SUPPLEMENTARY MATERIAL
NIR-assisted Video Enhancement via Unpaired 24-hour Data

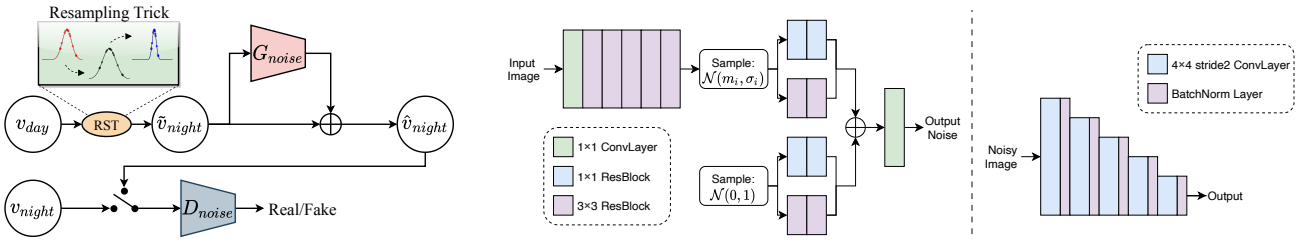
Muyao Niu, Zhihang Zhong, Yinqiang Zheng*
The University of Tokyo
muyao.niu@gmail.com, {zhong@is.s, yqzheng@ai}.u-tokyo.ac.jp

1. Details of NoiseGAN.

During the training of noise GAN G_{noise}, we simultaneously train a discriminator network D_{noise} to distinguish whether a given noisy image is synthesized from G_{noise} or sampled from the real nighttime VIS dataset. The two networks G_{noise} and D_{noise} can be optimized in an adversarial way:

$$\min_{G_{noise}} \max_{D_{noise}} \mathcal{L}_{adv} = \log(D_{noise}(v_{day})) + \log(1 - D_{noise}(\tilde{v}_{night} + G_{noise}(\tilde{v}_{night}))). \tag{1}$$

The training scheme of our noise GAN is shown in Figure 1(a). Once finishing training, G_{noise} can be used in the stage one pipeline to generate realistic pseudo nighttime VIS frames. Detailed architecture and training procedures of G_{noise} and D_{noise} are shown in Fig.1(b). The source codes for G_{noise} and D_{noise} are also included.



(a) (b)

Figure 1: (a) The training scheme of our noise GAN. (b) left: architecture of G_{noise}. Right: architecture of D_{noise}..

2. Details of Enhancement Network.

We present detailed components of our enhancement network G. "3x3 Conv" stands for a convolutional block in which the kernel size of convolution layers is 3×3, and the padding and the stride are all set to 1. The structure of Deformable Convolution completely follows [7]. The source codes of the enhancement network G are also included.

3. Results of unsupervised methods retrained on FMSVD.

We present the results of unsupervised methods after retrained on FMSVD, including Z-DCE [2], EnGAN [3], and SCI [5]. Results are reported in Tab.1 and Figure 3. We can see that EnGAN still fails to suppress the noise effect after training on our dataset. Z-DCE struggles to take balance between suppressing visible noise and enhancing light intensity. SCI produces unsatisfying results with obvious color bias and noise and fails to enhance the low-light VIS images.

*Corresponding author

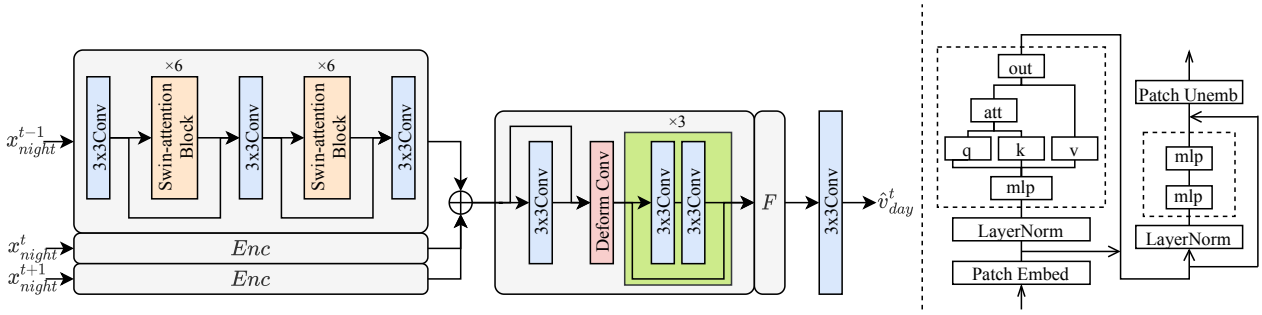


Figure 2: Left: Detailed architecture of our enhancement network. Right: Detailed component of swin-attention block.

Methods	FMSVD			Third Party		
	PI↓	NIQE↓	HSE↑	PI↓	NIQE↓	HSE↑
EnGAN	5.424	8.324	2.83	3.460	4.838	3.16
Z-DCE	5.373	8.785	2.16	4.741	7.509	2.33
SCI	9.132	16.071	1.83	8.035	13.951	1.66
Ours	**4.786**	**5.469**	**4.50**	**2.819**	**3.666**	**4.33**

Table 1: Numerical results of unsupervised methods after retrained on FMSVD. The best results are shown in **bold**.

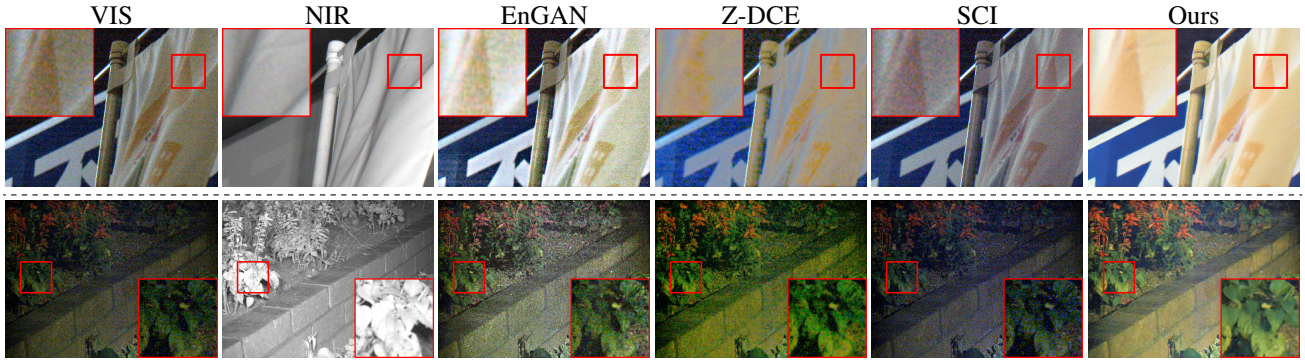


Figure 3: Visual comparisons with unsupervised methods after retraining on FMSVD (above dash) and the third-party dataset (below dash).

4. Comparison between heuristic and physic-based relighting algorithm.

As has been mentioned in the main paper, we use a heuristic relighting algorithm instead of a strictly physic-based relighting algorithm. The core obstacle lies in that existing intrinsic image decomposition algorithms can not deal with outdoor scenes under complex natural illumination, so it is impossible to extract the reflectance component for relighting. However, considering the co-located setting of camera and auxiliary illuminant at night, the shading component can be synthesized in a strictly physics-based way. In this section, we first describe this physic-based shading simulation for relighting, then compare its pseudo nighttime NIR results with our heuristic relighting algorithm. We also compare the effect of video enhancement models trained on data synthesized by these two algorithms respectively.

Physic-based shading simulation for relighting. To start with, we predict the depth map using MonoViT [6]. The network takes a daytime VIS frame v_{day} as input, and outputs a depth map $\mathcal{D} \in \mathbb{R}^{H \times W}$ for further usage:

$$\mathcal{D} = G_{depth}\left(v_{day}\right), \tag{2}$$

$$\hat{\mathcal{D}} = \frac{1}{\mathcal{D}}, \tag{3}$$

where smaller value in each position of $\hat{\mathcal{D}}$ represents closer distance to the camera.

Figure 4: Visual results for pseudo nighttime NIR from two types of relighting algorithms. (a) daytime NIR, (b) Depth Map, (c) pseudo nighttime NIR from our heuristic relighting algorithm, (d) pseudo nighttime NIR from our physic-based shading relighting algorithm.

According to the inverse-square law, the light field intensity I_{dis} is inversely proportional to the square of the distance d from the light source:

$$I_{dis} \propto \frac{1}{d^2}, \tag{4}$$

thus we calculate the intensity of each pixel in daytime NIR n_{day} according to the depth map $\hat{\mathcal{D}}$ as:

$$n_{day}^{dis}(i,j) = n_{day}(i,j) \cdot \frac{\delta^2}{\hat{\mathcal{D}}(i,j)^2}, \tag{5}$$

where $i \in \{1, 2, ..., H\}$, $j \in \{1, 2, ..., W\}$, and δ is the value that represents the closest distance from the camera.

Based on the original NIR frame n_{day}, the value of each pixel in n_{day}^{dis} gradually fades as the distance from the camera grows. This process basically reflects our consideration for physical distance, which is actually a quadratic decay model regarding the distance from the camera. Note that we have to approximate reflectance component by using NIR images since existing image intrinsic decomposition methods are inadequate for pre-processing, as has been discussed in the main paper.

To further consider the surface angle effect, we calculate the scale map $\mathcal{P}$ according to $\hat{\mathcal{D}}$, and use it to modulate n_{day}^{dis} as:

$$\hat{n}_{night} = n_{day}^{dis} \cdot \mathcal{P}, \tag{6}$$

where each pixel in $\mathcal{P}$ represents the reflected light intensity according to the Lambertian Reflectance Law [1, 4]:

$$I_{reflect} = \mathbf{L} \cdot \mathbf{N} = |\mathbf{L}||\mathbf{N}| \cos \alpha, \tag{7}$$

where $I_{reflect}$ is the light intensity after reflection, $\mathbf{L}$ is the light-direction vector, $\mathbf{N}$ is the surface's unit normal vector, and α is the angle between these two vectors. For each pixel in $\hat{\mathcal{D}}$, we assume $\mathbf{L}$ to be perpendicular to the image plane and calculate the unit normal vector of each pixel according to the $k \times k$-sized subarea centered on it.

Pseudo nighttime results of two types of relighting algorithms. Visual comparison results are shown in Figure 4. Results from the physic-based algorithm suffer from excessive content loss. Our heuristic algorithm successfully models the nighttime light distribution while preventing excessive content loss.

Results of enhancement models respectively trained on data synthesized by these two algorithms. The results are shown in Table 2 and Figure 5. We can see that our enhancement network obtains better numerical and visual results when trained on data synthesized via the heuristic relighting algorithm.

Therefore, we can conclude that, although the simulation of physics-based shading component is feasible, it will not perform better than our heuristic relighting algorithm. The reason is that, the daytime NIR image has to be used as reflectance, which is a rough approximation. So, without physics-based reflectance, to make the shading part only conform to physics will not help the enhancement algorithms.

5. Samples from FMSVD.

Figure 6 shows more visual samples from our dataset.

6. Pseudo Nighttime NIRs.

Figure 7 shows more visual results for pseudo nighttime NIR generated by our pseudo relighting algorithm.

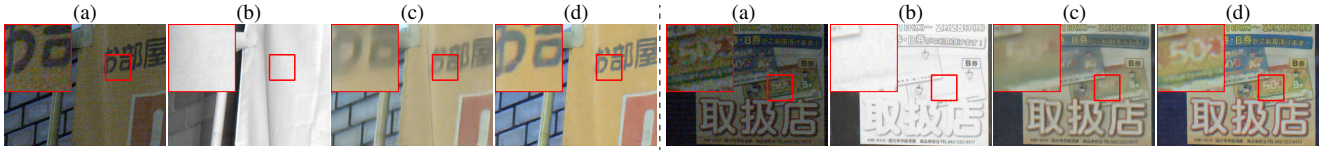


Figure 5: Quantitative results of enhancement model respectively trained on data synthesized by two types of relighting algorithms. (a) VIS, (b) NIR, (c) Results from model trained on data generated by our physic-based shading relighting algorithm, (d) Results from model trained on data generated by our heuristic relighting algorithm.

Table 2: Quantitative results of enhancement models respectively trained on data synthesized by two types of relighting algorithms.

Methods	FMSVD			Third-Party		
	PI↓	NIQE↓	HSE↑	PI↓	NIQE↓	HSE↑
Physics-based Shading	5.617	6.440	3.66	3.121	3.702	3.83
Heuristic	**4.786**	**5.469**	**4.50**	**2.819**	**3.666**	**4.33**

References

[1] Ronen Basri and David W Jacobs. Lambertian reflectance and linear subspaces. *IEEE transactions on pattern analysis and machine intelligence*, 25(2):218–233, 2003. 3

[2] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1780–1789, 2020. 1

[3] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30:2340–2349, 2021. 1

[4] Sanjeev J Koppal. Lambertian reflectance. *Computer vision: a reference guide*, pages 1–3, 2020. 3

[5] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5637–5646, 2022. 1

[6] Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *International Conference on 3D Vision*, 2022. 2

[7] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019. 1

Figure 6: Representative examples from our dataset. Above the dash: daytime; beneath the dash: nighttime. (a) NIR; (b) VIS.

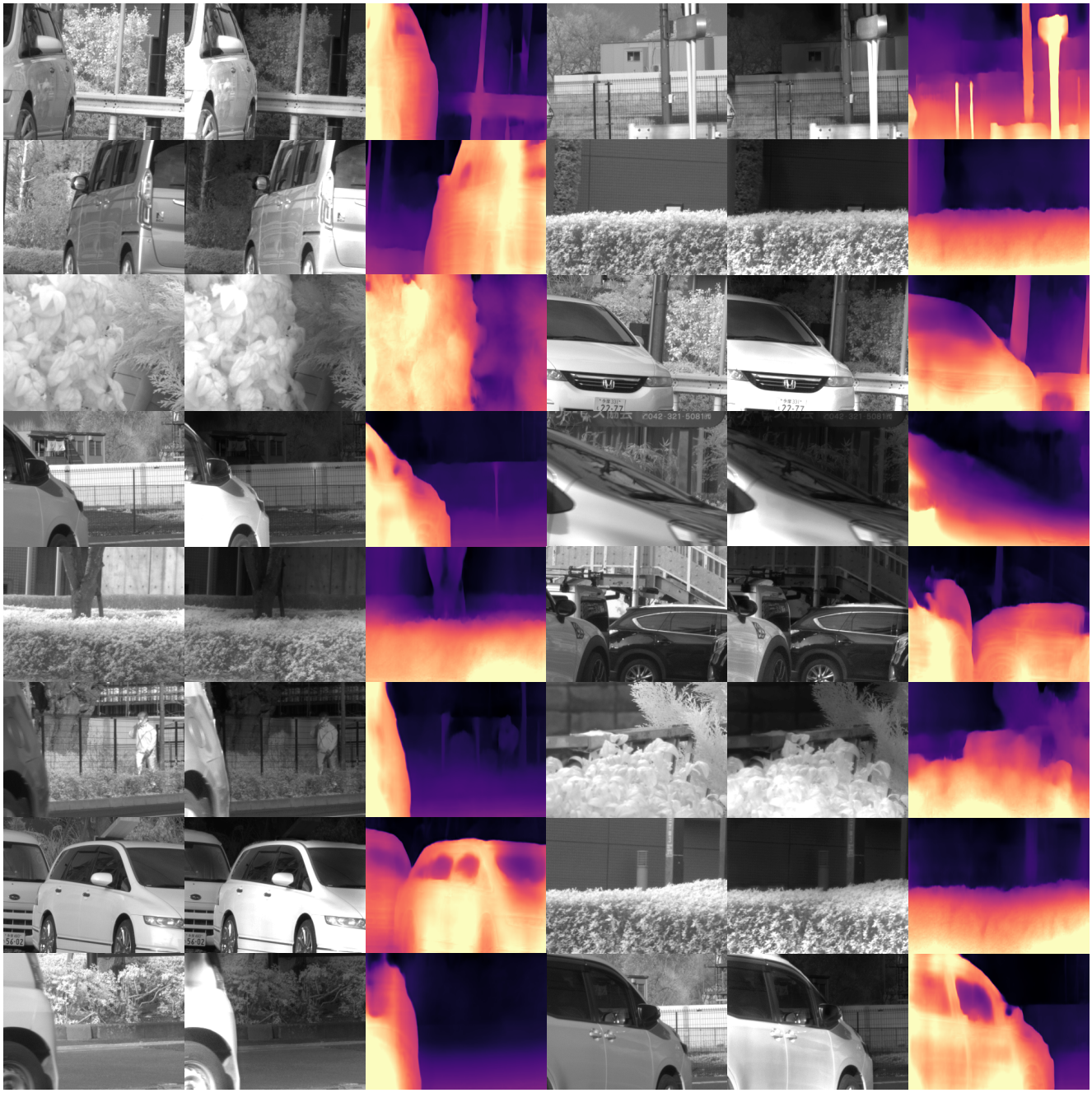


Figure 7: Visual results for pseudo NIR. (a) original daytime NIR; (b) pseudo nighttime NIR; (c) depth map estimated by G_{depth}.