Supplementary Material: Automatic adaptation of object detectors to new domains using self-training

Aruni RoyChowdhury Prithvijit Chakrabarty Ashish Singh SouYoung Jin
Huaizu Jiang Liangliang Cao Erik Learned-Miller
College of Information and Computer Sciences
University of Massachusetts Amherst
{arunirc, pchakrabarty, ashishsingh, souyoungjin, hzjiang, llcao, elm}@cs.umass.edu

Abstract

In addition to the experiments in the main paper, we include the following extended discussions and results:

- *Histogram specification for cross-domain score mapping (Sec. 1).*
- *Application to semi-supervised learning (Sec. 2).*
- *Pedestrian dataset statistics (Sec. 3).*
- *Additional qualitative results (Sec. 4).*

For further details, please visit the project webpage: http://vis-www.cs.umass.edu/unsupVideo/

1. Histogram specification

We provide a brief review, mostly adapted from Gonzales and Woods [2], of the histogram specification method used to map between the distribution of scores in source and target domains. The terms *original* and *desired* is used to denote the two distributions we want to map between, to reduce ambiguity with *source*-domain and *target*-domain.

Assuming continuous values for ease of exposition, let the original distribution have probability density function (p.d.f.) $p_r(r)$, with $0 \leq r \leq 1$. Let the desired distribution have p.d.f. $p_z(z)$, with $0 \leq z \leq 1$.

Let us consider the cumulative distribution functions (c.d.f.) as two transformations F and G, acting on the original and desired distributions, respectively.

$$s = F(r) = \int_0^r p_r(w)dw \tag{1}$$

$$v = G(z) = \int_0^z p_z(u)du \tag{2}$$

From Eq. 2, $z = G^{-1}(v)$, will give back the values z of the desired distribution $p_z(z)$. Instead of v, if the values of s in Eq. 1 are used, we can re-map values r from the original distribution $p_r(r)$ to values in the desired distribution:

$$z = G^{-1}(s) = G^{-1}[F(r)] \tag{3}$$

$\mathcal{T} \rightarrow \mathcal{S}$ score mapping. This involves making the distribution of detector scores on the target domain $\mathcal{T}$ resemble the distribution of scores on the source domain $\mathcal{S}$. The score values are binned between 0 and 1 with step-size of 0.01. The inverse mapping is done using linear interpolation.

$\mathcal{S} \rightarrow \mathcal{T}$ score threshold. This involves the reverse of the previous process – we choose a threshold based on labeled source data, and then "transfer" this to the target domain via histogram specification, as above.

2. Application to semi-supervised learning

Table 1: Semi-supervised learning results on BDD.

Method	AP	# images
Baseline	20.07 ± 0.00	12,477
Det	30.25 ± 0.34	100,001
HP	30.35 ± 0.58	100,001
HP-*cons*	31.36 ± 0.67	100,001
Ground-truth	35.38 ± 0.83	57,513

The general approach of re-training a model on using a mixture of labeled and unlabeled data is an instance of semi-supervised learning, without necessarily having a domain adaptation component, *i.e.* there is no domain shift between train and test datasets – we merely augment the labeled training set with additional unlabeled (pseudo-labeled) data. Table 1 shows results using BDD(*clear,daytime*) as labeled data, the rest of BDD as unlabeled data, and evaluations on the BDD(*clear,daytime*) test set. The extra self-labeled training data (`Det`) improves considerably over the baseline. Training on a smaller amount of perfect la-

bels (Ground-truth) is an expected upper-bound. Our soft-labeling strategy emphasizes hard examples from the target domain, which is specific to the domain adaptation task and not useful for general semi-supervised learning – we can see soft-labels (`HP-cons`) is not significantly better than hard labels (`HP`). Implicit domain-shift in the unlabeled data (lots of night-time videos, while we test on day) is a realistic confounding factor – we cannot always ensure that the unlabeled data is exactly matching the train and test distributions. Eliminating such confounding factors should improve the test performance.

3. BDD 100k dataset statistics

The BDD 100k dataset [5] provides detailed annotations on conditions such as time of day (*e.g.* dawn, dusk, night, day) and weather (*e.g.* rainy, clear, snowy, *etc.*), among others. Each of the 100k videos has *one* frame annotated with objects. We use these ground-truth annotations to divide the dataset into source and target domains, summarized in Figure 1. Note that when we perform pseudo-labeling on target domain videos, we are discarding all label information, except what was used to create the two domains.

4. Additional qualitative results

We show additional qualitative results on the BDD [5] pedestrian dataset (Figures 2, 3). Please note, due to restrictions and privacy concerns [4], we show only selected images from CS6/IJB-S in the main paper and do not include extensive qualitative results from surveillance videos in this supplemental.

References

[1] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3339–3348, 2018.

[2] R. C. Gonzalez, R. E. Woods, et al. Digital image processing, 2002.

[3] S. Jin, A. RoyChowdhury, H. Jiang, A. Singh, A. Prasad, D. Chakraborty, and E. Learned-Miller. Unsupervised hard example mining from videos for improved object detection. In *European Conference on Computer Vision (ECCV)*, 2018.

[4] N. D. Kalka, B. Maze, J. A. Duncan, K. OConnor, S. Elliott, K. Hebert, J. Bryan, and A. K. Jain. Ijb–s: Iarpa janus surveillance video benchmark.

[5] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018.

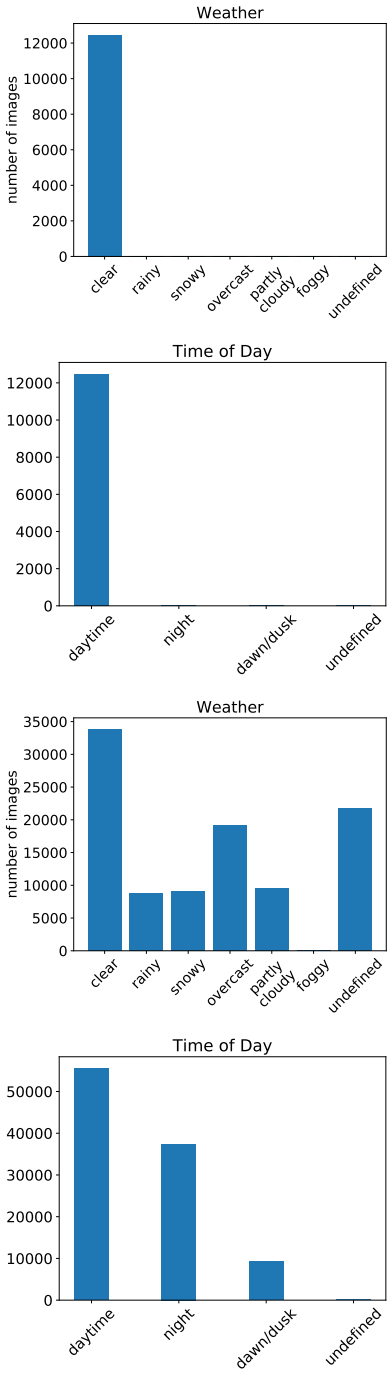

Figure 1: Images in BDD 100k [5] across weather conditions and time of day are used to create source and target domains. *Rows 1-2:* the number of images in source **BDD(*clear,daytime*)**. *Rows 3-4:* the number of images in target **BDD(*rest*)**, spanning all other weather and time-of-day conditions.

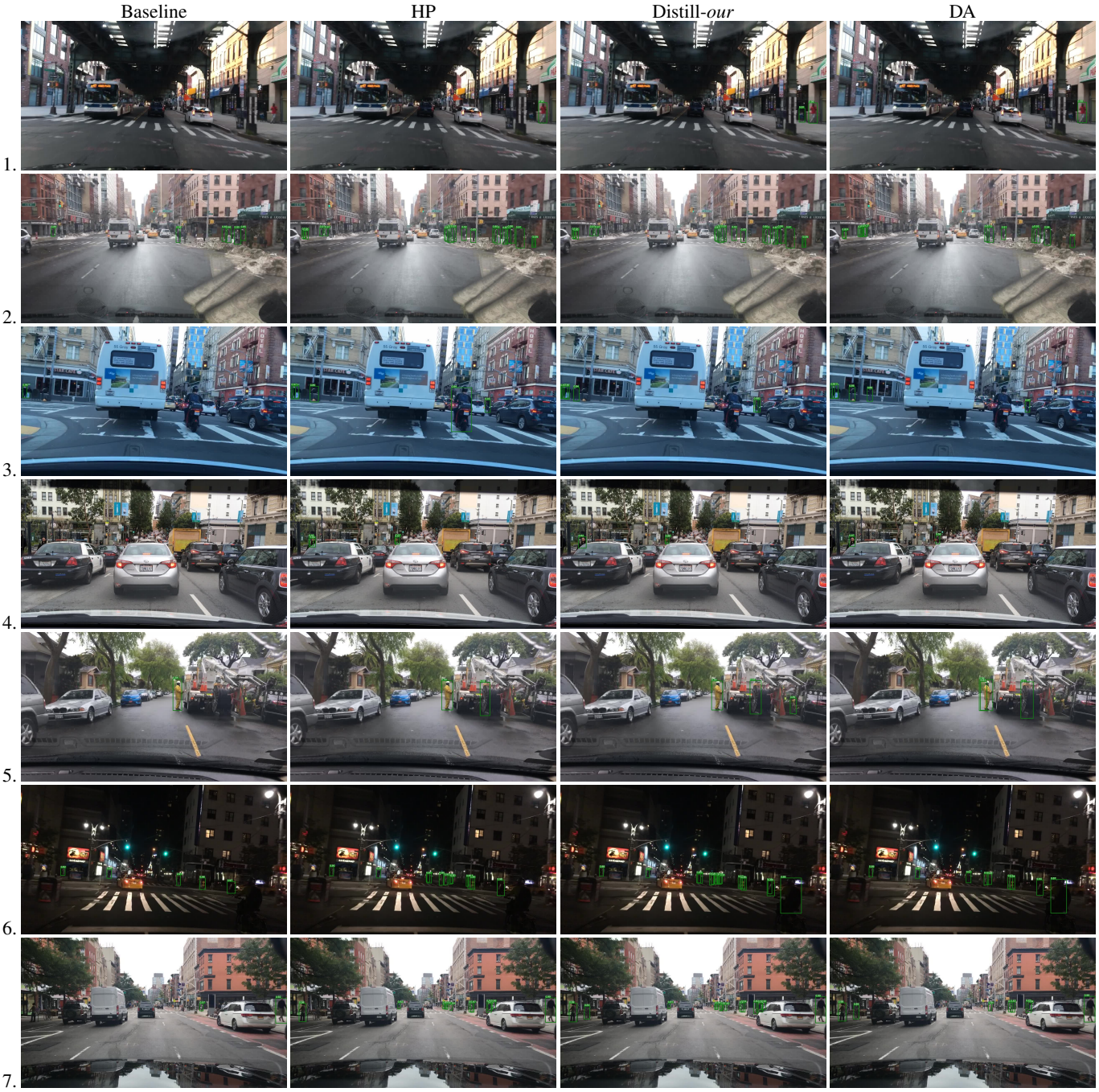

Figure 2: **Qualitative results**(best zoomed-in)**.** (a) Baseline; (b) HP [3]; (c) Distill-*ours*; (d) DA[1]. The domain adapted methods (HP, Distill, DA) pick up prominent objects missed by the baseline detector, along with a few false positives (*rows 1,4,7*). *Row 2:* Distill and HP get the prominent pedestrian on the right, while DA misses it. *Row 3:* the HP method detects a motorcycle rider as a pedestrian, while the soft-labeled Distill method gets this subtle difference correctly. Failure modes: part of the wheel detected as a pedestrian by all the domain adapted methods (*row 5)*; rider detected as pedestrian (*row 6*).

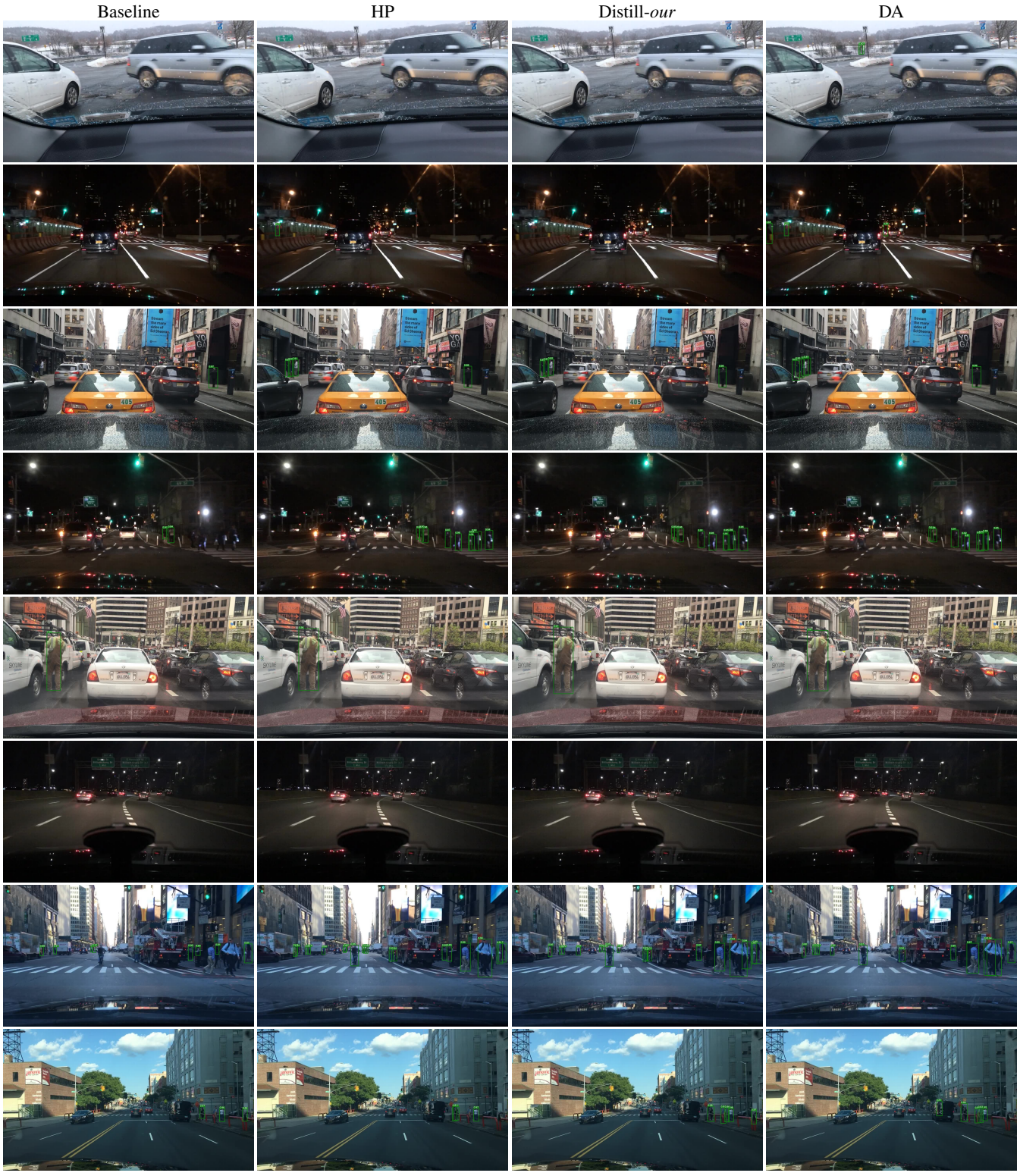

Figure 3: **More qualitative results**(best zoomed-in)**.** (a) Baseline; (b) HP [3]; (c) Distill-*ours*; (d) DA[1]. *Rows 1,2:* the domain adversarial method (DA) detects false positives, which are avoided by our Distill method. Pedestrians that are challenging for the baseline detector are picked up after domain adaptation in *rows 3,4,7,8*. *Row 5:* when conditions are well-lit and clear – similar to the training set of the Baseline mode, there is not much difference with the domain adapted models.