RaidaR: A Rich Annotated Image Dataset of Rainy Street Scenes Supplementary Material

Jiongchao Jin[1,2 *] Arezou Fatemi[1 *] Wallace Michel Pinto Lira[1] Fenggen Yu[1]
Biao Leng[2] Rui Ma[3,1 †] Ali Mahdavi-Amiri[1] Hao Zhang[1]
[1]Simon Fraser University [2]Beihang University [3]Jilin University

1. Introduction

Here, we provide more details about our masked image-to-image translation frameworks discussed in the main paper, along with additional results of the semantic segmentation and instance segmentation. A gallery of RaidaR image samples and their semantic and instance segmentation ground truth is also shown in Figure 5.

2. Masked Image-to-Image Translation Details

In this section, we first describe the loss functions of our masked CycleGAN and masked GANHopper. Then, we provide more implementation details and a preliminary comparison of inference time between the masked and the original image translation.

2.1. Masked CycleGAN

We have introduced a few adjustments to the original loss function of CycleGAN [3] to respect semantic segments and control their influence on the final translation. Here, Equation 1 describes the loss function for our masked CycleGAN.

$$\underset{cyc}{\mathcal{L}} = \sum_{i=1}^{m} \frac{\lambda_i}{p_i} |M_i \cdot F(G(u)) - M_i \cdot u|, \tag{1}$$

In our notation, u and v respectively represent samples in domains U and V and m is the number of labels in a segmentation mask. M_i is the binary mask for the i-th label, p_i is the number of pixels with label i, and λ_i is the weight for the i-th label. λ_i controls the importance of each label and p_i tries to give more influence to small but important regions. Note that G and F are the generators that translate images from domain U to V and vice-versa. We have chosen $m = 7$ labels in our segmentation masks: road, traffic lights, vegetation, sky, people, vehicles, and other. We set their respective λ_i to $2, 3, 1, 0.2, 1, 2, 1$ in our model to distinguish the importance of different categories.

2.2. Masked GANHopper

We further modify masked CycleGAN to masked GANHopper, whose loss is shown in Equation 2 as:

$$\underset{loss}{\mathcal{L}} = \gamma \underset{\text{cyc}}{\mathcal{L}} + \epsilon \underset{\text{adv}}{\mathcal{L}} + \delta \underset{\text{dom}}{\mathcal{L}} + \zeta \underset{\text{smooth}}{\mathcal{L}}, \tag{2}$$

same as the original GANHopper [1]. The cycle loss $\mathcal{L}_{\text{cyc}}$ and the smoothness loss $\mathcal{L}_{\text{smooth}}$ are both adapted to account for semantic segmentation masks provided in the dataset we propose in this paper, as shown in Equations 3 and 4. Let h represent the number of hops and G_n represent the transformation G that is applied consecutively n times, per GANHopper's framework. The cycle loss and the smoothness loss in Equations 3 and 4 that are defined for domain U have analogous counterparts for domain V.

$$\underset{cyc}{\mathcal{L}} = \sum_{n=1}^{h} \sum_{i=1}^{m} \frac{\lambda_i}{p_i} |M_i \cdot F(G_n(u)) - M_i \cdot G_{n-1}(u)| \tag{3}$$

$$\underset{smooth}{\mathcal{L}} = \sum_{n=1}^{h} \sum_{i=1}^{m} \frac{\lambda_i}{p_i} |M_i \cdot G_n(u) - M_i \cdot G_{n-1}(u)| \tag{4}$$

The general purpose of these losses is the same as in the original CycleGAN and GANHopper. While the cycle loss aims to enforce the cycle consistency from one hop to the next, the smoothness loss aims to preserve the input image as much as possible. These hop translations are represented as the functions G and F in this section. These two losses have conflicting goals that tend to reach at an equilibrium as the network converges, which enables it to find intermediary domains to facilitate the translation process.

We trained our model with the same dataset configuration as masked CycleGAN but for 24 epochs. Since we optimize the parameters in each hop, GANHopper needs fewer epochs to train. We set $\gamma = 10, \epsilon = 1, \delta = 1, \zeta = 1$ in our model. All other variables have the same values as in masked CycleGAN.

[*] Jiongchao Jin and Arezou Fatemi contribute equally to this work.
[†] Rui Ma is the corresponding author.

	Masked	Original
Generator	4.9 - 5.9 ms	3.8 - 4.8 ms
Discriminator	0.7 - 0.9 ms	0.5 - 1.5 ms

Table 1: Inference time for CycleGAN and masked CycleGAN. This table also represents inference for each hop in masked and original GANHopper. This is because the generator and discriminator from GANHopper have the same architecture as CycleGAN.

2.3. More Details and Inference Time Comparison

For training masked CycleGAN and masked GANHopper, we used 20,791 rainy and 15,925 sunny images from RaidaR for 70 epochs. Note that in this experiment, our goal is to train a generalizable model for the masked unpaired image-to-image translation. Hence, we chose to use the larger set of RaidaR images with Vpred segmentation masks instead of our ground truth segmentation which is on a smaller set of RaidaR images (5,000 rainy and 4,085 sunny). For testing and also generating the synthetic RaidaR dataset, we used the model trained above to a separate set of images with our ground truth segmentation masks.

We present a comparison of inference time for CycleGAN and masked CycleGAN in Table 1. The inference times of GANHopper and masked GANHopper also demonstrate similar results since each hop of GANHopper uses the same architecture as CycleGAN. The results show that masked versions of these networks require a slightly longer processing time, but they are still fast and efficient.

3. More Results

Here, we provide additional results for segmentation and masked image-to-image translation. We also provide sample data of our rainy and sunny images along with their semantic and instance segmentation. Figures 1 and 2 show the results of Vpred [4] trained on different training datasets and tested on BDD and RaidaR. The results share similar observations as the corresponding experiment conducted using HMSA [2]. Detailed explanations of the results can be found in the main paper. Figure 3 shows the comparison with the de-raining based segmentation results. It can be observed that the segmentation model trained directly on RaidaR outperforms the ones trained on Cityscapes or the de-rained version of RaidaR. This verifies the usefulness of RaidaR for providing annotated rainy images to facilitate the segmentation tasks on images with rainy artifacts. Figure 4 displays more RaidaR sample images and the corresponding results of image-to-image translation using masked GANHopper. In the results, the overall style of images are successfully translated into the target domain while the colors of important categories (e.g., traffic light, cars) are well preserved. Finally, Figure 5 shows some samples of rainy images along with semantic and instance segmentation ground truth in RaidaR.

References

[1] Wallace Lira, Johannes Merz, Daniel Ritchies, Daniel Cohen-Or, and Hao Zhang. GANHopper: Multi-hop gan for unsupervised image-to-image translation. In *ECCV*, 2020. 1

[2] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*, 2020. 2

[3] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. ICCV*, 2017. 1

[4] Yi Zhu, Karan Sapra, Fitsum A Reda, Kevin J Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *IEEE CVPR*, 2019. 2

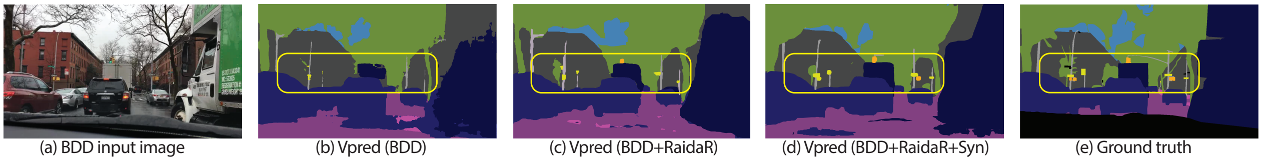

(a) BDD input image (b) Vpred (BDD) (c) Vpred (BDD+RaidaR) (d) Vpred (BDD+RaidaR+Syn) (e) Ground truth

Figure 1: Qualitative comparisons of Vpred trained on different training configurations and tested on BDD. Highlighted regions by the yellow boxes show that after adding RaidaR or RaidaR + Syn, the model is able to produce finer details (e.g., traffic signs, trees, etc).

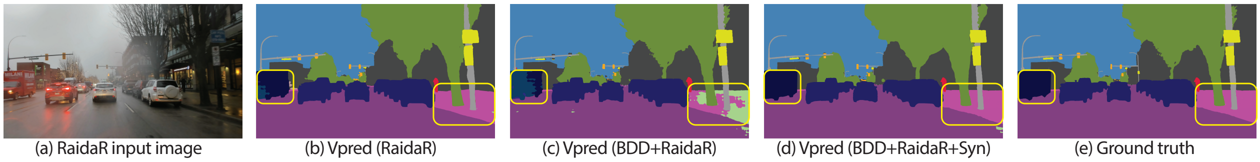

(a) RaidaR input image (b) Vpred (RaidaR) (c) Vpred (BDD+RaidaR) (d) Vpred (BDD+RaidaR+Syn) (e) Ground truth

Figure 2: Qualitative comparisons of Vpred trained on different training configurations and tested on RaidaR. Highlighted regions by the yellow boxes show that the model can produce satisfactory results on rainy images when trained on RaidaR, but inferior results when simply combined with BDD. After adding the synthetic images (Syn), the model can produce improved results.

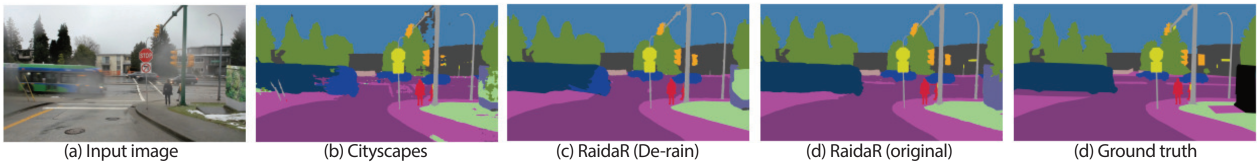

(a) Input image (b) Cityscapes (c) RaidaR (De-rain) (d) RaidaR (original) (d) Ground truth

Figure 3: Qualitative comparisons of Vpred trained on Cityscape, RadiaR (De-rain) and RaidaR (original).

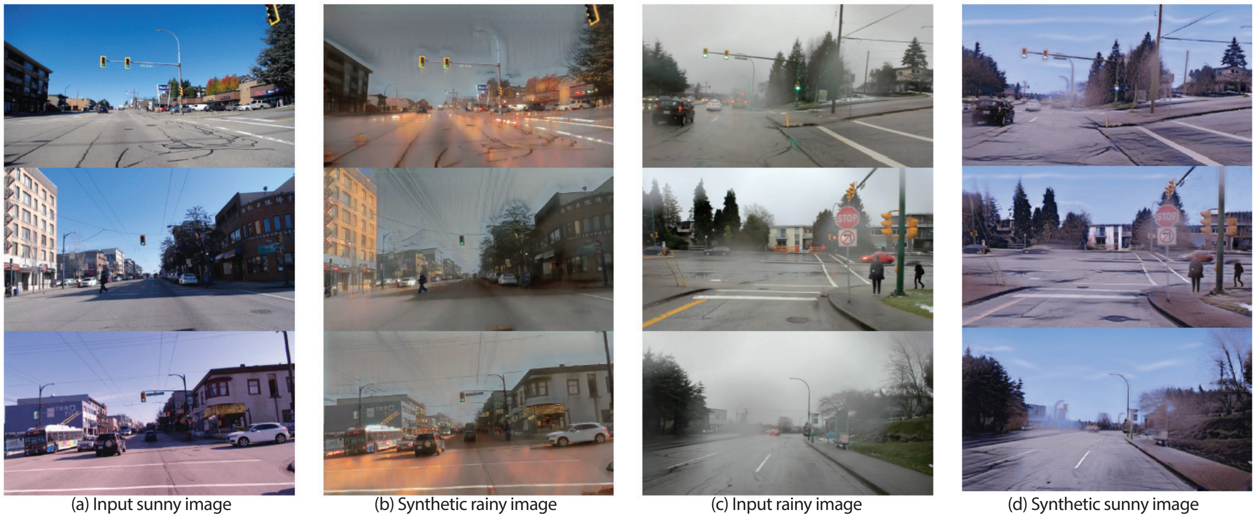
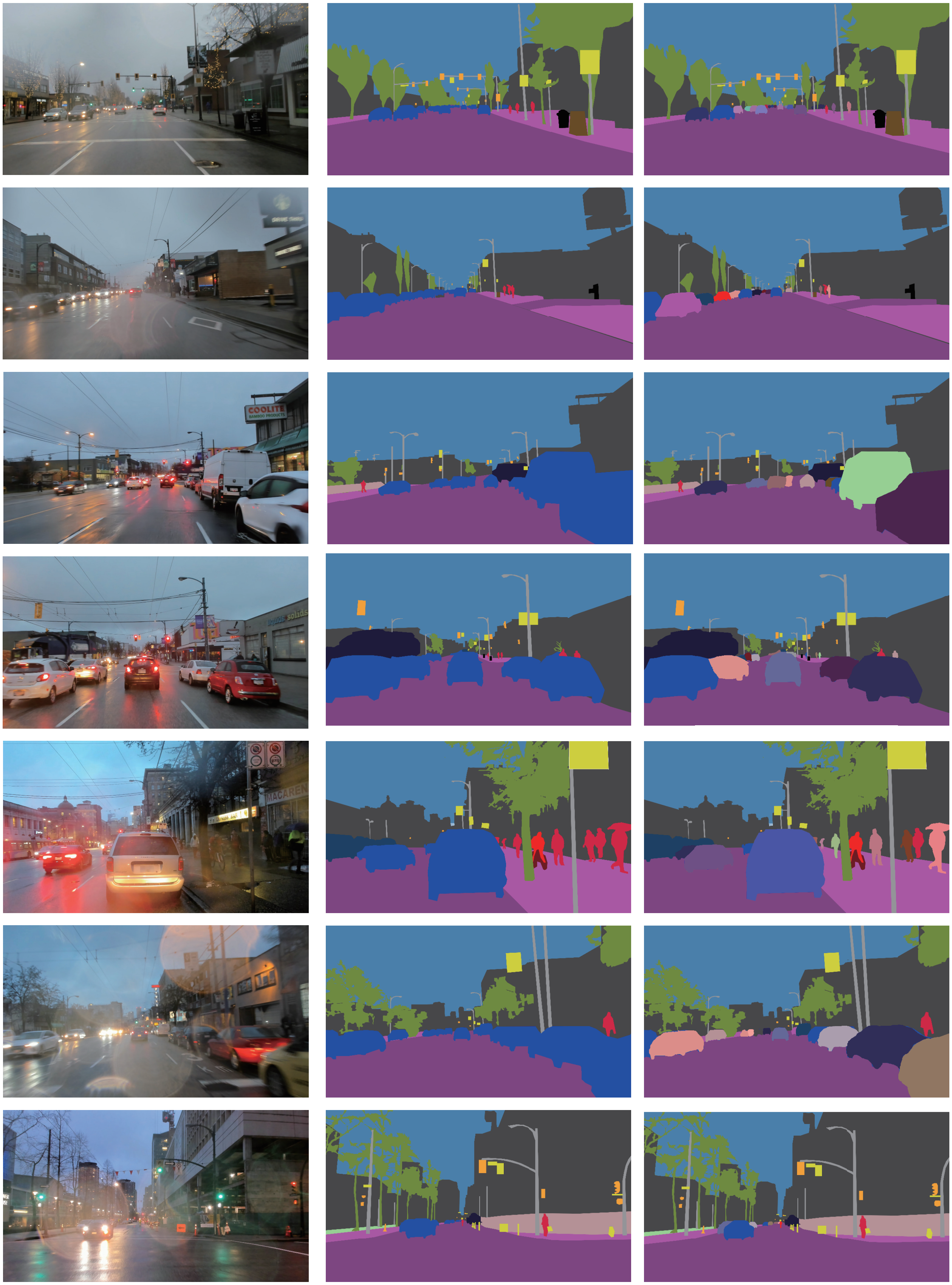

(a) Input sunny image (b) Synthetic rainy image (c) Input rainy image (d) Synthetic sunny image

Figure 4: More results of image translation using masked GANHopper trained on the RaidaR dataset.



(a) Rainy image (b) SSeg ground truth (c) ISeg ground truth

Figure 5: Samples of RaidaR images and their semantic and instance ground truth masks.