Supplementary Material for Harmonious Teacher for Cross-domain Object Detection

In this Supplementary, we present implementation details, additional experimental analysis, and discussion on limitations and the societal impact of our Harmonious Teacher (HT).

1. Implementation Details

Mean Teacher: The mean teacher (MT) framework was originally proposed by [7] for semi-supervised learning. Many cross-domain object detection (CDOD) works [2, 5] adopted the MT framework for its simplicity and effectiveness. The student model is updated by the SGD algorithm and supervised by the source labeled samples and pseudo-labeled target samples. The parameters of the teacher model are updated by the exponential moving average (EMA) of the student model as follows:

$$\theta_t := \epsilon\theta_t + (1-\epsilon)\theta_s, \tag{1}$$

where θ_t and θ_s are the parameters of the teacher model and the student model, respectively. ϵ is the EMA rate which controls the percentage of the parameters of the teacher network in the previous step to be retained. In our experiments, the ϵ is set to 0.9996. The update interval (*i.e.*, the interval for updating the teacher model) is set to 1 iteration for all experiments.

Self-training Loss for Regression: The self-training loss on the target domain contains two losses, *i.e.*, classification loss and regression loss. In our implementation, the classification loss uses the Quality Focal Loss [4] for the continuous label from the teacher model. The regression loss $\mathcal{L}_{reg}^t$ contains two parts including IoU loss $\mathcal{L}_{iou}$ and binary cross entropy (BCE) loss $\mathcal{L}_{bce}$. The IoU loss can be written as follows:

$$\mathcal{L}_{iou} = -ln(u), \tag{2}$$

where u is the IoU between the predicted bounding boxes at the same position in the feature map from the teacher model and the student model.

Optimization Pipeline: We depict the overall optimization pipeline of our HT in Algorithm 1, which contains pre-training stage and self-training stage. In the pre-training stage, we train the source model with the supervision $\hat{\mathcal{L}}_s$ of labeled source samples. This enables the model to have a harmonious initialization for the following self-training stage. In the self-training stage, we use the teacher model with weakly augmented target images to generate pseudo labels. While the student model with strong augmented target images is supervised by the generated pseudo labels from the teacher model. We design a harmony measure to provide accurate quality estimation for the predictions from the teacher model. It can be used to reweigh the self-training loss for all the predicted instances. In this way, all pseudo-labeled samples can contribute to the model training based on their prediction qualities, and the hard threshold is not needed anymore.

2. Experiments

The Inconsistency between Classification and Localization in CDOD: We present a further study on the inconsistency between classification and localization in cross-domain object detection (CDOD) scenarios.

The object detector may produce inconsistent predictions even in the in-domain object detection scenario. However, in the CDOD scenario, the inconsistent issue could become even worse, due to the domain discrepancy between the source domain and the target domain. To verify this, we show the three correlation coefficients between classification score and IoU with ground truth boxes on Cityscapes and Foggy Cityscapes using the source model trained on Cityscapes in Table 1. We can easily find that the correlations between classification and localization significantly decrease on the target domain because the domain discrepancy between the source and target domains leads to performance degradation. The presence of inconsistency between classification and localization harms the performance of the self-training strategy in CDOD. The main drawback of inconsistency in self-training is that the confidence score (*i.e.*, classification score) cannot correctly reflect the quality of the predicted bounding boxes. The noise pseudo labels are dangerous for the following self-training and thus significantly limit the detection performance on the target domain.

The Effect of Harmony Measure: The harmony measure explicitly encodes the classification prediction and localization score into a unified metric to estimate the quality of pseudo labels. It provides a more accurate ranking of

Table 1. The Pearson correlation coefficient (PCC), Spearman rank correlation coefficient (SCC) and Kendall rank correlation coefficient (KCC) between classification score and IoU with ground truth boxes, where the model is trained on the Cityscapes and evaluated on the Cityscapes and Foggy Cityscapes.

Test Domain	PCC (%)	SCC (%)	KCC (%)
Cityscapes [1]	39.2	33.3	22.8
Foggy Cityscapes [6]	35.9 (-3.3)	30.1 (-3.2)	20.5 (-2.3)

pseudo labels than the classification score. Here, we present the false negative (FN) rate comparison under different selection ratios between the classification score and our harmony measure in Fig. 1. The lower FN indicates the model detects more ground truth objects. From Fig. 1, we can observe that the proposed harmony measure achieves a lower FN rate than the classification score across multiple selection ratios. These results support our claim that the harmony measure provides a more accurate quality estimation of pseudo labels.

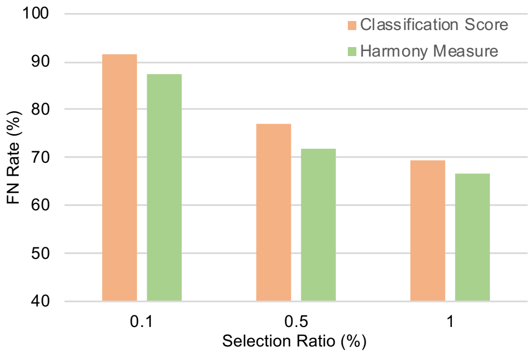

Figure 1. The false negative (FN) rate on Foggy Cityscapes for the source model pre-trained on Cityscapes under different sample selection ratios. The harmony measure shows consistently lower FN rates than the classification score on different sample selection ratios.

More Qualitative Results: We show more qualitative results on adaptation from Cityscapes to Foggy Cityscapes in Fig. 2. Compared with Source Only [8] and the state-of-the-art work SIGMA [3], our HT produces more accurate predictions. For example, our HT can correctly detect more objects (green box) than other baselines.

3. Limitation

Although our method outperforms many existing cross-domain object detection (CDOD) works, it still has some failure cases (Fig. 2). For example, we can see that our HT may miss some small-scale and obscure objects. We conjecture that this could be caused by the limited representation ability. On the one hand, we could utilize a more advanced backbone network to enhance the representation ability. On the other hand, we can leverage contrastive learning to improve the feature representation.

Algorithm 1 Harmonious Teacher.

```
Input: Source domain $\mathcal{S}$, Target domain $\mathcal{T}$.
Output: The parameters of the teacher model $\theta_t$ and the student model $\theta_s$.
Initialize the source model parameters $\theta$ with the ImageNet pre-trained weights.
while Pre-training Stage do
    Train the source model $\theta$ by calculating $\hat{\mathcal{L}}_s$ in Eq. (10)
end while
Initialize the student model $\theta_s \leftarrow \theta$
Initialize the teacher model $\theta_t \leftarrow \theta$
while Self-training Stage do
    Sample source images from the source domain $\mathcal{S}$
    Calculate supervised loss $\hat{\mathcal{L}}_s$ in Eq. (10)
    Sample target images from the target domain $\mathcal{T}$
    Inference on weakly augmented target images by using $\theta_t$
    Calculate harmony measure $h$ based on Eq. (7)
    Feed the strongly augmented target images to the student model $\theta_s$
    Calculate $\hat{\mathcal{L}}_t$ based on Eq. (9)
    Calculate $\mathcal{L}_u$ based on Eq. (6)
    Calculate the overall objective $\mathcal{L}$ based on Eq. (10)
    Train the student model $\theta_s$
    Update the teacher model $\theta_t$ via EMA
end while
return $\theta_t$, $\theta_s$
```

4. Societal Impact

With the ability to address cross-domain object detection, our Harmonious Teacher has potential positive impacts on the many downstream systems (*e.g.*, autonomous driving) that often face unseen domains in real-world applications. However, the large-scale source data may cause privacy issues, *e.g.*, surveillance videos and medical images. We plan to study source-free object detection, where we adapt a source pre-trained detector to an unlabeled target domain without accessing the source samples.

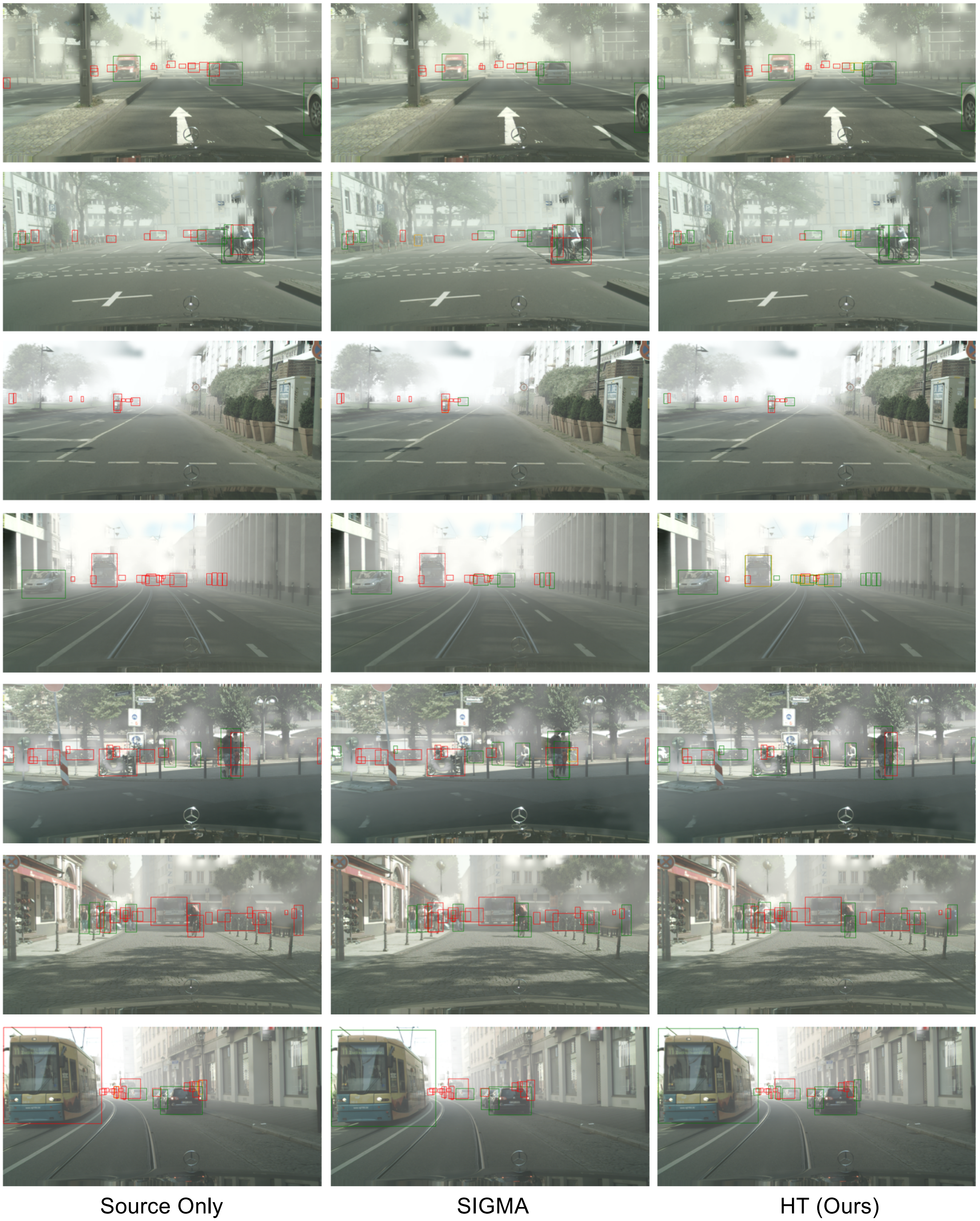

Figure 2. Qualitative results on the target domain of Cityscapes to Foggy Cityscapes. Green, red and orange boxes indicate true positive (TP), false negative (FN) and false positive (FP), respectively. We set the score threshold to 0.7 for better visualization. Best appreciated when viewed in color and zoomed up.

References

[1] 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 *CVPR*, pages 3213–3223, 2016. 2

[2] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, pages 4091–4101, 2021. 1

[3] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *CVPR*, pages 5291–5300, 2022. 2

[4] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *NeurIPS*, 33:21002–21012, 2020. 1

[5] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *CVPR*, pages 7581–7590, 2022. 1

[6] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 126(9):973–992, 2018. 2

[7] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 30, 2017. 1

[8] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, pages 9627–9636, 2019. 2