Supplementary Material for
Task-specific Inconsistency Alignment for Domain Adaptive Object Detection

Liang Zhao   Limin Wang ✉
State Key Laboratory for Novel Software Technology, Nanjing University, China
liangzhao@smail.nju.edu.cn, lmwang@nju.edu.cn

A. More Implementation Details

In this section, we present more details about the *Baseline* model. As mentioned in main text (Sec. 3.1), for higher semantic consistency, we adhere to the mainstream practice of aligning features on the source and target domains, at both mid-to-upper layers of the backbone (*i.e.* image-level) and ROI layer (*i.e.* instance-level), with the help of Gradient Reversal Layer (GRL) [6]. Concretely, in consistent with [13], for the features output from the last three blocks of VGG16 [14], or last three layers of ResNet101 [8], we feed them into separate discriminators (D_1, D_2 and D_3, their concrete architecture is shown in Tab. A.1) connected via a GRL to determine the domain to which the features belong. After that, three image-level domain adaptation losses are calculated as follows:

$$\begin{aligned}\mathcal{L}_{da}^{img1} = & \frac{1}{n_s \cdot H \cdot W}\sum_{i=1}^{n_s}\sum_{w=1}^{W}\sum_{h=1}^{H} D_1(x_i)_{wh}^2 \\ & + \frac{1}{n_t \cdot H \cdot W}\sum_{i=1}^{n_t}\sum_{w=1}^{W}\sum_{h=1}^{H} (1 - D_1(x_i)_{wh})^2,\end{aligned} \tag{1}$$

$$\begin{aligned}\mathcal{L}_{da}^{img2} = & \frac{1}{n_s}\sum_{i=1}^{n_s}\mathcal{L}_{ce}(D_2(x_i^{'}), d_i^s) \\ & + \frac{1}{n_t}\sum_{i=1}^{n_t}\mathcal{L}_{ce}(D_2(x_i^{'}), d_i^t),\end{aligned} \tag{2}$$

$$\begin{aligned}\mathcal{L}_{da}^{img3} = & \frac{1}{n_s}\sum_{i=1}^{n_s}\mathcal{L}_{fl}(D_3(x_i^{''}), d_i^s) \\ & + \frac{1}{n_t}\sum_{i=1}^{n_t}\mathcal{L}_{fl}(D_2(x_i^{''}), d_i^t),\end{aligned} \tag{3}$$

where x_i, $x_i^{'}$ and $x_i^{'''}$ denotes the features output from the last three blocks of the backbone for the i-th training image, d_i indicates the corresponding domain label, and n_s and n_t refer to the total number of images within a mini-batch in source and target domains, respectively. Besides, L_{ce} suggests the cross-entropy loss, while the L_{fl} indicates the focal loss, with its γ set to 5 following [11]. Likewise, the alignment of high-level feature patches (ROIs) is also employed. With the discriminator D_4 illustrated in Tab. A.1, the instance-level loss is formally as

$$\begin{aligned}\mathcal{L}_{da}^{ins} = & \frac{1}{n_s}\sum_{i=1}^{n_s}\mathcal{L}_{ins}(D(r_i), d_i^s) \\ & + \frac{1}{n_t}\sum_{i=1}^{n_t}\mathcal{L}_{ins}(D(r_i), d_i^t),\end{aligned} \tag{4}$$

where r_i denotes the i-th ROI and d_i indicates the corresponding domain label. As for $\mathcal{L}_{ins}$, we use cross-entropy loss for the *Normal-to-Foggy* and *Cross-Camera* scenarios and focal loss for the *Real-to-Artistic* scenario, with γ being also set to 5.

In conclusion, the overall training objective of *Baseline* becomes:

$$\mathcal{L} = \mathcal{L}_{det} + \lambda_1(\mathcal{L}_{da}^{img1} + \mathcal{L}_{da}^{img2} + \mathcal{L}_{da}^{img3} + \mathcal{L}_{da}^{ins}), \tag{5}$$

where λ_1 is set to 1.0. Additionally, we concatenate the image-level features processed by previous three discriminators with the high-level ROI representation after FCs, in a manner similar to [11, 13], to realize greater training stability.

B. Additional Ablation Study

For the localization-specific inconsistency alignment module, the effect of **different measures of dispersion** is further investigated here. To reveal more clearly their impact on the localization branch, we remove the classification branch. The results on the *Real-to-Artistic* scenario are displayed in Tab. B.1. It showcases that (1) a measure that is closer to the original scale is preferred; (2) L2-norm delivers a more appropriate and precise estimate to behavioral uncertainty among diverse localizers.

✉: Corresponding author.

Discriminator D_1
Conv $1 \times 1 \times 256$, stride 1, pad 0
ReLU
Conv $1 \times 1 \times 128$, stride 1, pad 0
ReLU
Conv $1 \times 1 \times 1$, stride 1, pad 0
Sigmoid

Discriminator D_2 and D_3
Conv $3 \times 3 \times 512$, stride 2, pad 1
Batch Normalization, ReLU, Dropout
Conv $3 \times 3 \times 128$, stride 2, pad 1
Batch Normalization, ReLU, Dropout
Conv $3 \times 3 \times 128$, stride 2, pad 1
Batch Normalization, ReLU, Dropout
Average Pooling
Fully connected 128×2

Discriminator D_4
Conv $3 \times 3 \times 512$, stride 2, pad 1
ReLU
Conv $3 \times 3 \times 128$, stride 2, pad 1
ReLU
Conv $3 \times 3 \times 128$, stride 2, pad 1
ReLU
Average Pooling
Fully connected 128×2

Table A.1. Architecture of discriminators.

Measurement	mAP
Mean absolute deviation	42.7
Variance	41.8
Standard deviation	43.2

Table B.1. Ablation study on different measures of dispersion.

C. Visualization

We provide some detection results of vanilla detector (*i.e. Source Only* [10]), state-of-the-art adaptive detectors (*e.g. HTCN* [1] and *UMT* [4]), and our framework TIA. Fig. F.1 illustrates the comparison of detections on the PASCAL VOC [5] $\rightarrow$ Clipart [9] benchmark. It is observed that our proposed TIA outperforms both *Source Only* and *UMT* [4], and produces more accurate detection results, *i.e*., more foreground objects are identified (Row 1&2), and higher quality bounding boxes are provided along with accurate categorization (Row 3-5). Qualitative results on the Cityscapes [3] $\rightarrow$ Foggy Cityscapes [12] benchmark represented by Fig. F.2 also demonstrates the superiority of our TIA. For example, in the first row, for the two cars on the left, the bounding box given by *HTCN* is relatively off-target, while ours method present more compact boundaries, compared to *Source Only*'s.

D. Limitations

The discrepancy between source and target domains in the label space, *i.e*., label shift, substantially affects the design philosophy and severely limits the performance of existing domain adaptive detectors. In this subsection, we will provide in-depth analysis of how label shift limits our TIA for each dataset benchmark.

The benchmarks used in *Normal-to-Foggy* (Cityscapes [3] $\rightarrow$ Foggy Cityscapes [12]) and *Real-to-Artistic* (PASCAL VOC [5] $\rightarrow$ Clipart [9]) are essentially appropriate and they allow a good evaluation of the performance of various domain adaptive detectors. Specifically, the former case is ideal, since it shares an **identical label space** between the source and target domains, while the latter one has its label shift diluted due to the scale of the source domain. In this context, it is observed that, our framework exceeds the upper bound indicated by *Target Only* on the former benchmark and easily achieves state-of-the-art performance on the latter benchmark.

It is quite different in the *Cross-Camera* scenario. We find that the label shift of the benchmark (KITTI [7] $\leftrightarrow$ Cityscapes) employed in this scenario is dominated by the imbalance in the foreground-background ratio, namely the inconsistency in the average number of objects between the source and target domain data. In fact, the average numbers of instances of Cityscapes and KITTI are 9.1 and 3.8, respectively. This directly leads to two serious problems. On the one hand, we observe that the *Source Only* model undergoes severe overfitting issue during training, which means that we underestimate the lower bound of the benchmark; on the other hand, it imposes higher demands on the cross-domain performance of RPN, and this straightforwardly undermines the effectiveness of the existing mainstream approaches that focus on feature alignment for it.

In summary, two arguments are made. First, existing methods are highly inefficient in coping with label shift. In light of [15], although the execution of domain alignment alone reduces the divergence between domains (the second term in Theorem 1), it leads to arbitrary increases in λ^* (the third term in Theorem 1), hence eventually, the target errors of detectors cannot be well-guaranteed. For this reason, taking into account the detectors' empirical predictions on the target domain, or namely, the behavior of label predictors, is gradually emerging as a necessity. Moreover, compared to classification tasks, the label shift in object detection task is considerably complicated. It is no longer limited to the differences in category proportions, but is more widely distributed in spatial differences in scale, position, etc. of bounding boxes. These two facts drive the proposal of TIA on a different aspect.

Second, in view of the fact that the label shift cannot be well estimated nor truly eliminated, we argue that there is a gap between the true upper bound and the present upper bound specified by *Target Only*, according to [16]. Under such circumstances, the close performance of the domain adaptive detectors in the *Cross-Camera* benchmark can be reasonably explained.

E. Societal Impact

Domain adaptive object detection is a prevalent visual scene understanding task and we follow the convention experimental setting as in [1, 2, 4, 11]. Hence if the method is used properly, there is no negative social impact.

F. Code and Dataset License

Our code is built on open-sourced object detection code with MIT license. As for the datasets, the Cityscapes [3] and its modification Foggy Cityscapes [12] are *made freely available to academic and non-academic entities for non-commercial purposes such as academic research, teaching, scientific publications, or personal experimentation*; and the PASCAL VOC [5] includes images obtained from the "flickr" website; Clipart [9] is meant for education and research purposes only; in addition, KITTI [7] is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License.

References

[1] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8869–8878, 2020. 2, 3, 5

[2] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc 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. 3

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 3

[4] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021. 2, 3, 4

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 3

[6] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 1

[7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 2, 3

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[9] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018. 2, 3

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 2

[11] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019. 1, 3

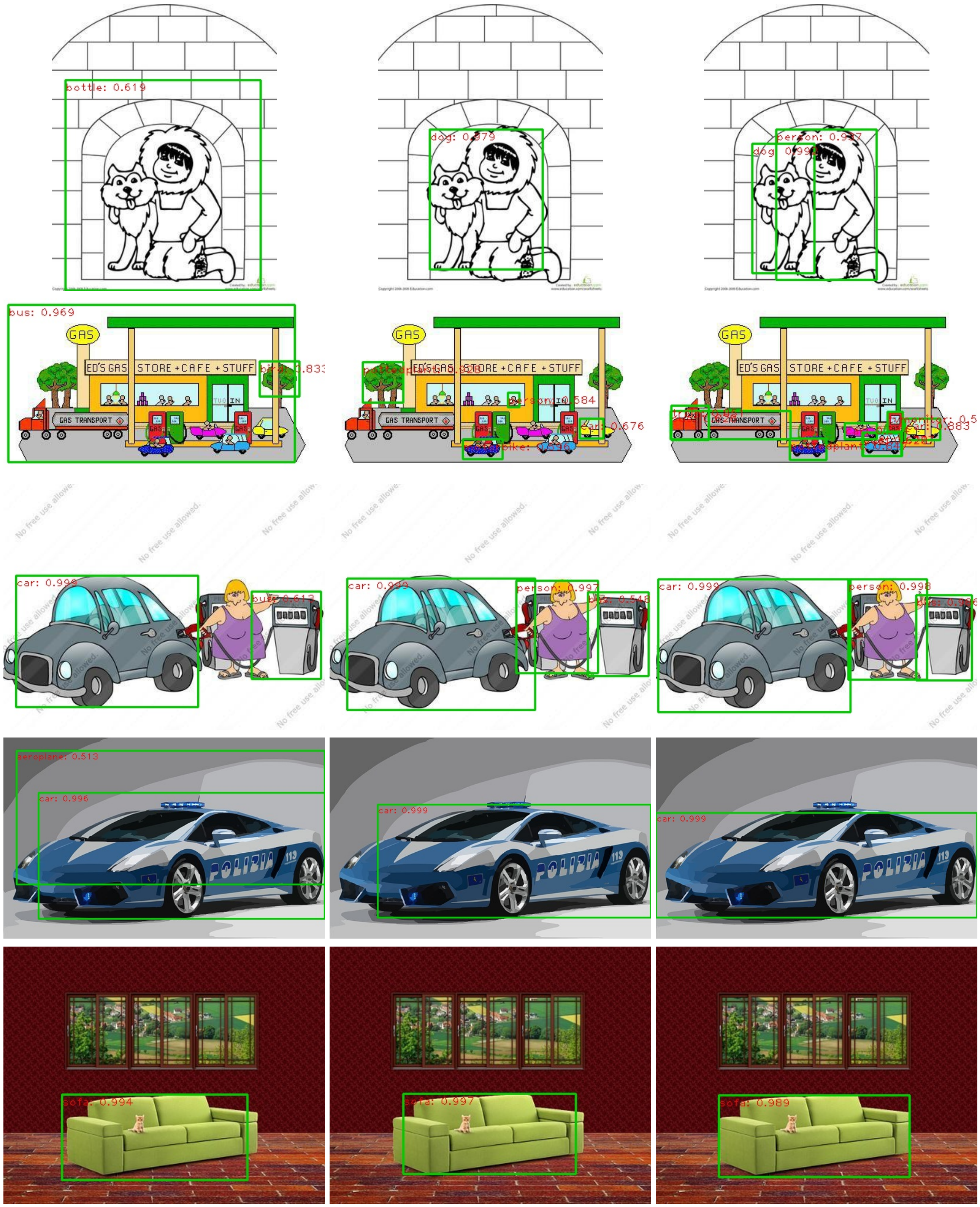
[12] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018. 2, 3

[13] Zhiqiang Shen, Harsh Maheshwari, Weichen Yao, and Marios Savvides. Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. *arXiv preprint arXiv:1911.02559*, 2019. 1

[14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

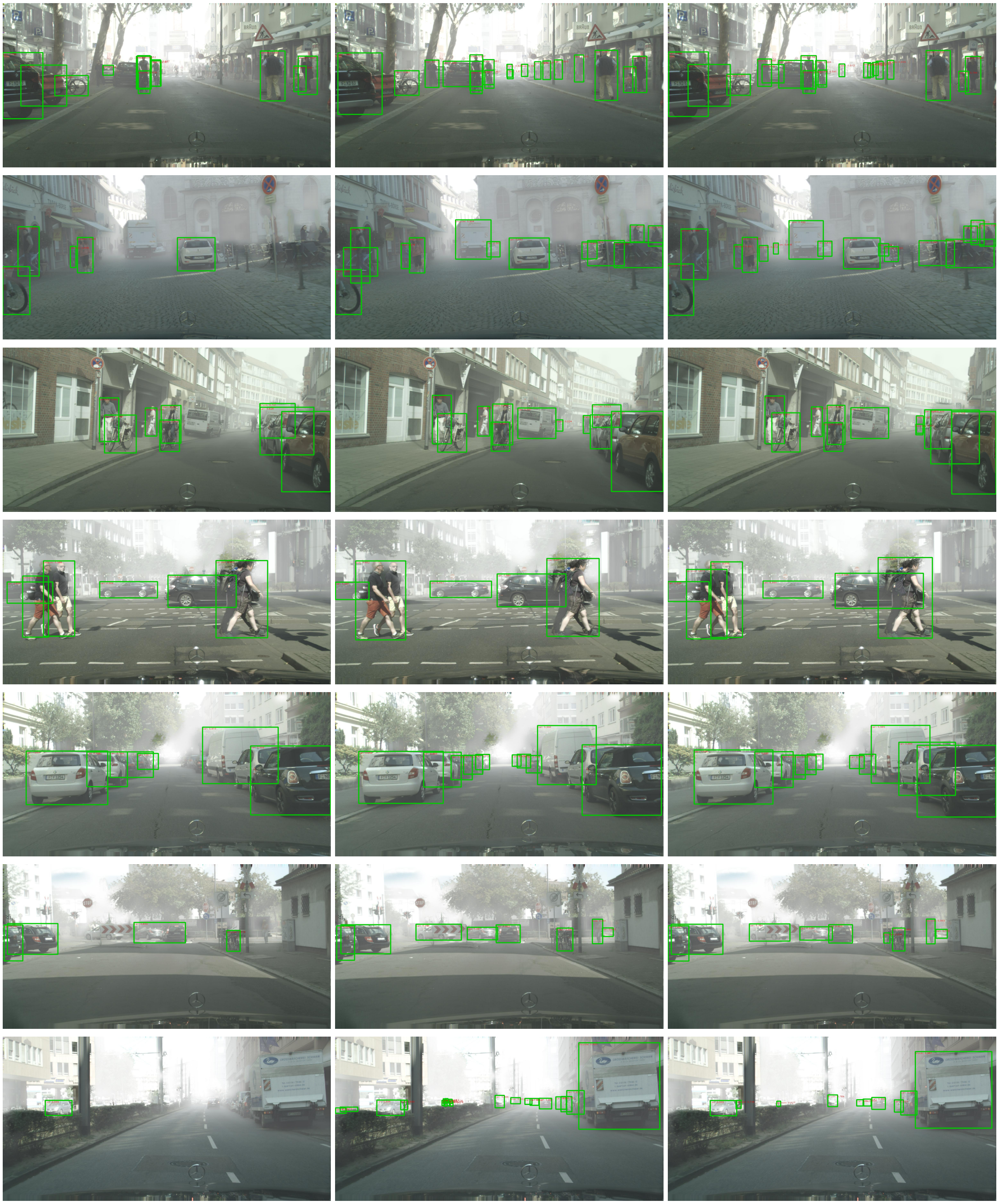
[15] Yifan Wu, Ezra Winston, Divyansh Kaushik, and Zachary Lipton. Domain adaptation with asymmetrically-relaxed distribution alignment. In *International Conference on Machine Learning*, pages 6872–6881. PMLR, 2019. 2

[16] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pages 7523–7532. PMLR, 2019. 2



(a) Source Only (b) UMT [4] (c) Our TIA

Figure F.1. Illustration of the detection results on the PASCAL VOC → Clipart benchmark. Compared to *Source Only*, *UMT*'s localization performance is worse, while ours is better.



(a) Source Only (b) HTCN [1] (c) Our TIA

Figure F.2. Illustration of the detection results on the Cityscapes → Foggy Cityscapes benchmark. Our TIA identifies more objects and delivers more accurate bounding boxes.