Adverse Weather Image Translation with Asymmetric and Uncertainty-aware GAN

-Supplementary material-

Jeong-gi Kwak
kjk8557@korea.ac.kr
Youngsaeng Jin
youngsjin@korea.ac.kr
Yuanming Li
lym7499500@korea.ac.kr
Dongsik Yoon
kevinds1106@korea.ac.kr
Donghyeon Kim
kis6470@korea.ac.kr
Hanseok Ko
hsko@korea.ac.kr

Korea Univerity
Seoul, Korea

6 Appendix

In this section, we supplement additional analysis and experimental results that are not presented in the main paper. We first describe the details of our architecture for reproducibility (Sec. 6.1) and then analyze the effectiveness of the uncertainty-aware cycle consistency loss (Sec. 6.2). In addition, we conduct broader analysis of our method, i.e., experiments on higher resolution (512×1024 and failure case (Sec. 6.3). Finally, we present additional comparison results (Sec. 6.4) and extra qualitative results of our model (Sec. 6.5).

6.1 Implementation

We report details of each module of our model and figures are depicted in Fig. 1. In the following, we explain each module.

Encoder The encoders of two domains $\{G^E_{\mathcal{A}\to\mathcal{B}}, G^E_{\mathcal{B}\to\mathcal{A}}\}$ have same network architecture. They consist of three convolutional layers and four residual blocks [2] with dilated convolution [3] (D.Resblk). Therefore, an input image, i.e., $x_{\mathcal{A},\mathcal{B}} \in \mathbb{R}^{256\times512\times3}$, is converted to encoded feature with the output size in $\mathbb{R}^{64\times128\times256}$. In addition, we utilize Instance Normalization [6] (IN). in all layers of the encoder.

T-net As mentioned in main text, feature transfer network (T-net) is inserted in $G_{\mathcal{A}\to\mathcal{B}}$. It consists of four residual blocks (Resblk) thus the size of input and output is same as in $\mathbb{R}^{64\times128\times256}$.

Decoder The two decoders $\{G^D_{\mathcal{A}\to\mathcal{B}}, G^D_{\mathcal{B}\to\mathcal{A}}\}$ have same structure except for the last two layers. They have a symmetrical structure with the encoders, thus the input feature $\in \mathbb{R}^{64\times128\times256}$ is transformed to RGB output image $\in \mathbb{R}^{256\times512\times3}$ by transposed convolution (Deconv). Unlike $G^D_{\mathcal{B}\to\mathcal{A}}$, $G^D_{\mathcal{A}\to\mathcal{B}}$ has an additional branch that generates the uncertainty map σ. With the mid-feature $\in \mathbb{R}^{128\times256\times128}$ in decoder, the branch outputs the uncertainty map$\in \mathbb{R}_+{}^{256\times512}$ by including Softplus in the last layer.

Discriminator The discriminators $\{D_\mathcal{A}, D_\mathcal{B}\}$ have the form of multi-scale [7] and PatchGAN [3] discriminators. The resolution of the output activations are in $\mathbb{R}^{16\times32}$ and $\mathbb{R}^{8\times16}$. As similar with the generators, we use Instance Normalization in all layers of each discriminator except for the last layer.

6.2 Analysis of uncertainty-aware cyclic loss

To demonstrate the effectiveness of the uncertainty-aware cycle consistency loss $\mathcal{L}^\mathcal{A}_{cyc}$, we analyze the role of the loss in training. We compare the translated images ($\mathcal{A}\to\mathcal{B}$), the reconstructed images ($\mathcal{A}\to\mathcal{A}$) and the cyclic reconstructed images ($\mathcal{A}\to\mathcal{B}\to\mathcal{A}$) of two variants of our model, i.e., with $\mathcal{L}^\mathcal{A}_{cyc}$ (Ours) and without $\mathcal{L}^\mathcal{A}_{cyc}$ (Ours w.o. *un*.). In the latter case, we use the standard cycle consistency loss [11]. The results are presented in Fig. 2. As mentioned in the main paper, the cyclic reconstructed image is not obliged to possess artifacts same as that of the original if the disentanglement is conducted successfully. As shown in Fig. 2, the cyclic reconstructed image when $\mathcal{L}^\mathcal{A}_{cyc}$ is not in use has unnecessary artifacts or reflections. As a result, some of artifacts also appear in the transferred day image. However when using $\mathcal{L}^\mathcal{A}_{cyc}$, the problem is alleviated because the regions with artifacts or reflections have less confidence and thus they are removed clearly in the converted day image.

6.3 Broader analysis

6.3.1 Experiments on higher resolution

Although the resolution of train and test images in our method is 256×512, we additionally train our model with higher resolution images (512×1024). We use BDD100K dataset only because its original resolution is 720×1280 (Alderley: 260×640). We just added one more layer in each encoder, decoder and discriminator while keeping others (*e.g.* hyper-parameter and network architecture) unchanged. Although our model can translate adverse domain but it is not converged well so shows inferior visual quality and generates some artifact as shown in Fig. 3. We remain it for future work that finding proper hyper-parameters and network architecture to train high resolution images.

6.3.2 Failure case

We also analyze the failure case and limitation of our model. In Fig. 4, we show two examples of translation results (night $\to$ day) by our model. The regions of road or car that usually appear in dataset show satisfactory translation result. However, in the case of dark building

Generator

l	$G^E_{A\to B}$, $G^E_{B\to A}$
1	Conv(3, 64, 7, 1), IN, ReLU
2	Conv(64, 128, 3, 2), IN, ReLU
3	Conv(128, 256, 3, 2), IN, ReLU
4	D.Resblk(256, 256, 3, 1)
5	D.Resblk(256, 256, 3, 1)
6	D.Resblk(256, 256, 3, 1)
7	D.Resblk(256, 256, 3, 1)

Encoder

l	T-net
1	Resblk (256, 256, 3, 1)
2	Resblk (256, 256, 3, 1)
3	Resblk (256, 256, 3, 1)
4	Resblk (256, 256, 3, 1)

***T*-net**

l	$G^D_{B\to A}$
1	D.Resblk(256, 256, 3, 1)
2	D.Resblk(256, 256, 3, 1)
3	D.Resblk(256, 256, 3, 1)
4	D.Resblk(256, 256, 3, 1)
5	Deconv(256, 128, 3, 2), IN, ReLU
6	Deconv(128, 64, 3, 2), IN, ReLU
7	Conv(64, 3, 7, 1), Tanh

l	$G^D_{A\to B}$	
1	D.Resblk(256, 256, 3, 1)	
2	D.Resblk(256, 256, 3, 1)	
3	D.Resblk(256, 256, 3, 1)	
4	D.Resblk(256, 256, 3, 1)	
5	Deconv(256, 128, 3, 2), IN, ReLU	
6	Deconv(128, 64, 3, 2), IN, ReLU	Deconv(128, 64, 3, 2), IN, ReLU
7	Conv(64, 3, 7, 1), Tanh	Conv(64, 1, 7, 1), Softplus

Decoder

Discriminator

l	D_A, D_B
1	D.Conv(3, 64, 4, 2), IN, LReLU
2	D.Conv(64, 128, 4, 2), IN, LReLU
3	D.Conv(128, 256, 4, 2), IN, LReLU
4	D.Conv(256, 512, 4, 2), IN, LReLU
5	Conv(512, 1, 4, 1)

Figure 1: Details of proposed modules. Conv, Resblk, D.Resblk, Deconv denotes convolutional layer, residual block, residual block with dilated convolution and transposed convolution respectively. (c_{in}, c_{out}, k, s) denotes input channels, output channels, kernel size, stride respectively.

or completly dark areas, our model sometimes generates artifacts and unrealistic results such as “wooded building” or “tree on the road” (red boxes in translated results of Fig. 4). This is because our model is biased by dataset in that many images contain street trees. we believe that further work jointly exploiting region-based spatial attention methods with our model alleviates this problem.

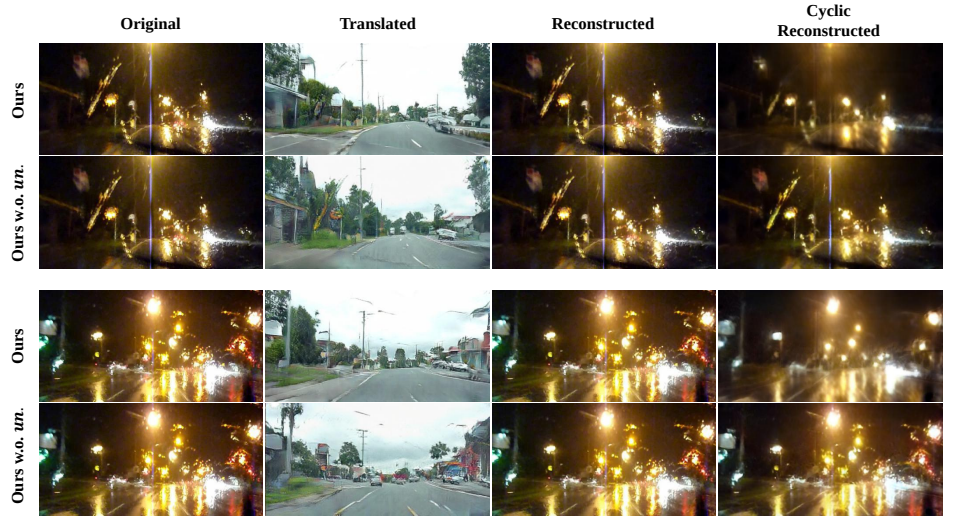

Figure 2: Experiments results of the variants of our model, i.e., the uncertainty-aware loss is used or not.

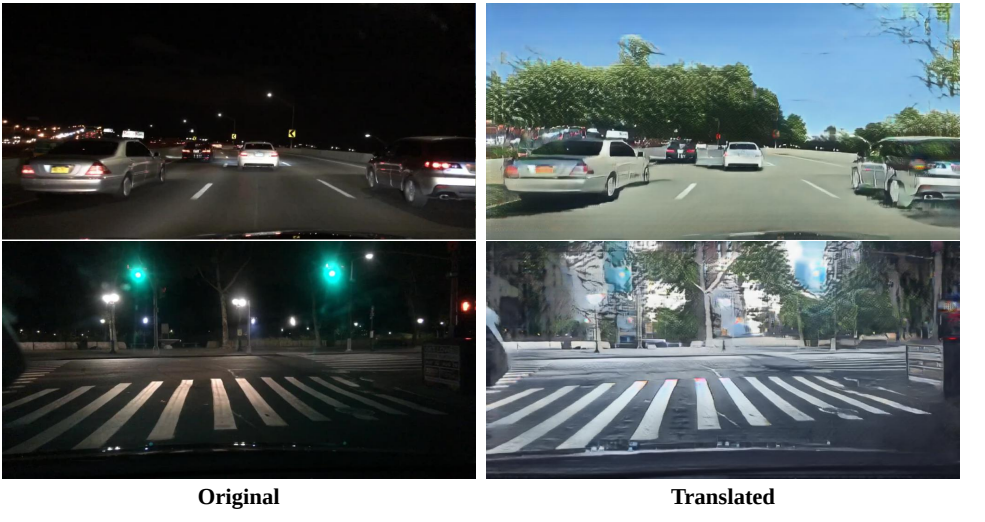

Figure 3: Experiments with higher resolution (512×1024) images of BDD100K

6.4 Additional comparison result and training details

In this section, we present additional comparison of qualitative results with same methods used in the main paper and we use official implementation and settings provided by the

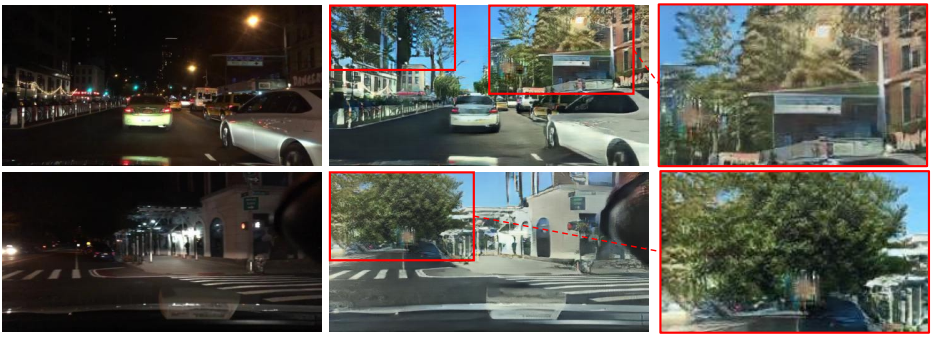

Figure 4: Failure case of our method

authors, i.e., CycleGAN [1] [10], UNIT [2] [4], ToDayGAN [3] [1] and ForkGAN [4] [10]. All methods are trained on NVIDIA RTX Titan GPU with same datasets, i.e., Alderley [5] and BDD100K [9] that are cropped and resized to 256×512. The number of iteration for training is about 100,000 with batch size 4 and if a model could not converge and fell into mode collapse, we picked earlier checkpoint which generates reasonable results. As shown in Fig. 5, our model performs domain translation with superior visual quality while preserving objects compared to other methods.

6.5 Extra qualitative results

Finally, we supplement the extra qualitative results (day $\leftrightarrow$ night) of our model on the datasets BDD100K [9] and Alderley [5]. Although the main purpose of our method is about adverse weather image translation, our model also can conduct the translation on opposite direction reasonably as shown in the right half of Fig. 6.

[1] https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix
[2] https://github.com/mingyuliutw/UNIT
[3] https://github.com/AAnoosheh/ToDayGAN
[4] https://github.com/zhengziqiang/ForkGAN

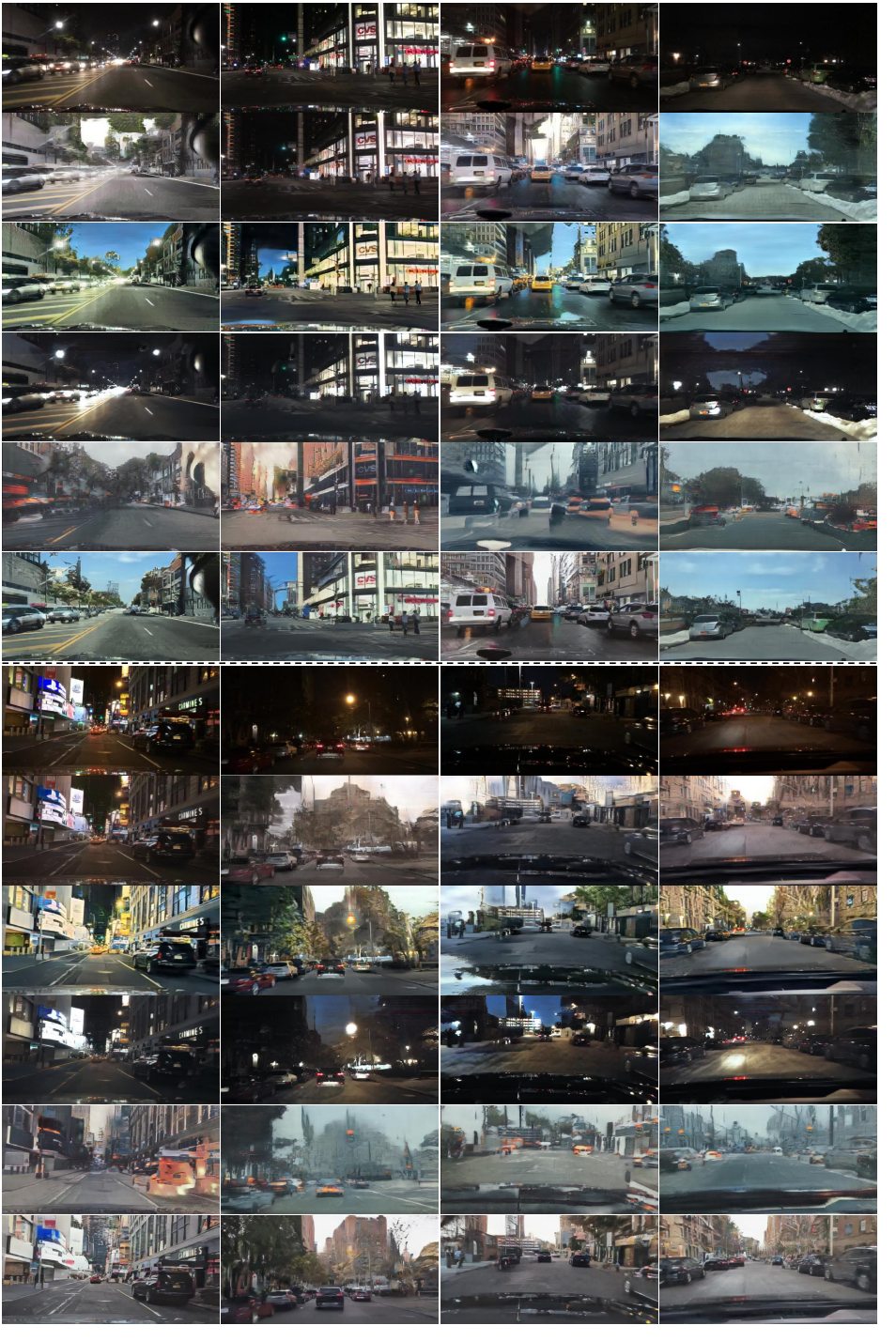

Figure 5: Additional results of qualitative comparison. **Please zoom in to see more details.**

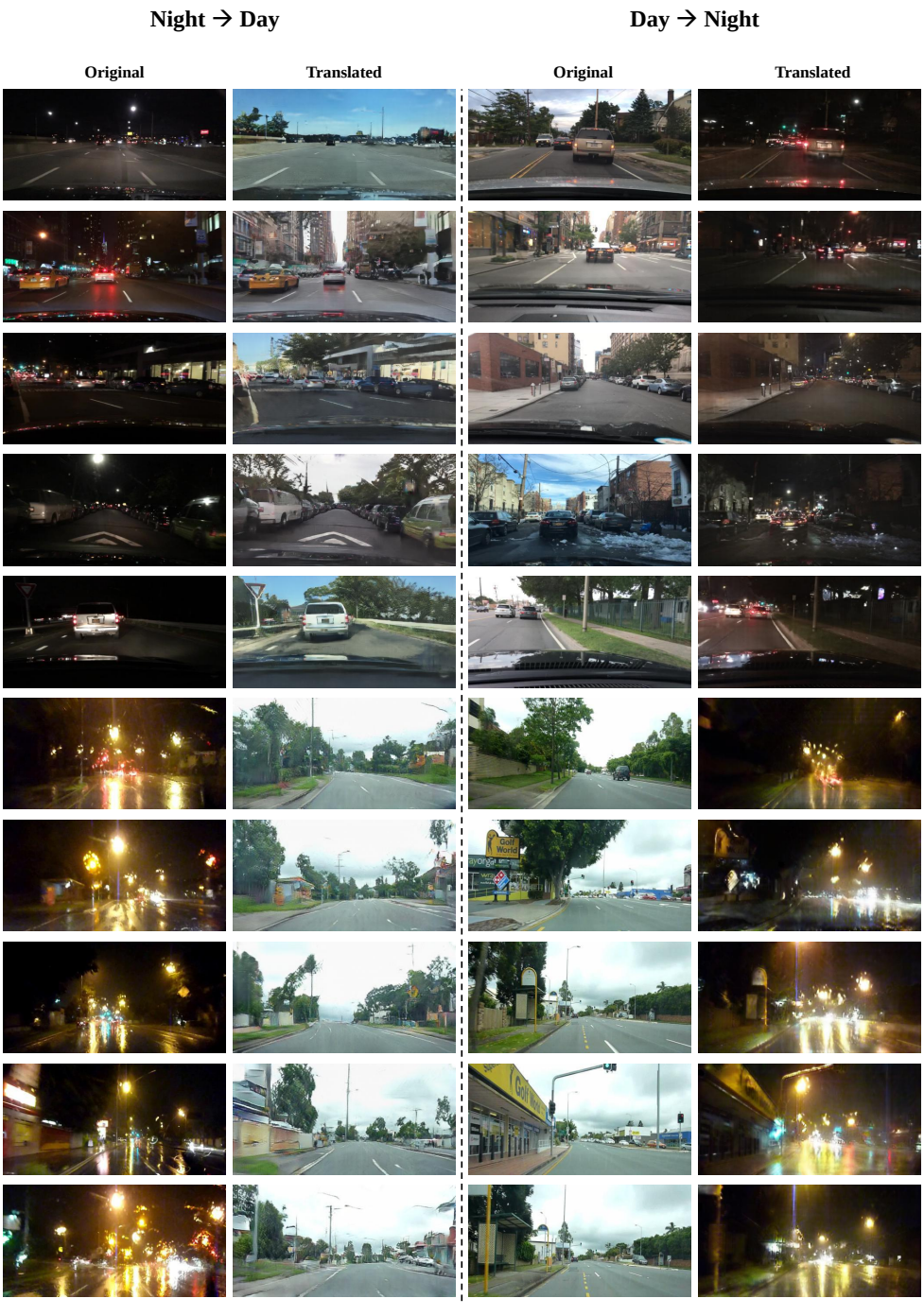

Figure 6: Extra qualitative results of our model. **Please zoom in to see more details.**

References

[1] Asha Anoosheh, Torsten Sattler, Radu Timofte, Marc Pollefeys, and Luc Van Gool. Night-to-day image translation for retrieval-based localization. In *International Conference on Robotics and Automation (ICRA)*, 2019.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-To-Image Translation With Conditional Adversarial Networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[5] Michael J. Milford and Gordon. F. Wyeth. Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In *International Conference on Robotics and Automation (ICRA)*, 2012.

[6] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[7] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[8] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations (ICLR)*, 2015.

[9] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018.

[10] Ziqiang Zheng, Yang Wu, Xinran Han, and Jianbo Shi. ForkGAN: Seeing into the rainy night. In *European Conference on Computer Vision (ECCV)*, 2020.

[11] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *International Conference on Computer Vision (ICCV)*, 2017.