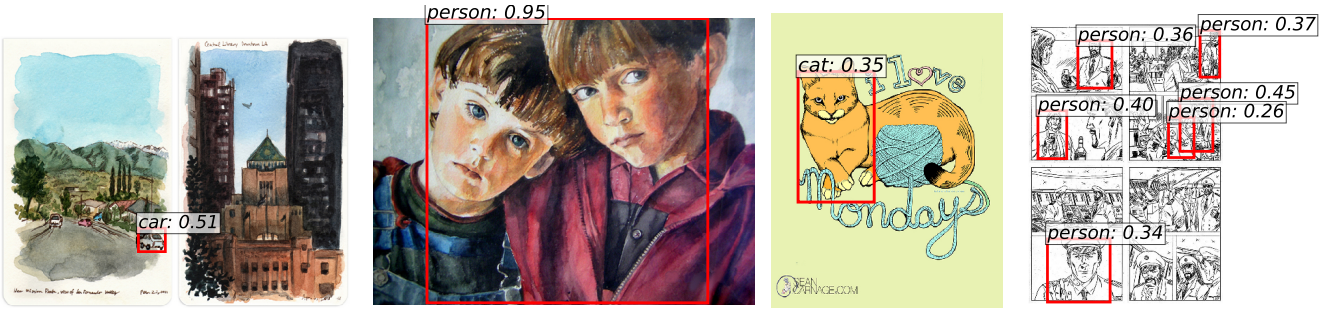
As shown in Table 1, the distribution of the number of the instances for each class in Clipart1k is unbalanced, as is also seen in PASCAL VOC [1]. In Table 2, the number of instances in Watercolor2k and Comic2k is shown. In all datasets, the person class is dominant.

Table 1: The number of instances in Clipart1k.

Name	#instances	Name	#instances
Aeroplane	73	Dining table	115
Bicycle	36	Dog	54
Bird	265	Horse	79
Boat	129	Motorbike	17
Bottle	121	Person	1185
Bus	21	Potted plant	178
Car	202	Sheep	76
Cat	50	Sofa	52
Chair	340	Train	46
Cow	46	TV/monitor	80

Table 2: The number of instances in Watercolor2k and Comic2k.

Dataset	Bicycle	Bird	Car	Cat	Dog	Person	Total
Watercolor2k	27	486	101	102	116	2483	3315
Comic2k	87	270	107	233	192	5500	6389



(a) Ignoring small objects (b) Merging two objects (c) Localizing most discriminative parts only (d) Highly-deformed objects

Figure 2: Typical detection errors by DT+PA using SSD300 as the baseline FSD. The images are from the test set of Clipart1k and Comic2k.

2. Visualization of Detections

We discuss the detection results produced by our methods. We will show the typical detection errors of our methods in Fig. 2. The errors are often caused due to ignoring small objects (Fig. 2a), merging highly-overlapped objects which belong to the same object class (Fig. 2b), localizing only the most discriminative part of an object (Fig. 2c), or being unable to recognize highly-deformed objects (Fig. 2d). The detections results obtained by our methods are shown in Fig. 3, Fig. 4, and Fig. 5. We confirm that our method is generally applicable and valid for various depiction styles.

3. Implementation Details

3.1. Domain Transfer

All the images were loaded and resized to 286 $\times$ 286. In the fine-tuning phase, the images were randomly cropped to the size 256 $\times$ 256. In the test phase, the images were loaded, transferred, and converted back to the original size. We used all 16,551 images in VOC2007-trainval and VOC2012-trainval and obtained the domain-transferred images.

3.2. Configurations for training FSDs

For YOLOv2 [3], we used the original implementation and employed a learning rate of 1.0×10^{-5}. The input images were resized to 416 $\times$ 416. With the IoU threshold (0.45) and the confidence threshold (0.001) employed, YOLOv2 was fine-tuned for five epochs and one hundred epochs for the DT and other experiments, respectively.

For Faster R-CNN [4], we used the reimplementation provided in ChainerCV [2]. We employed a learning rate of 1.0×10^{-5}. The length of the shorter edge of the input image was scaled to 600. After the scaling, if the length of the longer edge was longer than 1,000, the image was scaled so that the length of the longer edge came down to 1,000. With the IoU threshold (0.3) and the confidence threshold (0.05) employed, Faster R-CNN was fine-tuned for one epoch and one hundred epochs for the DT and other experiments, respectively.

References

[1] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2), 2010. 1

[2] Y. Niitani, T. Ogawa, S. Saito, and M. Saito. ChainerCV: a library for deep learning in computer vision. In *ACM Multimedia*, 2017. 2

[3] J. Redmon and A. Farhadi. YOLO9000: Better, Faster, Stronger. In *CVPR*, 2017. 2

[4] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 2

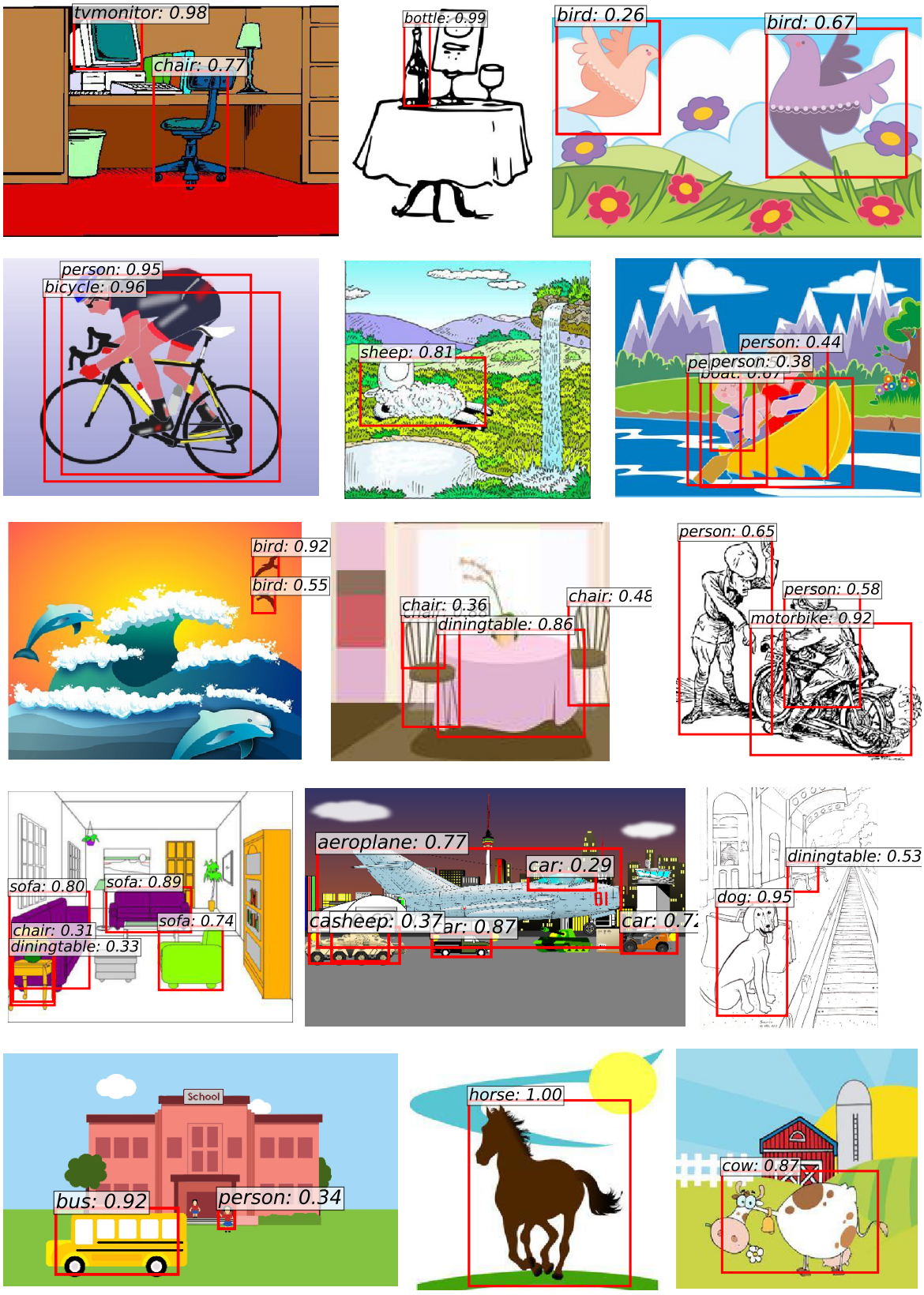


Figure 3: Example outputs for our DT+PA using SSD300 as the baseline FSD in the test set of Clipart1k. We only showed windows whose scores are above 0.25 so as to maintain visibility.

Figure 4: Example outputs for our DT+PA using SSD300 as the baseline FSD in the test set of Watercolor2k. We only showed windows whose scores are above 0.25 so as to maintain visibility.

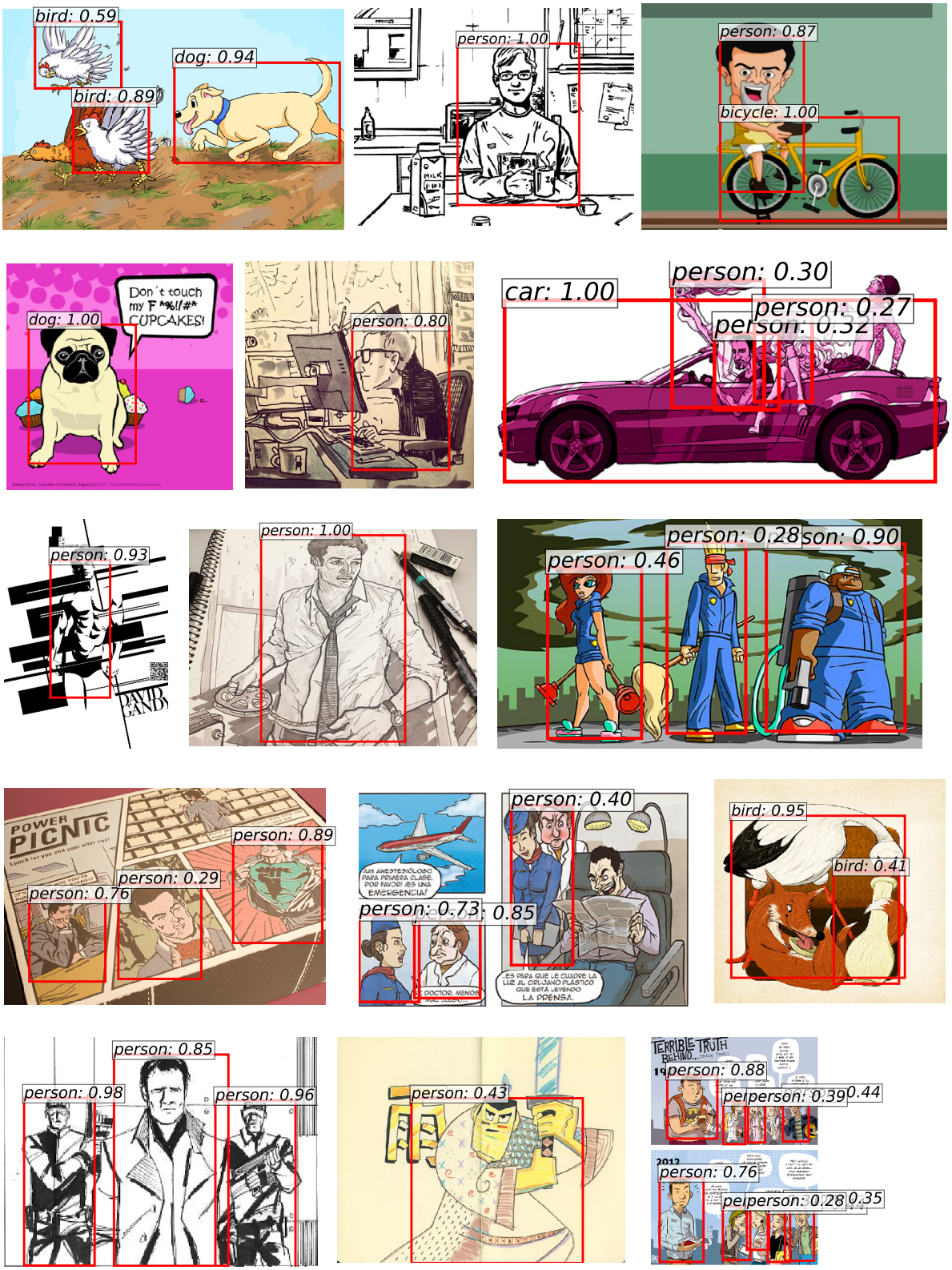


Figure 5: Example outputs for our DT+PA using SSD300 as the baseline FSD in the test set of Comic2k. We only showed windows whose scores are above 0.25 so as to maintain visibility.