Supplementary of
Model-based occlusion disentanglement for image-to-image translation

Fabio Pizzati[1,2], Pietro Cerri[2], and Raoul de Charette[1⋆]

[1] Inria, Paris, France
{fabio.pizzati, raoul.de-charette}@inria.fr
[2] VisLab, Parma, Italy
pcerri@ambarella.com

This file provides the reader with additional information on training pipeline (Sec. 1) and qualitative results (Sec. 2) of our method. We refer to the supplementary video for better visual understanding.

1 Training details

We experienced the perceptual VGG loss having a major impact on the obtained image quality in MUNIT [3]. With the default weight in [1] the perceptual loss prevents drastic changes to the input image, and this led to unsatisfying results on clear $\mapsto$ rain. On the other hand, avoiding to use a perceptual loss leaded to artifacts generation. Hence, we set `vgg_w` to 0.1 on all datasets. For a fair comparison, we train the baseline with the same configuration. Rather than cropping the input images, we train our network on entire images, downsampled 4x, in order to maintain contextual information (needed for example to render realistic reflections). We train our disentangled MUNIT for 150000 iterations. Other parameters follow the default MUNIT configuration [1]. Our training routine (refer to main paper Sec. 4.1) is represented in Fig. 1.

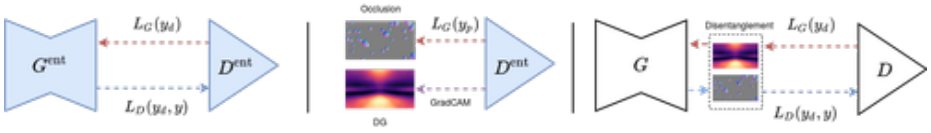

Fig. 1: Our three-stages pipeline. First, we obtain an entangled discriminator D^{ent} with a naive training, along with an entangled generator G^{ent} which is not used. Then, we use it to extract occlusion parameters and DG. Finally, we retrain the generator injecting occlusions.

⋆ Corresponding author.

2 Additional qualitative results

In Figs. 2 and 3, we provide more qualitative results for the clear $\mapsto$ rain translation on the nuScenes [2] dataset, and the comparison with several baselines such as MUNIT [3], CycleGAN [9], U-GAT-IT [4], DRIT [5] and AttentionGAN [7]. The experiment is fully detailed in Sec. 4.2 of the article. Our method is the only able to learn the scene transformation in a disentangled manner (*Ours - disentangled* rows), while simultaneously enabling realistic modeling of occlusions as in the target dataset (*Ours - target*) or in different styles (*Ours - dashcam 1*, *Ours - dashcam 2*). Instead, in Figs. 4 and 5 we report additional images for performances with general occlusions. We disentangle soiling in the WoodScape [8] dataset and learn the clean_gray $\mapsto$ color transformation. Finally, we add synthetic occlusions in Synthia [6] and learn clear $\mapsto$ snow. Again, we refer to the article, Sec. 4.3, for details. Here, we also report images with injected occlusions (*Ours - target*).

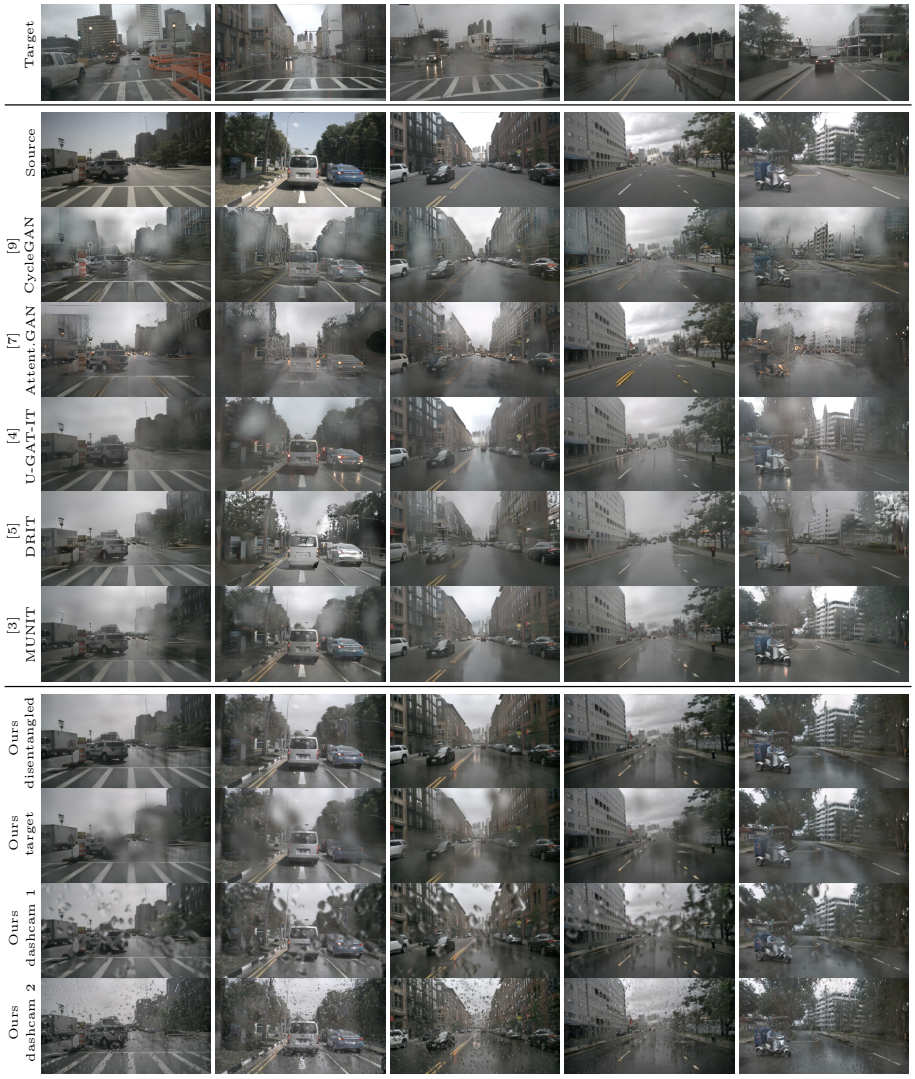

Fig. 2: Additional images for the qualitative comparison between the baselines and our approach on the nuScenes clear $\mapsto$ rain transformation.

Fig. 3: Additional images for the qualitative comparison between the baselines and our approach on the nuScenes clear $\mapsto$ rain transformation.

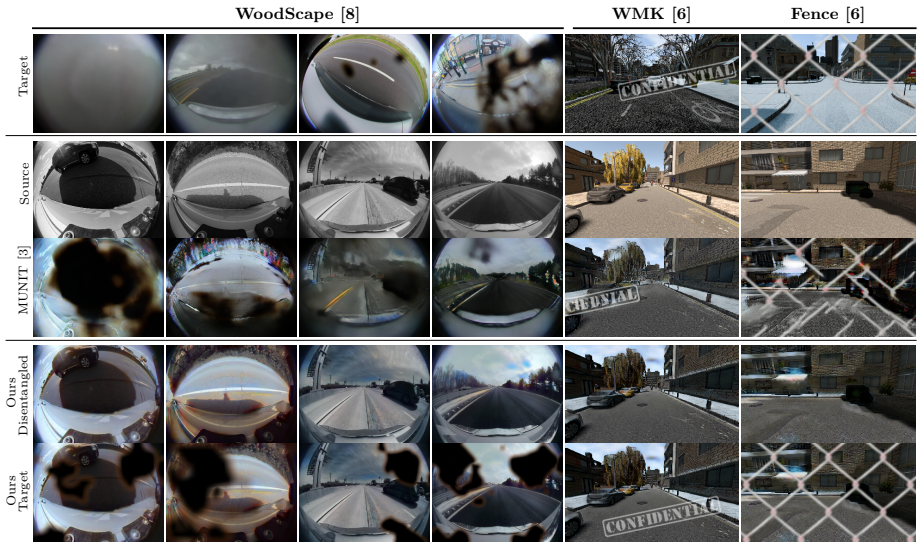

Fig. 4: Additional qualitative results on general occlusions.

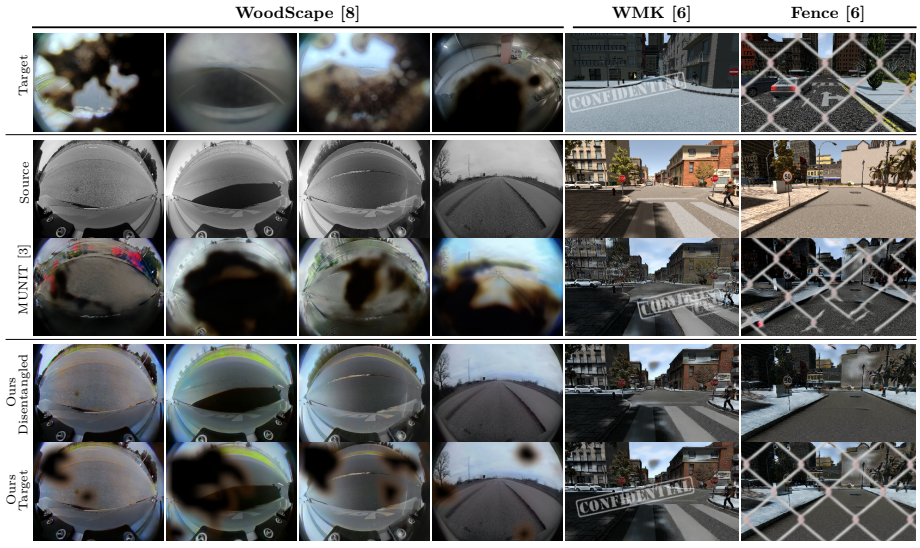

Fig. 5: Additional qualitative results on general occlusions.

References

1. Munit configuration. https://github.com/NVlabs/MUNIT/blob/master/configs/synthia2cityscape_folder.yaml
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020)
3. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: ECCV (2018)
4. Kim, J., Kim, M., Kang, H., Lee, K.: U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In: ICLR (2020)
5. Lee, H.Y., Tseng, H.Y., Mao, Q., Huang, J.B., Lu, Y.D., Singh, M., Yang, M.H.: Drit++: Diverse image-to-image translation via disentangled representations. arXiv preprint arXiv:1905.01270 (2019)
6. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: CVPR (2016)
7. Tang, H., Xu, D., Sebe, N., Yan, Y.: Attention-guided generative adversarial networks for unsupervised image-to-image translation. In: International Joint Conference on Neural Networks (IJCNN) (2019)
8. Yogamani, S., Hughes, C., Horgan, J., Sistu, G., Varley, P., O'Dea, D., Uricár, M., Milz, S., Simon, M., Amende, K., et al.: Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In: ICCV (2019)
9. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: CVPR (2017)