

# Supplementary Materials: Stochastic Context Consistency Reasoning for Domain Adaptive Object Detection

Anonymous Authors

## A A CHALLENGE OF MASKING STRATEGY

### A.1 Challenge Formulation

MIC [8] proves the masking strategy can improve performance across different visual recognition tasks. Here, we explore a challenge in the single masking strategy, where entire objects may be masked (as shown in Figure 1), leading to unfair penalization during training. This unfair penalization means that the teacher model generates pseudo labels from the complete images, but the student model is penalized with pseudo labels for failing to predict objects completely occluded by the mask. As training progresses, the error accumulation of the student model is passed on to the teacher model through the Exponential Moving Average (EMA), creating a vicious cycle. We utilize different masked patch sizes and conduct experiments on several common cross-domain datasets [5, 10, 15] to evaluate the percentage of objects that are entirely obscured, as shown in Figure 2. We observe this situation is prevalent across the above datasets and gets worse as the masked patch size increases.

### A.2 Our Solution

Different from MIC, our three modules address this challenge from the following technical perspectives:

**In SCM**, we use the stochastic complementary masks to generate complementary masked images (i.e., two views  $\{\mathcal{V}_1, \mathcal{V}_2\}$ ), and feed them to the student model, separately. The complementary mechanism ensures that the student model receives complete target image information in each iteration, preventing half of the image information from being omitted due to a single masking strategy. This ensures fairer penalization for the student model.

**In Inter-CCR**, to prevent mismatches resulting from enforcing consistency, we employ a higher IoU threshold for bounding box matching between the two views to ensure that the matched bounding boxes come from the same object. As illustrated in Figure 3, we calculate the IoU matrix for the student model’s predictions from two views  $\{\mathcal{V}_1, \mathcal{V}_2\}$ , and select the maximum value along the longest side as their matching score. If this score exceeds the preset  $\tau$ , the bounding boxes are considered successfully matched.

**In the Intra-CCR**, predictions of the student model for each of the two views are computed with pseudo labels to calculate the loss. Because the actual number of objects in  $\{\mathcal{V}_1, \mathcal{V}_2\}$  is different, to balance the contributions of the two branches in Intra-CCR, we set  $\lambda = \frac{N}{N+M}, \mu = \frac{M}{N+M}$  in Eq. 12 and Eq. 13, i.e., the bigger number of objects will contribute more to the calculated loss  $\mathcal{L}_{cons}^{Intra}$ . We also set  $\lambda = 1 - r, \mu = r$ , which assumes that the number of objects is inversely proportional to the mask ratio. As shown in Figure 4, the former (full line) performs better when the mask rate  $r$  is closer to 0.5, while the latter (dashed line) performs better at  $r < 0.4$ , indicating that the latter is more suitable for handling extreme cases. As  $r = 0.5$  achieves the best performance, we choose the setting of  $\lambda = \frac{N}{N+M}, \mu = \frac{M}{N+M}$ , ultimately.

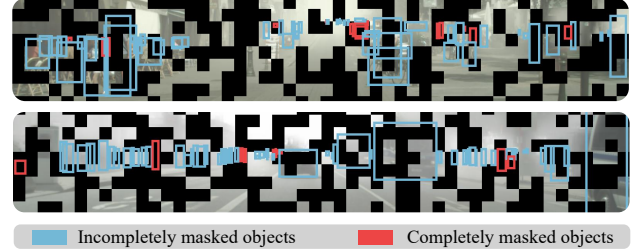


Figure 1: The masking strategy may result in entire objects being occluded. (where masked patch size  $b = 32$ ).

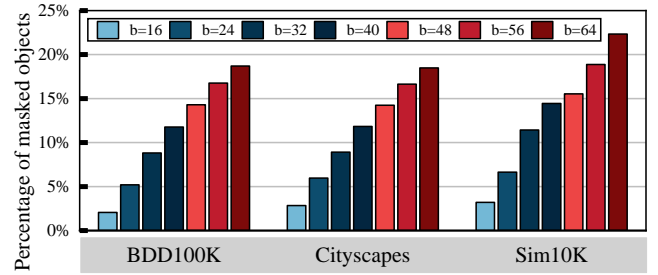


Figure 2: Percentage of completely masked objects to the total number of objects under different masked patch sizes.

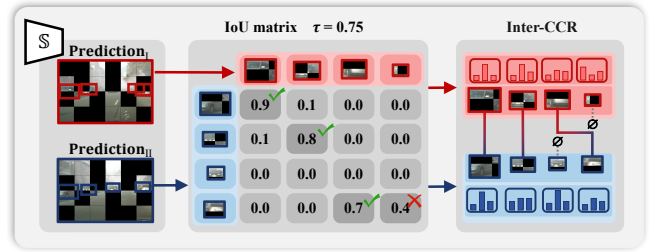


Figure 3: IoU-based matching mechanism of Inter-CCR.

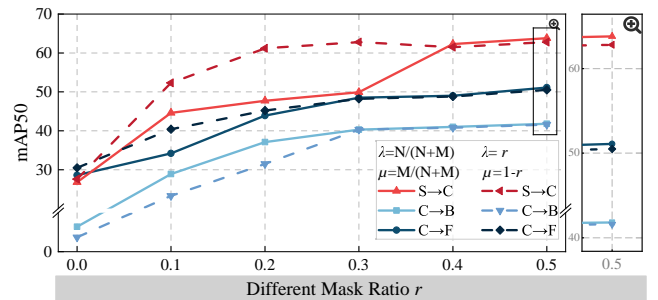


Figure 4: We explore the performance of SOCCER at different mask ratios under two loss weight settings.

## B MORE DIFFERENT MASKING MODE

### B.1 Masking Out Data Augmentation.

We further discuss the difference between our SCMasking strategy and other masking out methods (as shown in Figure 5). Chen *et al.* [2] proposes that avoiding excessive deletion and reservation of continuous regions is the core requirement for masking out methods. Recently, researchers intend to find a balance between deleting and reserving regional information on the images. However, this is very difficult for two reasons: On the one hand, excessively deleting one or a few regions may result in the removal of complete object and context information. Thus remaining regions are not enough to be detected by the network. On the other hand, excessive preserving regions could make some objects untouched, which cannot provide challenging detection scenarios for the network to learn the context information. Neither Cutout [7], Random Erasing [17], nor GridMask [2] can achieve true balance, because creating challenging detection scenarios for the network will always mask a large number of objects. It always caused the unfair penalization, as discussed in Section A. Our stochastic complementary masking strategy provides a good solution for the balance between deleting and reserving of regions.

### B.2 Regular Pattern Masking.

From the above experiments, we observe that stochastic masking can prevent the network from over-relying on specific visual clues and can train the model to predict the context reasoning ability of the masked region. However, would using regular pattern masking lead to better results?

As illustrated in Figure 6, we employ three types of regular pattern masking: checkerboard, horizontal stripes, and vertical stripes. We conduct experiments with the above masking types on Sim10k to Cityscapes, keeping all other settings unchanged, and the results are summarized in Table 1. We observe that regular pattern masking does not effectively contribute to enhancing the detection performance. During the training process, the network may only become sensitive to fixed positions in the context region. In other words, the network learns to take shortcuts to predict only the fixed unmasked region and loses the ability to actively utilize the context information. A similar result is obtained when we fix the stochastic mask. We also conduct experiments by randomly applying these three regular pattern masking during the training of the network. we observe that random application of the above regular pattern masking can also obtain a similar performance to the stochastic complementary masking strategy. The result indicates that the stochastic mechanism is better than the fixed masking to encourage the network to explore the masked region, and it is independent of the masking pattern (stochastic or regular).

## C DAY-TO-NIGHT DAOD BENCHMARK

Recently, the safety of autonomous driving at night has focused on more and more researchers and domain adaptation technology is undoubtedly the most effective solution. So we further evaluate our SOCCER's domain adaptive performance in the day-to-night direction. Note that our approach is mainly designed for general cross-domain adaptation, rather than day-to-night adaptation.

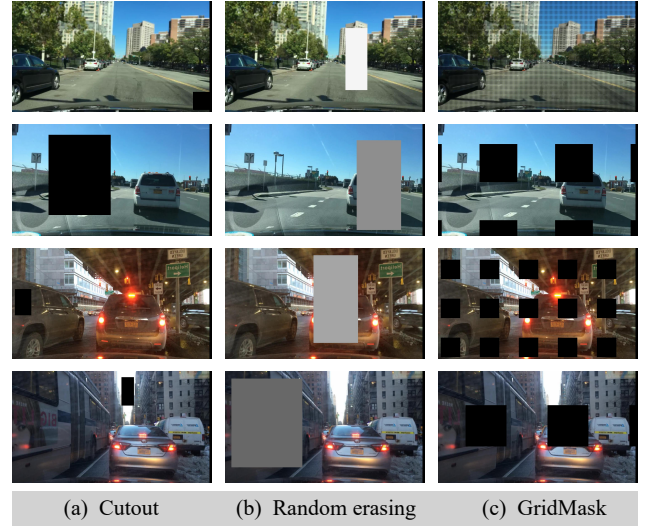


Figure 5: More examples of different masking out augmentation methods: Cutout [7], Random Erasing [17], and GridMask [2]. We observed that none of the above data augmentation measures strike a good balance between deletion and retention regions.



Figure 6: Three types of regular pattern masking: checkerboard, horizontal stripes, and vertical stripes (where masked patch size  $b = 32$ ).

Table 1: The impact of different types of regular pattern masking on the performance of SOCCER. "random" denotes randomly applying the above three regular pattern masking in the training of SOCCER, SCMasking denotes the stochastic complementary masking strategy, and fixed SCMasking denotes the stochastic masking is fixed. All types of regular pattern masking are detrimental to performance, but random application of these three regular pattern masking leads to similar results as SCMasking.

S→C (car)	AP <sub>0.5</sub>	AP <sub>0.75</sub>	AP <sub>0.5:0.95</sub>
checkerboard	60.7	34.1	33.9
horizontal stripes	61.1	34.0	34.0
vertical stripes	61.0	34.8	34.4
fixed SCMasking	61.4	34.9	34.8
random	63.7	38.3	<b>36.9</b>
SCMasking(ours)	<b>63.8</b>	<b>38.4</b>	36.8

**Table 2: Results of Day-to-Night domain adaptation on the SHIFT dataset. Faster RCNN is used as the Source and is trained on the labeled daytime subset. Where "D2N" denotes the detector specifically designed for day-to-night DAOD tasks. For a fair comparison, all the compared methods employ the self-training framework with Faster RCNN as their base detectors.**

Method	Venues	Detector	D2N	person	car	truck	bus	mcycle	bicycle	mAP
Source [13]	NeurIPS'15	FRCNN	No	40.4	44.5	49.9	53.7	14.3	46.7	41.6
DA FR [3]	CVPR'18	FRCNN	No	43.0	48.8	47.8	52.1	19.9	55.8	43.7
UMT [6]	CVPR'21	FRCNN	No	7.7	47.5	18.4	46.8	16.6	49.2	31.1
AT [12]	CVPR'22	FRCNN	No	25.8	33.0	54.7	49.5	20.7	52.3	38.9
2PCNet [11]	CVPR'23	FRCNN	Yes	51.4	54.6	54.8	56.6	23.9	54.2	49.1
CoS [9]	ICME'24	FRCNN	Yes	50.8	56.0	57.2	64.5	22.2	55.5	51.0
ISP-Teacher [16]	AAAI'24	FRCNN	Yes	51.6	59.1	<b>58.7</b>	62.3	24.1	<b>58.3</b>	52.4
SOCCEr(ours)	-	FRCNN	No	<b>58.8</b>	<b>59.4</b>	58.3	<b>66.6</b>	<b>29.6</b>	57.7	<b>55.1</b>

## C.1 Dataset

SHIFT [14] is a simulated autonomous driving dataset that contains scenes in various environments, and it includes discrete shifts (e.g. urban, village, and rural) and continuous shifts (e.g. daytime to night) in cloudiness, rain, and fog weather. Following the previous methods [11, 16], we use images with the "day" and "night" labels as our source and target data respectively, which contain 19,452 daytime images and 8,497 nighttime images for training, and 1,200 nighttime images for validation. Additionally, we also ensure that the weather label is "clear" to isolate other weather conditions from the evaluation.

## C.2 Evaluation

As illustrated in Table 2, our SOCCEr outperforms all other methods, even the latest Day-to-Night-SOTA: ISP-Teacher [16]. In addition, we maintain an absolute advantage in the detection performance of the "person" category with 7.2% mAP than SOTA. Rare classes such as "bus" and "mcycle" also show a significant improvement. This shows that our strategy is effective in enhancing the network's context reasoning capacity even in the night scene.

## D PSEUDO-CODE FOR SOCCEr

In Algorithm 1, we present a pseudo-code pipeline of our stochastic context consistency reasoning (SOCCEr) network. The core process is 5-9. We also provide detailed hyper-parameters in Table 3.

## E QUANTITATIVE RESULTS

In Figure 7, we provide more detailed performance of our SOCCEr with our baseline SADA [4], MIC [8], and CMT [1] on Cityscapes to Foggy Cityscapes. We observe that small object detection is still a challenging direction. The mAP of small objects is all below 10% among the four approaches, and CMT is only 1.8%. Compared with the baseline SADA and SOTA CMT, SOCCEr leverages the context relation information to detect small objects more accurately. For MIC, our SOCCEr can reduce unfair penalization and avoid error accumulation for detecting small objects. As illustrated in the performance of mAPm, SOCCEr can detect medium objects better than others with a significant gain of 6.3% over SADA. We also observe that mAPl is very similar to mAP0.5, indicating that

### Algorithm 1: The training pipeline of SOCCEr

**Input:** Object detectors: Student  $\mathbb{S}(\cdot; \theta_s)$ , Teacher  $\mathbb{T}(\cdot; \theta_t)$  and  $\theta_s/\theta_t$  are the model parameters of Student and Teacher. Domain discriminator  $\mathbb{D}(\cdot; \theta_d)$  and domain labels  $d_s, d_t$ . Total number of iterations:  $T_{max\_iterations}$ . Hyper-parameters: Momentum  $\alpha$  in EMA, pseudo label confidence threshold  $\delta$ , and learning rate  $\eta$ .

**Output:** Student  $\mathbb{S}(\cdot; \theta_s)$ , Teacher  $\mathbb{T}(\cdot; \theta_t)$  after the training of stochastic context consistency reasoning.

**for**  $iteration \leftarrow 1$  **to**  $T_{max\_iterations}$  **do**

// 1. Load data mini-batch

Sample source batch  $B_s = \{(x_s^i, b_s^i, c_s^i)\}_{i=1}^{N_s} \in \mathcal{D}_s(X_s, B_s, C_s)$

Sample target batch  $B_t = \{(x_t^i)\}_{i=1}^{N_t} \in \mathcal{D}_t(X_t)$

// 2. Compute base losses

$B'_s, C'_s = \mathbb{S}(X_s; \theta_s)$

$\mathcal{L}_{sup} = \mathcal{L}_{sup}^{reg}(B'_s, B_s) + \mathcal{L}_{sup}^{cls}(C'_s, C_s)$

$\mathcal{L}_{adv} = \max_{\mathbb{D}} \min_{\mathbb{D}} \mathbb{D}(\mathcal{L}_{BCE}(X_s, X_t, d_s, d_t); \theta_d); \theta_s)$

// 3. Update Teacher by EMA

$\theta_t \leftarrow \alpha \theta_t + (1 - \alpha) \theta_s$

// 4. Generate pseudo labels by Teacher

$\hat{B}, \hat{C} = \text{Filter}(\mathbb{T}(X_t; \theta_t), \delta)$

**if**  $\hat{B}, \hat{C} \neq \emptyset$  **then**

// 5. Generate stochastic complementary masked images

$X_t^M = \{\mathcal{M}(x_i)\}_{i=1}^{N_t}$      $X_t^{\bar{M}} = \{\bar{\mathcal{M}}(x_i)\}_{i=1}^{N_t}$

// 6. Student reasoning stage

$B_t^M, C_t^M = \mathbb{S}(X_t^M; \theta_s)$      $B_t^{\bar{M}}, C_t^{\bar{M}} = \mathbb{S}(X_t^{\bar{M}}; \theta_s)$

// 7. Compute Inter-CCR loss

$\mathcal{L}_{consis}^{Inter} = \mathcal{L}_{Inter}^{reg}(B_t^M, B_t^{\bar{M}}) + \mathcal{L}_{Inter}^{cls}(p^M, p^{\bar{M}})$

// 8. Compute Intra-CCR loss

$\mathcal{L}_{consis}^{Intra} = \mathcal{L}_{Intra}^{reg}(B_t^M, B_t^{\bar{M}}, \hat{B}) + \mathcal{L}_{Intra}^{cls}(p^M, p^{\bar{M}}, \hat{C})$

▷ Where  $p^M/\bar{M}$  denotes probability vector of  $C_t^M/\bar{M}$ .

// 9. The optimization objective

$\mathcal{L} = \lambda_0 \cdot \mathcal{L}_{sup} + \lambda_1 \cdot \mathcal{L}_{consis}^{Inter} + \lambda_2 \cdot \mathcal{L}_{consis}^{Intra} + \lambda_3 \cdot \mathcal{L}_{adv}$

**else**

// 10. If the predictions of the Teacher are not sufficient to generate pseudo labels, then only the base losses are optimized

$\mathcal{L} = \lambda_0 \cdot \mathcal{L}_{sup} + \lambda_3 \cdot \mathcal{L}_{adv}$

**end if**

Take SGD step:  $\theta_s = \theta_s - \eta \nabla_{\theta_s} \mathcal{L}$

**end for**

**Table 3: Detailed hyper-parameters of SOCCER for each benchmark. "City2BDD" denotes Cityscapes to BDD100k(daytime), "Sim2City" denotes Sim10k to Cityscapes(car), and "City2Foggy" denotes Cityscapes to Foggy Cityscapes(0.02)**

Hyperparameter	Description	City2BDD	City2Sim	City2Foggy
$N_c$	Number of shared cross-domain categories	7	1	8
$\alpha$	EMA update ratio	0.9	0.9	0.9
$\delta$	confidence threshold of pseudo labels	0.8	0.8	0.8
$\lambda_0$	Weight for Supervised Loss	1	1	1
$\lambda_1$	Weight for Inter-CCR Loss	1	1	1
$\lambda_2$	Weight for Intra-CCR Loss	1	1	1
$\lambda_3$	Weight for Image/Instance-Level Adversarial Loss	0.025/0.1	0.025/0.1	0.025/0.1
$b$	Mask patch size of SCM	32	32	32
$\sigma$	Transition point of Huber Loss	0.6	0.6	0.6
$\tau$	IoU matching threshold of Inter-CCR	0.75	0.75	0.75
$r$	Mask ratio of SCM	0.5	0.5	0.5
$lr$	Initial learning rate	0.0025	0.0025	0.0025
-	Learning rate weight decay	0.0001	0.0001	0.0001
-	Total training iterations	60000	60000	60000

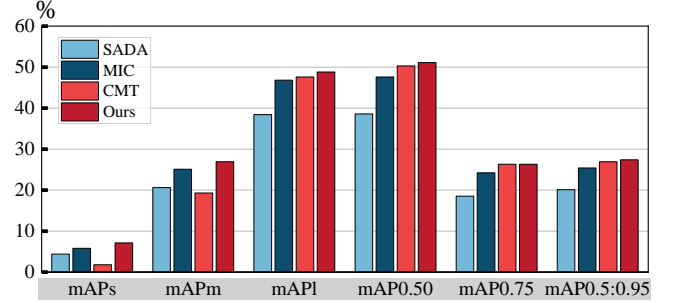
the performance of detectors is dominated by the detection performance of large objects. Our SOCCER can balance the detection performance for small, medium, and large objects.

## F QUALITATIVE RESULTS

In Figure 8, we present more qualitative results of our SOCCER with SOTA: MIC [8] and CMT [1] on the task Cityscapes  $\rightarrow$  Foggy Cityscapes. As shown in this figure, MIC and CMT struggle with accurate car localization, whereas SOCCER precisely determines car positions even in the presence of occlusion (rows 1 and 3). By leveraging context knowledge within the images, SOCCER can accurately distinguish confusing categories in foggy weather (rows 5, 7, 8, and 10). Compared to MIC, the stochastic complementary masking module in SOCCER effectively avoids the unfair penalization of the single mask strategy on small objects, thereby improving the recognition rate for small objects (rows 2 and 10). Compared to CMT, SOCCER exhibits a lower false positive ratio, enabling accurate differentiation between foreground and background (rows 4 and 5).

## G DETECTION RESULTS UNDER DIFFERENT TRAINING ITERATION

In this section, we provide the detection results of our method across different iterations. In Figure 9, we show the detection results in 15k, 30k, 45k, and final 60k iterations on the task Cityscapes  $\rightarrow$  Foggy Cityscapes. Similarly, in Figure 10 and Figure 11, we show the detection results in 10k, 20k, 30k, and final 60k iterations on the tasks BDD100k  $\rightarrow$  Cityscapes and Sim10k  $\rightarrow$  Cityscapes. Due to the significant domain gap between synthetic and real-world data, we additionally present the detection results at different iterations on the source domain Sim10k for the Sim10k  $\rightarrow$  Cityscapes task, as shown in Figure 12. As the iteration progresses, the true positive rate of the model continues to increase, enabling the recognition of some easily-confused categories. Moreover, the localization of objects becomes progressively more accurate.



**Figure 7: Quantitative results. We compare the detection results of SADA [4], MIC [8], CMT [1], and our SOCCER on Cityscapes to Foggy Cityscapes. Where mAPs, mAPm, and mAPl denotes the mAP of small, medium, and large objects.**

## H DETECTION RESULTS UNDER TWO VIEWS

We also report the detection results of SOCCER for two views  $\{\mathcal{V}_1, \mathcal{V}_2\}$  under different masked patch sizes, aiming to validate the robustness of our method under varied masking conditions. In Figure 13, we observe that excessively small masked patch sizes can interfere with the predictions, leading to incorrect predicted categories. Conversely, overly large mask sizes may result in missed detection due to complete object coverage. For a moderate masked patch size, our network demonstrates the ability to detect results similar to the original image accurately. We also observe that, in certain situations, the network is capable of predicting completely occluded objects.



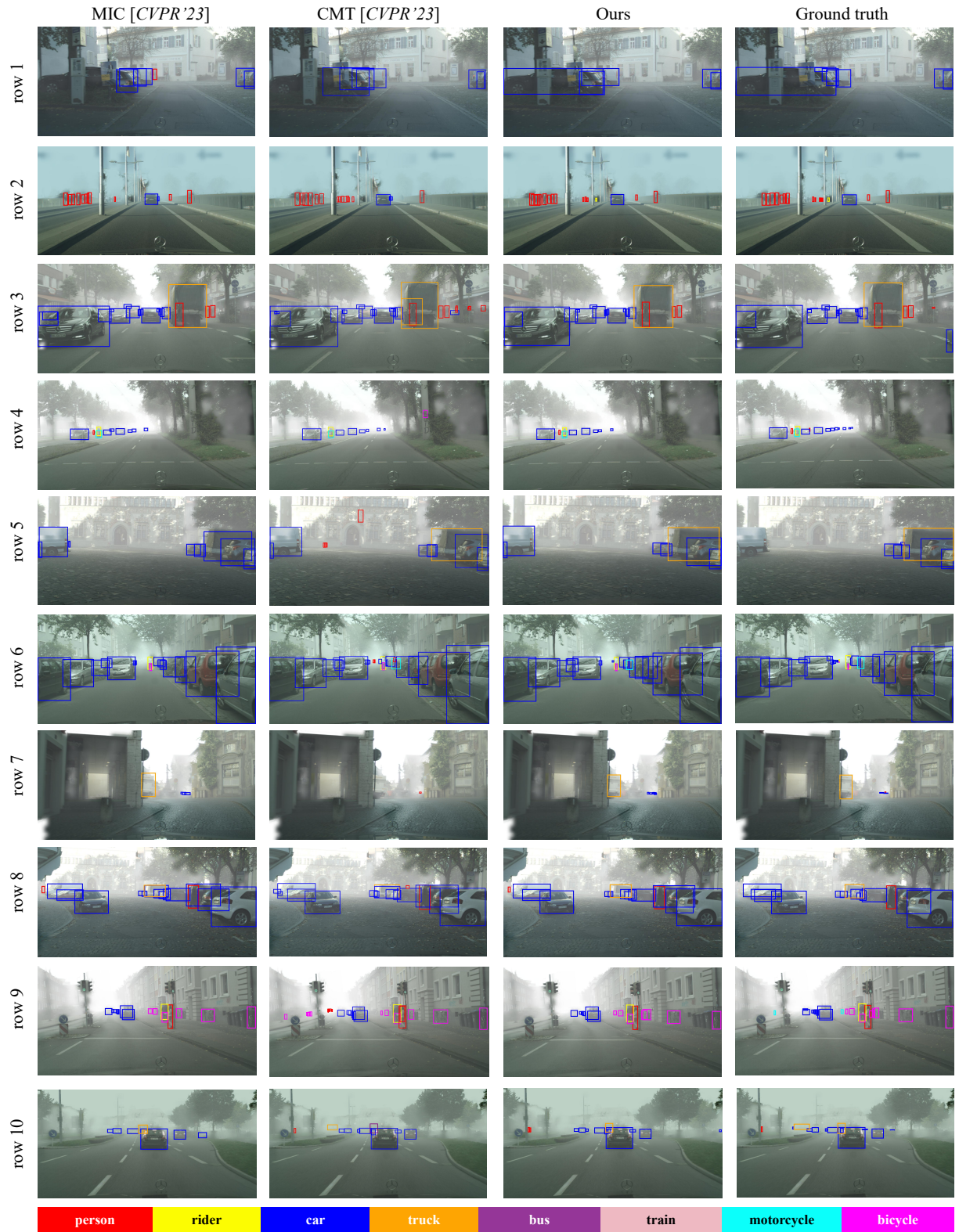


Figure 8: The qualitative results on the task Cityscapes → Foggy Cityscapes, in which SOTA MIC [8] and CMT [1], as well as our method are evaluated.

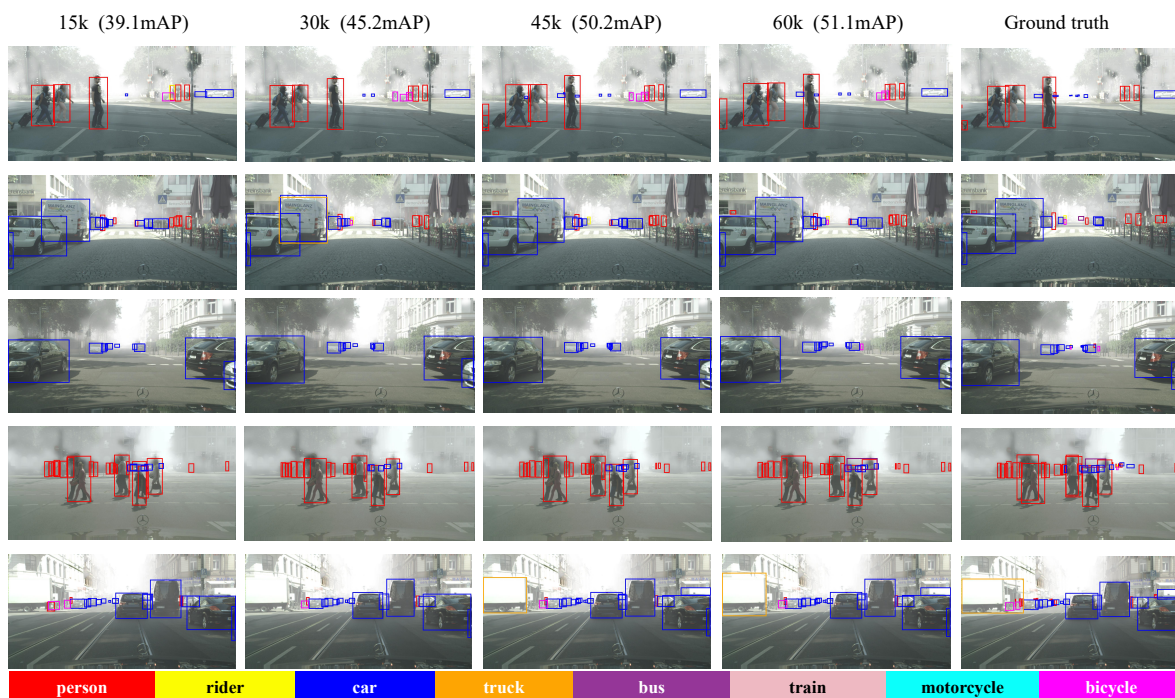


Figure 9: Iterative results analysis on the task Cityscapes → Foggy Cityscapes.



Figure 10: Iterative results analysis on the task Cityscapes → BDD100k.



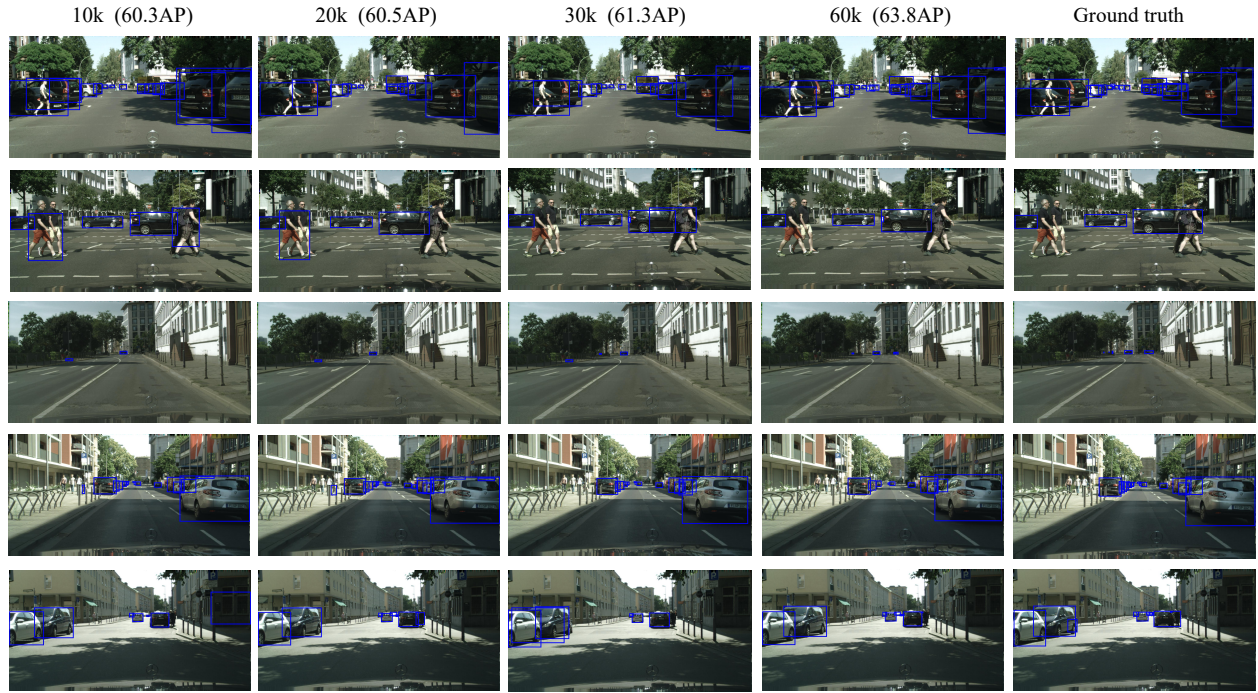


Figure 11: Iterative results analysis on the task Sim10k → Cityscapes, and this task only detects cars.

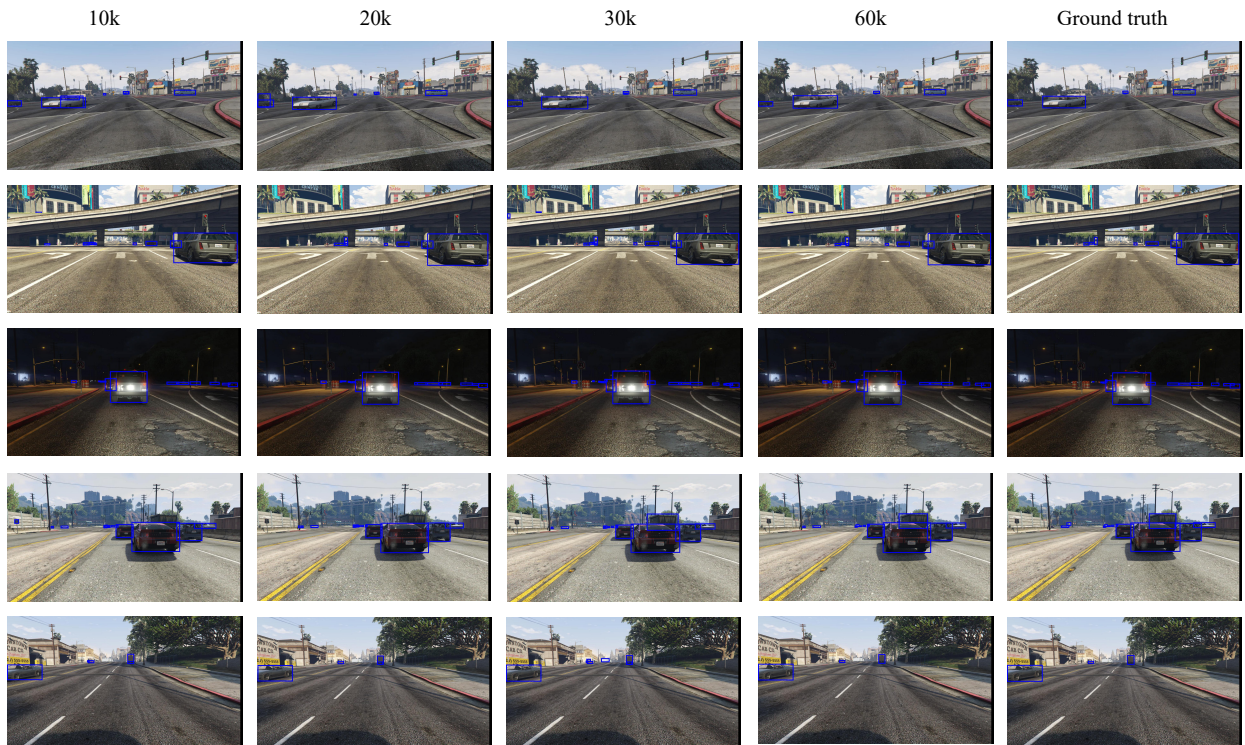


Figure 12: Iterative results analysis on the Sim10k dataset (Sim10k → Cityscapes), and only report the detection results of cars.



Figure 13: Stochastic complementary masking detection results. We compare the detection results of SOCCER under different masked patch sizes.



## REFERENCES

- [1] Shengcao Cao, Dhiraj Joshi, Liang-Yan Gui, and Yu-Xiong Wang. 2023. Contrastive Mean Teacher for Domain Adaptive Object Detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23839–23848.
- [2] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. 2020. Gridmask data augmentation. *arXiv preprint arXiv:2001.04086* (2020).
- [3] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. 2018. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3339–3348.
- [4] Yuhua Chen, Haoran Wang, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. 2021. Scale-aware domain adaptive faster r-cnn. *International Journal of Computer Vision* 129, 7 (2021), 2223–2243.
- [5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3213–3223.
- [6] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. 2021. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4091–4101.
- [7] Terrance DeVries and Graham W Taylor. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552* (2017).
- [8] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. 2023. MIC: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11721–11732.
- [9] Yuan Jicheng, Le-Tuan Anh, Hauswirth Manfred, and Le-Phuoc Danh. 2024. Cooperative Students: Navigating Unsupervised Domain Adaptation in Nighttime Object Detection. *2024 IEEE International Conference on Multimedia and Expo (ICME)* (2024).
- [10] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. 2016. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv preprint arXiv:1610.01983* (2016).
- [11] Mikhail Kennerley, Jian-Gang Wang, Bharadwaj Veeravalli, and Robby T Tan. 2023. 2PCNet: Two-Phase Consistency Training for Day-to-Night Unsupervised Domain Adaptive Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11484–11493.
- [12] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. 2022. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7581–7590.
- [13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).
- [14] Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. 2022. SHIFT: a synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21371–21382.
- [15] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. 2020. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2636–2645.
- [16] Yin Zhang, Yongqiang Zhang, Zian Zhang, Man Zhang, Rui Tian, and Mingli Ding. 2024. ISP-Teacher: Image Signal Process with Disentanglement Regularization for Unsupervised Domain Adaptive Dark Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 7387–7395.
- [17] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 13001–13008.