CoMoGAN: continuous model-guided image-to-image translation
- supplementary file -

Fabio Pizzati
Inria, Vislab
fabio.pizzati@inria.fr

Pietro Cerri
Vislab
pcerri@ambarella.com

Raoul de Charette
Inria
raoul.de-charette@inria.fr

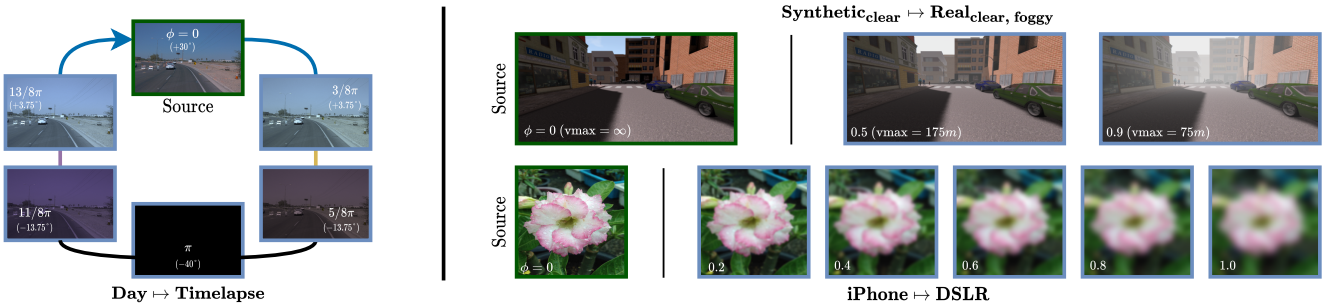

Figure 1: Model guidance for training for sample ϕ values (white text inset).

We provide the reader with additional insights about the importance of model guidance with details and ablations (Sec. 1), further training implementation details (Sec. 2). Additional qualitative results are reported in this document (Sec. 3). Refer to the video for additional visualizations.

1. Model guidance

Models – shown in Fig. 1 – are intentionally providing only a coarse training guidance and not intended for realistic translations. This is a fundamental difference with prior works [7, 10] as it allows learning complex visual effects *non-modeled* in the guidance. In particular from above figure, Day↦Timelapse model provides a tone mapping guidance that *intentionally* does not encompass realistic dawn/dusk/night visual appearance. Similarly, iPhone ↦ DSLR is a naive blurring guidance, and Synthetic$_{clear}$ ↦ Real$_{clear, foggy}$ provides guidance only on the foggy appearance while ignoring the synthetic-to-real changes. In CoMoGAN, the learning relies on our DRB block (main paper Sec. 3.2) to disentangle features so as to relax the model and learn the complex *non-modeled* features from unsupervised target data.

1.1. Details

Day↦Timelapse. We render intermediate conditions by interpolating the tone-mapped model from [9], written $\Omega(.)$. Since the latter was originally designed only for night time rendering, we replace the target color in $\Omega(.)$ by the average of the Hosek sky radiance model [5], denoted HSK(ϕ). For implementation reason, we accordingly map ϕ to $[0, 2\pi]$ so that max and min sun elevation angles corresponding to $30°$ and $-40°$, respectively. The complete model writes

$$M(x, \phi) = (1-\alpha)x + \alpha\Omega(x, \text{HSK}(\phi) + \text{corr}(\phi)) + \text{corr}(\phi)\,, \tag{1}$$

with α the interpolation coefficient defined as,

$$\alpha = \frac{1 - \cos(\phi)}{2}$$

and corr(ϕ) an asymmetrical hue correction to arbitrarily account for temperature difference at dusk and dawn. It writes

$$\text{corr}(\phi) = \begin{cases} [0.1, 0.0, 0.0]\sin(\phi) & \text{if } \sin(\phi) > 0, \\ [0.1, 0.0, 0.1](-\sin(\phi)) & \text{Otherwise.} \end{cases} \tag{2}$$

The effect of corr(.) is visible in Fig. 1 at $\phi = 5/8\pi$ and $\phi = 11/8\pi$, which both maps to elevation of $-13.75°$ for dusk (right image) and dawn (left image). We found that it slightly pushes the network towards better discovery of the red-ish and purple-ish appearance of dusk and dawn, respectively. Its importance is evaluated in Sec. 1.2.

Synthetic$_{clear}$ ↦ Real$_{clear, foggy}$. As mentioned, the guidance only account for fog without modeling *real* traits. We

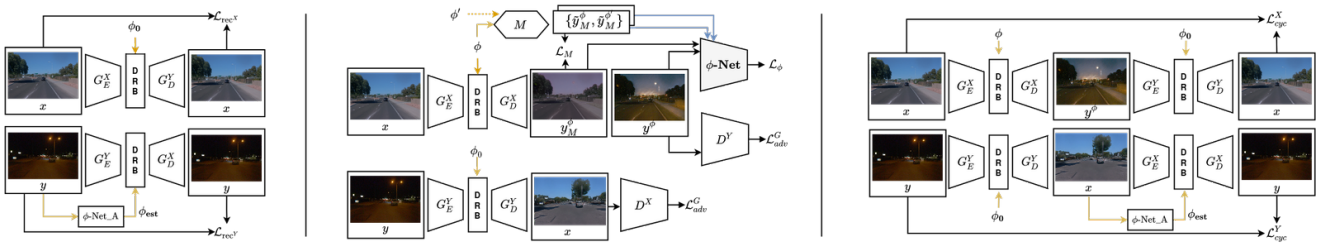

Figure 2: The complete training strategy for CoMo-MUNIT and CoMo-CycleGAN is composed by reconstruction, adversarial and cycle consistency constraints. The adversarial pipeline is adaptable to other GAN architectures seamlessly.

Model	IS↑	CIS↑	LPIPS↑
MUNIT	1.43	1.41	**0.583**
w/o color	1.42	1.35	0.577
w/o corr	1.56	1.45	0.579
CoMo-MUNIT	**1.59**	**1.51**	0.580

Table 1: We ablate the importance of a correct model for the cyclic scenario in Day $\mapsto$ Timelapse. Not distinguishing between dusk and dawn (*w/o corr*) brings the optimization to a simpler minimum, resulting in lower variability but still outperforming baseline MUNIT on IS/CIS. In the much harder guidance with only grayscale images (*w/o color*), the network gets slightly outperformed in image quality and diversity by baseline, still we are able to learn a reasonable data organization. CoMo-MUNIT performs best, using the complete model in Eq. 1.

use the model $f(x, d)$ from [4] to augment clear image with fog, assuming a depth map d. We use depth maps from either Cityscapes [2] or Synthia [8] projects pages. More in depth, [4] renders fog by applying a standard optical extinction model. The model writes

$$f(x, d) = xe^{-\beta(\phi)d} + L_\infty(1 - e^{-\beta(\phi)d}), \quad (3)$$

with L_∞ arbitrarily set to 0.859. We obtain the so-called extinction coefficient $\beta(\phi)$, by applying a step linear function following standard fog literature to map the maximum visibility from ∞ (*clear weather*) to 75m (thick fog). In formulas,

$$\beta(\phi) = \begin{cases} 0 & \text{if } \phi <= 0.2, \\ (\phi - 0.2) \cdot (\frac{0.045}{1-0.2}) & \text{Otherwise.} \end{cases} \quad (4)$$

iPhone $\mapsto$ DSLR. As model for guidance, we simply use gaussian blurring, with kernel radius in pixels accordingly mapped to ϕ values, as

$$M(x, k) = G(k) * x, \quad (5)$$

being G the Gaussian kernel, x input and k kernel size, which is directly mapped from $\phi \in [0, 1] \mapsto k \in [0, 8]$.

1.2. Model ablations

To evaluate the importance of model guidance, we ablate the model for Timelapse translation as it is the most complex translation task addressed. Performance is reported in Tab. 1.

Departing from the complete model in Eq. 1, we removed 1) the *corr* term (*w/o corr*), hence not distinguishing between dawn and dusk, and 2) color from the model (*w/o color*), hence providing only brightness information as guidance. From results in Tab. 1, while the complete model (*CoMo-MUNIT*) performs best, we still perform similar or better than the backbone by achieving controllable output even with symmetrical guidance (i.e. *w/o corr*) or naive brightness guidance (*w/o color*). This demonstrates that simple guidance is sufficient to reorganize the unsupervised target manifold.

2. Training details

Exploiting pairwise data. While losses presented in the paper are often sufficient to achieve convergence, we experienced that adding additional constraints with the available pairwise data further regularizes training to $\mathcal{L}_\phi$, such as

$$\begin{aligned} \mathcal{L}^G_{\phi M} &= ||\phi\text{-Net}(y^\phi_M, \tilde{y}^\phi_M)||_2 \\ &\quad + ||\phi\text{-Net}(y^\phi_M, \tilde{y}^{\phi'}_M) - \Delta\phi||_2, \\ \mathcal{L}_0 &= ||\phi\text{-Net}(y^\phi, y^\phi_M)||_2, \end{aligned} \quad (6)$$

We use those in all our trainings, adding them to $\mathcal{L}_\phi$.

Detailed training representation. In Fig. 2, we represent in details all constraints needed for CoMo-MUNIT/CoMo-CycleGAN training, which is composed by (1) reconstruction, (2) adversarial training and (3) cycle consistency. Additional regularization losses described above are omitted

for clarity. Again, steps (1) and (3) are necessary for cycle-consistency based network, still we assume the adversarial training (2) will be adaptable to any baseline.

Hyperparameters. We balance the contributions of the losses by weighting them when training CoMo-MUNIT in CoMo-CycleGAN. Specifically, for $\mathcal{L}_M$ and $\mathcal{L}_\phi$ we use a weight of 10 and for $\mathcal{L}_{reg}$ a weight of 1. The learning rate is set to $lr = 1e-4$ for CoMo-MUNIT and $lr = 2e-4$ for CoMo-CycleGAN as in [6] and [12], respectively.

Image size. We train Day $\mapsto$ Timelapse and $\text{Synthetic}_{\text{clear}} \mapsto \text{Real}_{\text{clear, foggy}}$ on x4 downsampled images, and train iPhone $\mapsto$ DSLR on 256x256 resized images. All training use horizontal flip data augmentation.

3. Additional qualitative results

We show additional qualitative results for Day $\mapsto$ Timelapse (Figs. 3,4,5,6 and video), $\text{Synthetic}_{\text{clear}} \mapsto \text{Real}_{\text{clear, foggy}}$ (Fig. 7) and iPhone $\mapsto$ DSLR (Fig. 8). *Note again that all Day $\mapsto$ Timelapse baselines use an additional supervision at Dusk/Dawn, which we do not require.*

Additional results are aligned with the main ones, with noticeable benefit over baselines such as: accurate manifold discovery (note the stable appearance of night in Figs. 3,4,5,6), the discovery of non-modeled features (note lights at night in Figs. 3,4,5,6, real traits in Fig. 7 and the depth-of-field like effect in Fig. 8).

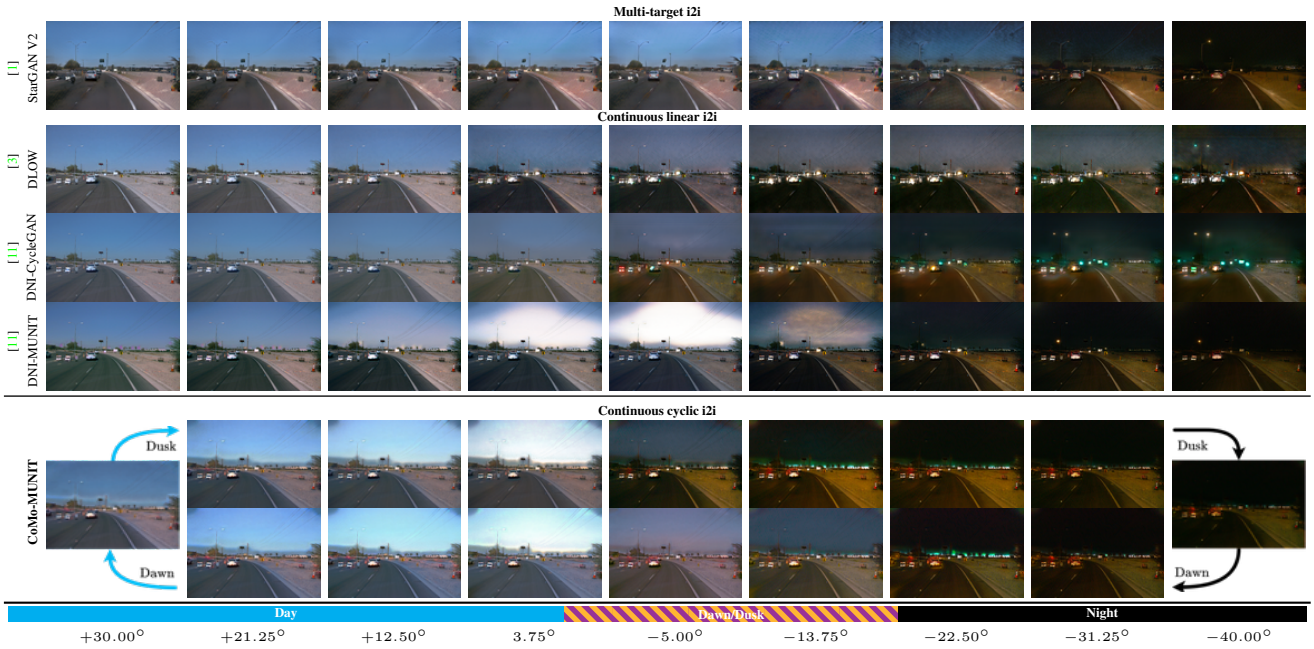

Figure 3: Additional qualitative results for Day ↦ Timelapse translations and baselines. Note all baselines (StarGAN v2, DLOW, DNI-CycleGAN, DNI-MUNIT) use additional Dusk/Dawn supervision.

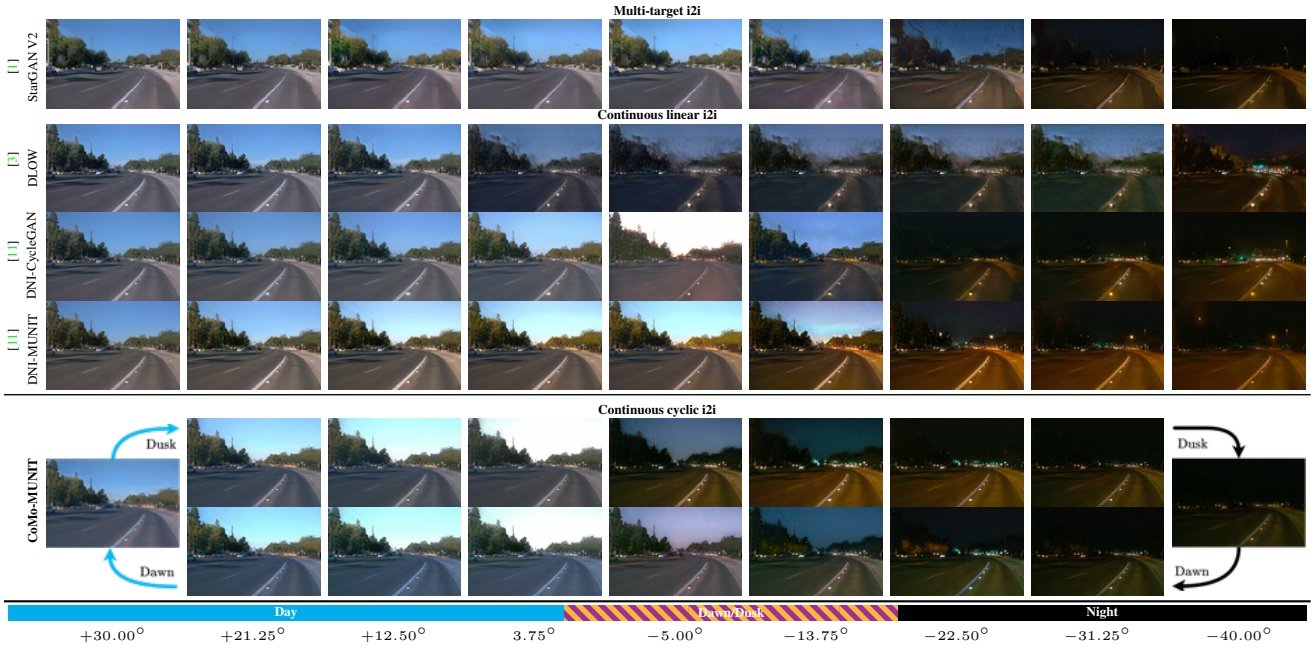

Figure 4: Additional qualitative results for Day ↦ Timelapse translations and baselines. Note all baselines (StarGAN v2, DLOW, DNI-CycleGAN, DNI-MUNIT) use additional Dusk/Dawn supervision.

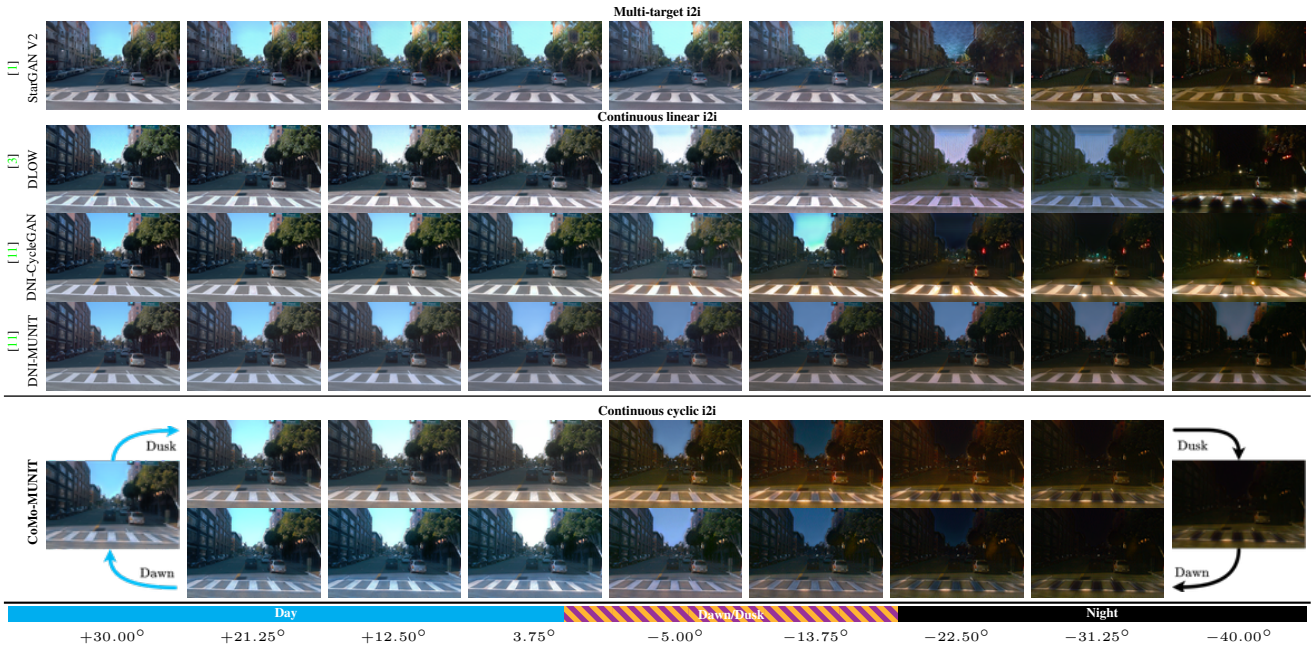

Figure 5: Additional qualitative results for Day ↦ Timelapse translations and baselines. Note all baselines (StarGAN v2, DLOW, DNI-CycleGAN, DNI-MUNIT) use additional Dusk/Dawn supervision.

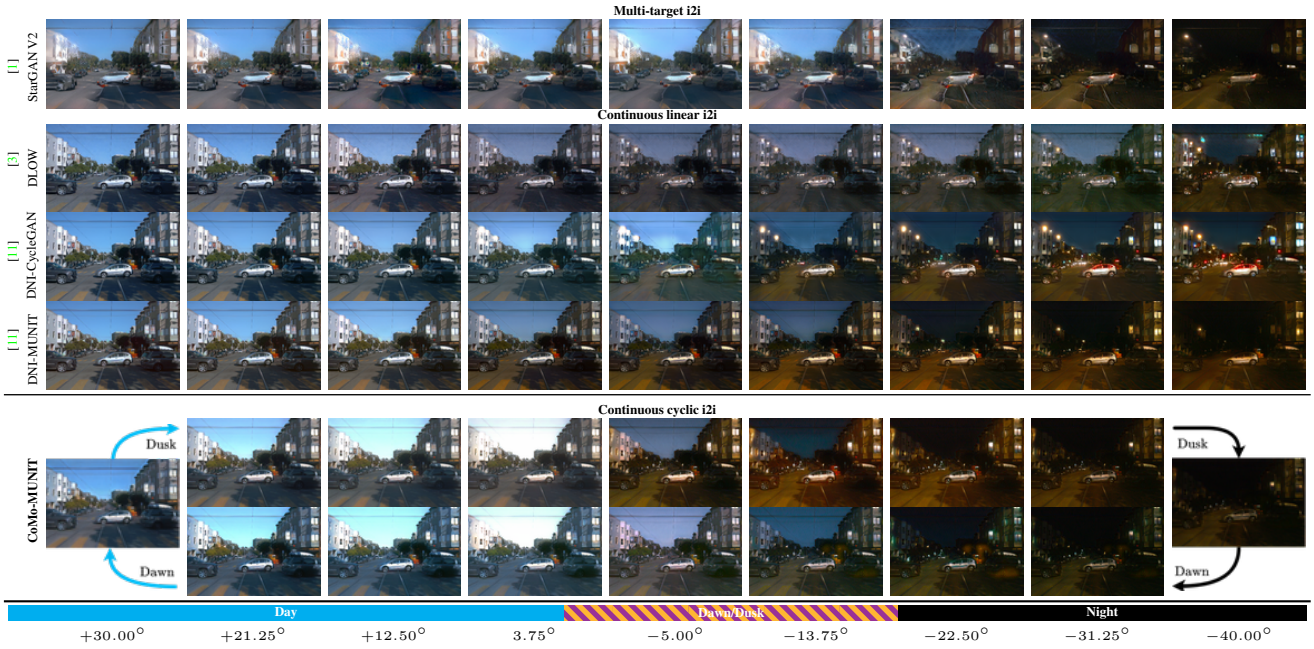

Figure 6: Additional qualitative results for Day ↦ Timelapse translations and baselines. Note all baselines (StarGAN v2, DLOW, DNI-CycleGAN, DNI-MUNIT) use additional Dusk/Dawn supervision.

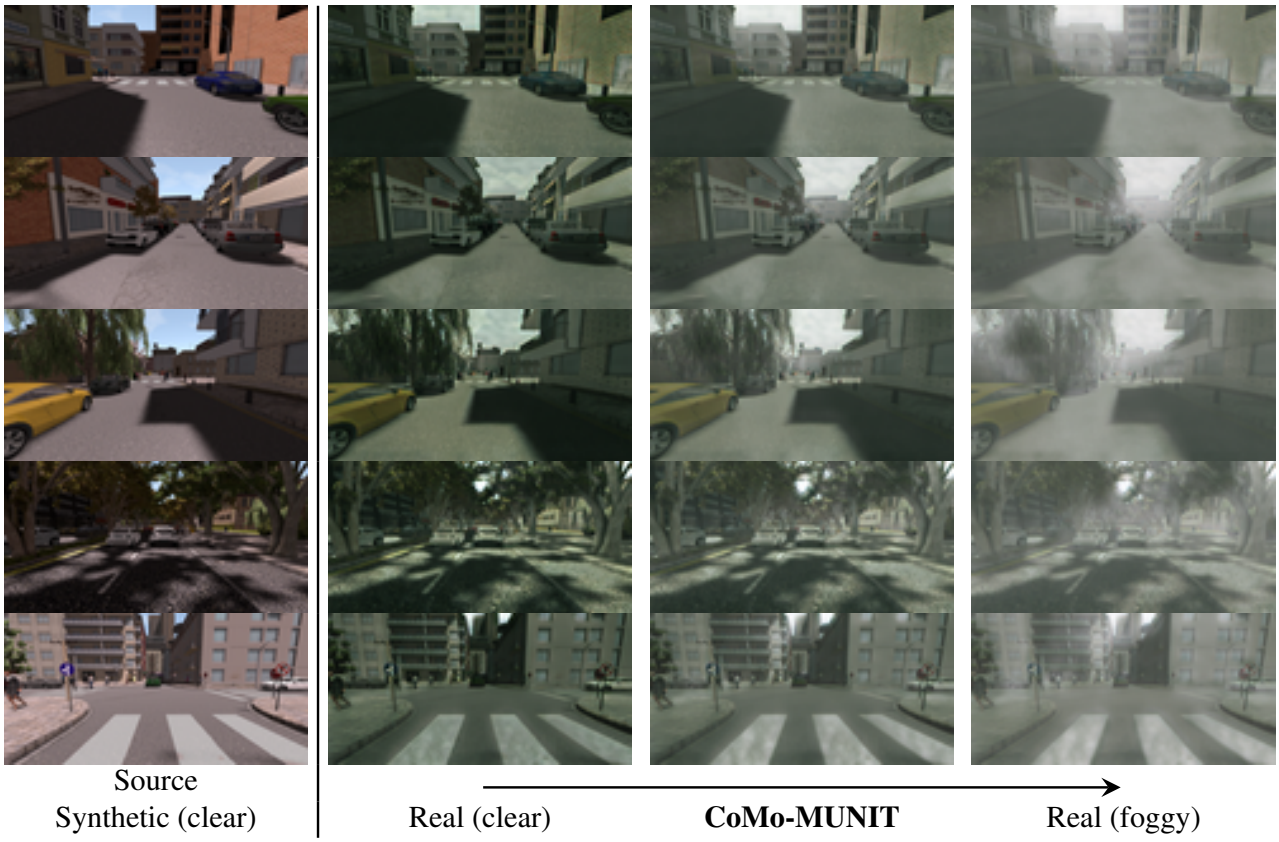

Figure 7: Additional qualitative results for $\text{Synthetic}_{\text{clear}} \mapsto \text{Real}_{\text{clear, foggy}}$.

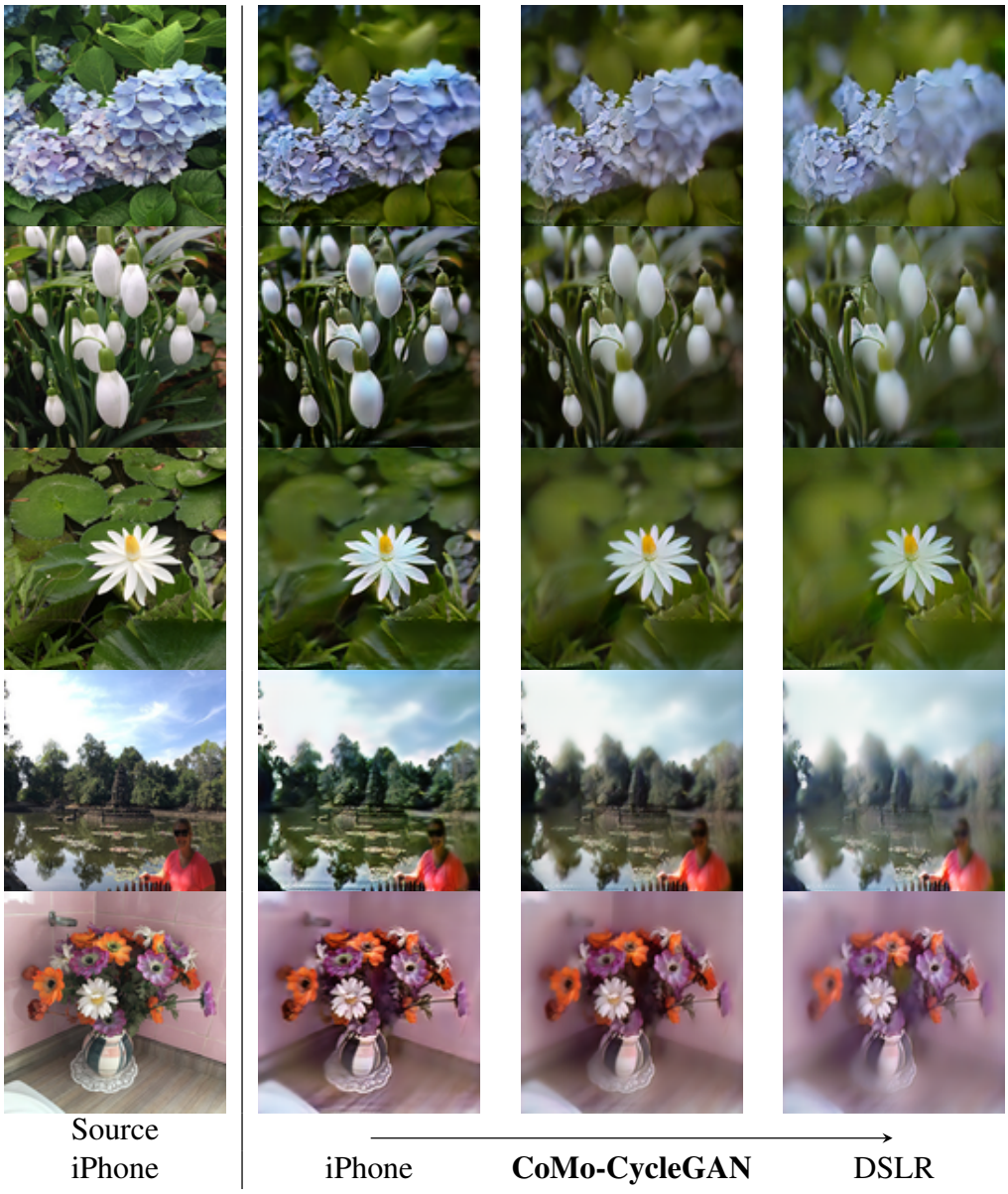

Figure 8: Additional qualitative results for iPhone $\mapsto$ DSLR.

References

[1] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 4, 5

[2] 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*, 2016. 2

[3] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *CVPR*, 2019. 4, 5

[4] Shirsendu Sukanta Halder, Jean-François Lalonde, and Raoul de Charette. Physics-based rendering for improving robustness to rain. In *ICCV*, 2019. 2

[5] Lukas Hosek and Alexander Wilkie. An analytic model for full spectral sky-dome radiance. *TOG*, 2012. 1

[6] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 3

[7] Ali Jahanian, Lucy Chai, and Phillip Isola. On the "steerability" of generative adversarial networks. In *ICLR*, 2020. 1

[8] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 2

[9] William B Thompson, Peter Shirley, and James A Ferwerda. A spatial post-processing algorithm for images of night scenes. *Journal of Graphics Tools*, 2002. 1

[10] Maxime Tremblay, Shirsendu Sukanta Halder, Raoul de Charette, and Jean-François Lalonde. Rain rendering for evaluating and improving robustness to bad weather. *IJCV*, 2020. 1

[11] Xintao Wang, Ke Yu, Chao Dong, Xiaoou Tang, and Chen Change Loy. Deep network interpolation for continuous imagery effect transition. In *CVPR*, 2019. 4, 5

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