Learning to See in Nighttime Driving Scenes with Inter-frequency Priors

Zhentao Fan[1] Xianhao Wu[1] Xiang Chen[2] Yufeng Li[1*]
[1]College of Electronic and Information Engineering, Shenyang Aerospace University
[2]School of Computer Science and Engineering, Nanjing University of Science and Technology

Abstract

Currently, image-to-image translation methods have achieved significant performance with the help of deep CNNs and GANs, but most existing models are not suitable for autonomous driving scenarios due to their interpretability. In this paper, we develop a high-quality image translator (i.e. night → day), N2D-LPNet, to facilitate nighttime driving scene dperception. Instead of pursuing a complex network structure, we first attempt to explore the inter-frequency relation knowledge to simplify image translation process. Specifically, the lightweight Laplace pyramid is introduced as the backbone architecture to decompose the feature maps of nighttime image into high- and low-frequency components. Considering the similar morphological properties of different frequency components, we design a flexible inter-frequency guiding strategy which utilizes each lower frequency information to refine the higher frequency feature in a progressive manner. Benefiting from the advantage of constraint from inter-frequency priors, our method can process the nighttime image better under the practical driving system while still persevere competitive results. We also discuss the potential value of N2D-LPNet for other high-level vision tasks.

1. Introduction

In nighttime scenarios, autonomous vehicles are often at risk because the cameras are not able to accurately identify more details in the dark. How to solve the "sluggishness" of self-driving at nighttime is the direction that many researchers have been working on [18]. Therefore, it is of great interest to develop a high-quality image translator (i.e. night → day, N2D) to work well under any light conditions.

Recently, numerous deep learning-based methods [1, 3, 11, 13, 14, 20, 23, 25, 28, 29] have been proposed to realize N2D task, which are mainly divided into two categories. The early way is to make nighttime images brighter with

*Corresponding author.

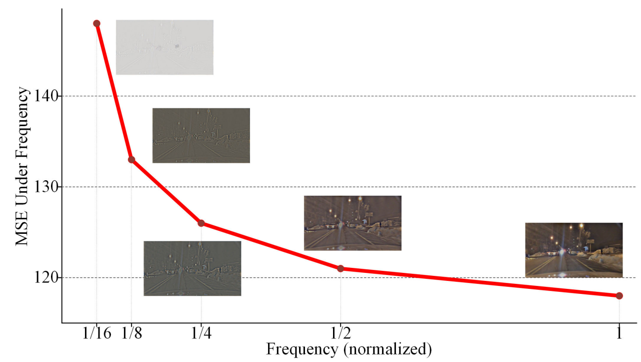

Figure 1. The MSE of image frequencies under different Laplace pyramid levels. We can observe that the morphological similarity between different frequency components of images, which can be explored to guide N2D task.

low-light image enhancement strategies [11, 14, 23], which propose to regularize the network training using the information extracted from the input data. However, these methods which baesd on enhancement cannot be well extended to nighttime driving scenes due to their performance bottlenecks.

The another popular way has been designed to simplify the N2D problem by directly learning day-to-night image translation via generation adversarial networks. CycleGAN [29] has been successfully applied to address the unpaired image-to-image translation, in order to achieve an unsupervised transfer process. To the best of our knowledge, there are only a few efforts for nighttime driving images translation. For exemple, TodayGAN [1] first utilizes image retrieval technology and migrates it to the visual localization system. Zheng et al. [28] present a ForkGAN to disentangle domain-invariant and domain-specific features between the nighttime domain and the daytime domain to realize N2D task. Lately, AU-GAN [13] has been developed based on an asymmetric CycleGAN architecture for adverse domain translation. Although these models show great performance, there are still the following inevitable problems in the actual nighttime driving scenes:

1) The existing approaches merely perform the GAN structure, while do not make full use of image priors which could be helpful to better boost the performance of deep models.

2) The learning-based models take the amount of computing resources and network parameters as the cost to achieve the satisfactory translation results. Due to the poor real-time performance, especially when processing high-resolution images, their application in automatic driving is limited.

3) These GAN-based translation algorithms have inadequate interpretability, and their results tend to leads to the destruction of the structure and semantic information of original images. Taking these images as input will bring trouble for subsequent high-level vision tasks such as vehicle detection [9] and road segmentation [22], which is contrary to our original intention.

To overcome the above problems, we introduce the inter-frequency priors with the lightweight Laplace pyramid pattern to design a high-quality image translator, named N2D-LPNet, which can contain fewer parameters than GANs. Taking into consideration the inherent properties of N2D task, we introduce Laplace pyramid framework as the backbone architecture to decompose the feature maps of nighttime image into high- and low-frequency components. Explicitly, the image illumination and color are mainly concentrated on the low-frequency components, while the image content details are mainly concentrated on the high-frequency component. In other words, the difference between daytime and nighttime images is mostly dominated in the low-frequency component. Motivated by the similar intuition in [10], Fig. 1 shows that the energy distribution of the nighttime image frequency obeys the reciprocal power law at different pyramid levels. In general, different frequency levels share the similar image morphological characteristics (*i.e.*, structure, texture, edge, etc.). Therefore, this motivates us to explore the inter-frequency relation knowledge in the network learning, so as to further simplify image translation process. To better process low-frequency components and reconstruct the images from Laplace pyramid, we build a flexible inter-frequency guiding strategy which utilizes each lower frequency information to refine the higher frequency feature in a progressive manner. The overall framework is trained end-to-end in an unsupervised fashion via unpaired adversarial learning strategy.

Several advantages can be gained from the proposed method. Firstly, our N2D-LPNet can process nighttime images and realize N2D, which successfully closes the gap in real-world automatic driving applications. Secondly, we make full use of low-frequency information to refine the high-frequency components, which is conducive to formulating a more compact network. Compared with other approaches, quantitative and qualitative experiments indicate the performance advantages of our developed method. We also show how N2D-LPNet can improve downstream tasks such as object detection.

2. Related Work

2.1. Image-to-Image Translation

The idea of image-to-image translation can be regarded as first proposed by Hertzmann et al.'s Image Analogies [7], who adopt a nonparametric texture model [4] on input- and output-image pair when training. Recently, for improving the time efficiency, many models [6, 15, 27] are adviced. [6] introduces two new layers, a multiplicative operation for affine transformation and a data-dependent lookup that enables slicing into the bilateral grid. [27] designs an algorithm for user-guided image colorization using deep learning. The system maps a grayscale image, along with sparse, local user "hints" to an output colorization with convolutional neural networks. Gharbi et al. [15] presents a framework for analyzing universal style transfer methods, which includes an effective and efficient algorithm that learns linear transformations. In addition, most researchers attempt to employ GANs to improve the quality of image translation. For example, Zhu et al. [29] applies a symmetric architecture based on bidirectional cycle consistency loss, which is considered a standard framework for unsupervised domain transfer methods. The work of [13] presents an asymmetric and uncertainty-aware GAN model to address image translation in adverse weather. The transcription and reconstruction of adverse domain images were separated by using a feature transfer network to enhance disentanglement of encoded features. Instead, we aim to explore the inter-frequency priors to simplify image translation process.

2.2. Laplacian Pyramid

In digital image processing, Laplacian pyramid [2] have long been used which main idea is that an image is linearly decomposed into a set of high- and low-frequency bands, then it is possible to reconstruct the original image exactly. The work of [5] has shown that the decomposition of Laplacian pyramid demonstrates to be capable of capturing a narrower distribution of describing both rain streaks and objects' details at different scales. The gradual masking strategy employed by [16] refines the high-frequency components effectively by using a Laplacian translation network. Inspired by these methods, we construct a lightweight Laplace pyramid network to take advantage of inter-frequency relation knowledge on N2D task.

3. Proposed Method

Fig. 2 illustrates the overall architecture of the proposed N2D-LPNet. In addition, the inter-frequency guiding strat-

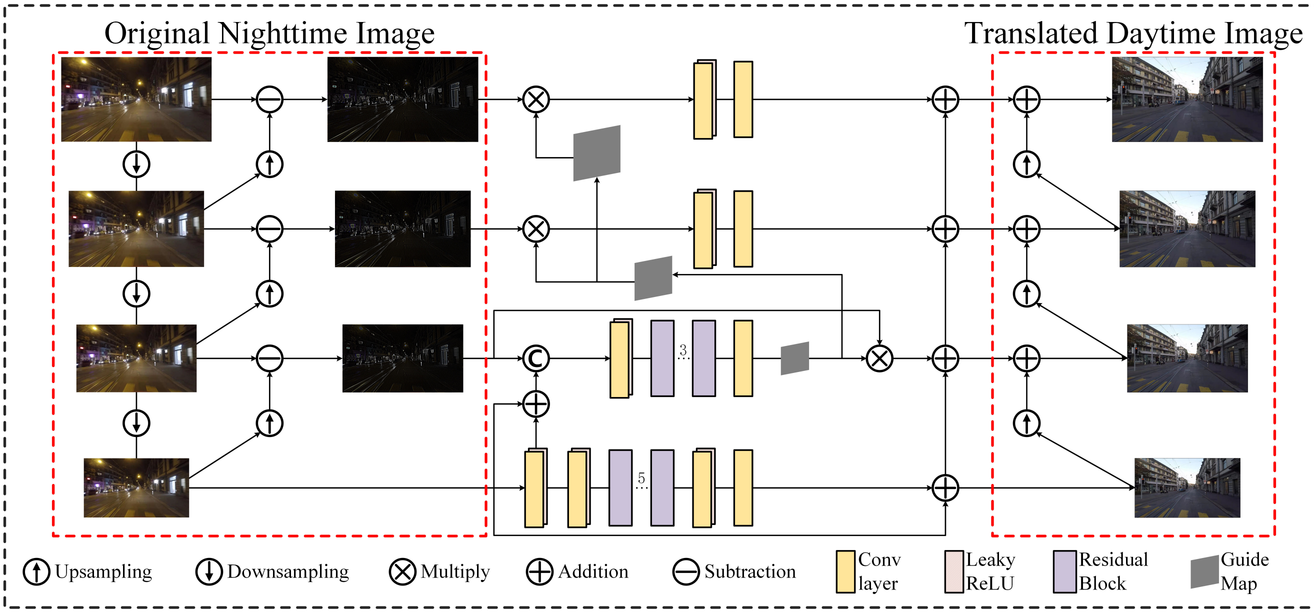

Figure 2. The framework of our developed N2D-LPNet for nighttime driving scenes. The red boxes are the input and output streams of the pyramid network. We make full use of inter-frequency priors which utilize each lower frequency information to refine the higher frequency feature in a progressive manner.

egy is designed to make each low-frequency component progressively refine the high-frequency components at the adjacent pyramid level.

3.1. Pyramid Network Architecture

We first decompose the input nighttime image $P_0 \in \mathbb{R}^{h \times w \times 3}$ into its Laplacian pyramid [2], obtaining a set of high-frequency parts $P_H = [F_0, F_1, \cdots, F_{L-1}] \in \mathbb{R}^{\frac{h}{2^{L-1}} \times \frac{w}{2^{L-1}} \times 3}$ and a low-frequency component $P_L \in \mathbb{R}^{\frac{h}{2^L} \times \frac{w}{2^L} \times 3}$, where L is the number of decomposed levels of the network. Here, we use the Laplacian pyramid as the backbone of the network, and the objective is twofold. First, N2D-LPNet performs multi-scale decomposition, allowing the model to take advantage of each level of sparsity. On the other hand, Laplacian Pyramid is a mature operation with small computation, which can be easily implanted into the driving system with GPU acceleration. For the low-frequency sub-band, we translate the initial features of pyramid decomposition using a lightweight network. For other sub-bands with higher resolution, we progressively interact with the learned inter-frequency guide map to adaptively refine the high-frequency information. The guidance operation is performed multiple times on the different frequency sub-bands. Finally, after all refinement and combination, N2D-LPNet reconstructs high-quality output results from nighttime input images. The inter-frequency priors are described below.

3.2. Inter-frequency Priors Guidance

Different from previous methods [5] that directly convolute the large-scale high frequency features, we explore the inter-frequency priors by excavating similarity information from each lower frequency information to refine the higher frequency feature, enforcing the communication within different pyramid levels. Thus, we exploit enough prior information to translate domain-specific features (*i.e.*, colors or illuminations in the N2D task) on low-frequency components with reduced resolution, which can significantly reduce the number of learnable parameters while maintaining remarkable performance. Specifically, we first use 1×1 convolution layer to extend the feature map channel of P_L. To compensate for the loss of information caused by feature decomposition, we stack five residual blocks to further enhance low-frequency feature propagation. Each residual block consists of two 3×3 convolution layers with stride number of 1 and a Leaky ReLU activation function. After achieving the translation of low-frequency sub-bands, the high-frequency components should also be carefully refined from P_L to $\hat{P}_L$, so as to achieve high-quality reconstruction. The final output $\hat{P}_0$ is reconstructed through using $\hat{P}_L$ and the refined feature $\left[\hat{F}_0, \hat{F}_1, \cdots, \hat{F}_{L-1}\right]$. Thus, we propose a simple yet effective inter-frequency guiding strategy for P_{L-1} and progressively share the feature tensor for refining high-frequency sub-bands, thanks to the morphological similarity between different frequency components of images. As shown in Fig. 2, we progressively upsample the

Table 1. Quantitative comparison of running speed and model size of different algorithms on BDD100K and Dark Zurich datasets.

Methods	BDD100K (1K) / s	Dark Zurich (2K) / s	Average Speed / s	Size / M
UNIT [17]	5.740	7.730	6.375	/
CycleGAN [29]	6.806	15.541	11.174	11.378
TodayGAN [1]	3.584	6.856	5.220	6.262
CUT [20]	1.965	5.359	3.662	11.378
EnlightenGAN [11]	0.968	3.622	2.295	8.237
Ours	**0.280**	**0.715**	**0.498**	**0.589**

Figure 3. Visual comparison of translation results on the BDD100K dataset.

pixel attention-based guide map $\left[\hat{A}_{L-2}, \hat{A}_{L-1}, \cdots, \hat{A}_0\right]$ at the adjacent pyramid level. We can refine all high-frequency components of the pyramid network by the following operations:

$$\hat{P}_{L-1} = P_{L-1} \otimes A_{L-1}, \tag{1}$$

where $\otimes$ represents the pixel-wise multiplication. Note that bilinear interpolation is employed in the upsampling operation.

3.3. Loss Function

Since the proposed N2D-LPNet is performed in an unpaired training setting, we adopt jointly optimize the reconstructive loss and adversarial loss in the image space for high quality. **Reconstructive Loss**. We apply a common-used reconstructive loss between the input image P_0 and the recovered image $\hat{P}_0$ to facilitate faithful translation, which can be given as:

$$\mathcal{L}_{recon} = \mathbb{E}\left\|P_0 - \hat{P}_0\right\|_2^2. \tag{2}$$

Adversarial Loss. To make the generated images look more realistic, the adversarial loss is introduced to match the target distribution, which is formulated based on the Least-Square GAN (LSGAN) [19] and the multi-scale dis-

Figure 4. Comparison of object detection results on the BDD100K dataset.

criminator [24]:

$$L_{adv} = \mathbb{E}[\log D(P_0^*)] + \mathbb{E}\left[\log\left(1 - D\left(G(P_0)\right)\right]\right., \quad (3)$$

where P_0^* denotes the unpaired ground-truth images.

Total Loss. The overall objective function used for training our network is thus calculated as:

$$\mathcal{L} = \mathcal{L}_{recon} + \lambda \mathcal{L}_{adv}, \quad (4)$$

where λ is a weight parameter. Here, we empirically set $\lambda = 0.1$ to balance the two losses.

4. Experiments

In this section, the proposed N2D-LPNet is evaluated. We first describe the datasets and training details. Then we compare our method with other algorithms quantitatively and qualitatively. Finally, the ablation study and user study are provided to show the effectiveness of the N2D-LPNet.

4.1. Datasets Setup

To evaluate our network and other algorithms, we collect two experimental datasets from **BDD100K** [26] and **Dark Zurich** [21]. 1) BDD100K (1K image resolution) consists of 100,000 images of the urban roads for autonomous driving which is used 3,600 images to be training set and 400 images as a test data. 2) Dark Zurich (2K image resolution) includes 2,890 daytime images and 2,567 nighttime images which is used to evaluate the generalization ability of the model.

4.2. Training Details

Our entire network is implemented in PyTorch and we use Adam optimizer [12]. The pyramid level L is set to 3. All of our experiments use a batch size of 16 and they are performed on one Tesla V100 GPU. All images are loaded in original size then cropped to 256×256 patches. Random flipping or rotation is applied for data augmentation. The final model is trained for a total of 300,000 iterations with the learning rate of 0.0001. Note that we shuffle above-mentioned images randomly in order to achieve unpaired supervised learning.

4.3. Comparison Results

We first present the results of qualitative comparison with five N2D methods, *i.e.*, UNIT [17], CycelGAN [29], TodayGAN [1], CUT [20], and EnlightenGAN [11]. Fig. 2

Table 2. Ablation study on different pyramid levels.

Number of Level	L = 3 (default)	L = 4	L = 5
FID / Run Time	**69.149 / 0.280**	80.093 / 0.382	74.289 / 0.397

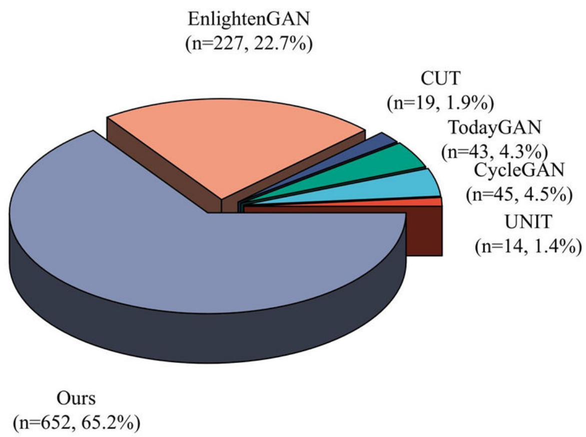

Figure 5. Comparison of user study on the BDD100K dataset.

shows two visual examples for comparison. Obviously, our developed method obtains the best image translation performance with less artifacts, blur, and reflections. In addition, we also perform quantitative analysis. As recorded in Tab. 1, other methods tend to take longer time to run which can be fatal for autonomous driving. It is heartening that our model uses fewer number of parameters and less running time. According the value of FID [8] in Fig. 7 (a), our method get the lower score which means the output has a higher image translation quality.

For further general verification in the high-level vision tasks, the translated images are further evaluated for the task of object detection. As can seen in Fig. 4, other approaches have poor capability for object detection, due to their translation results cause the destruction of the semantic information of the original image. However, our method can detect more objects and get higher scores. The above results can surely reflect the practicability and robustness of the designed N2D-LPNet.

4.4. Ablation Study

In order to verify the design rationality of model parameters, we perform ablation experiments to validate the influence of different number of pyramid levels L on the N2D task. For fair comparison, training settings of these variants are the same, and the values of FID and running time are listed in Tab. 2. We can easily find that N2D-LPNet receives the more superior consequences on all tested indicators with $L = 3$ than that with $L = 4$ or $L = 5$. We also provide visual observation in Fig. 6. By observing zoomed

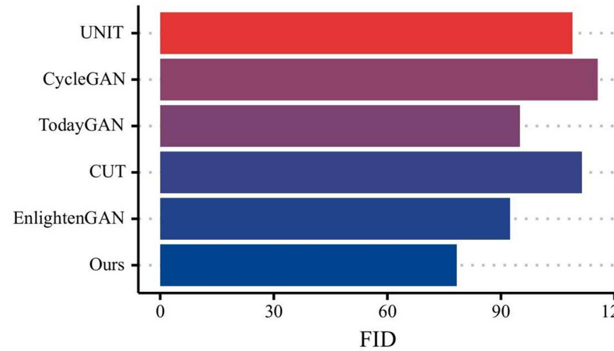

Figure 6. Comparison of FID quantitative results on the BDD100K dataset.

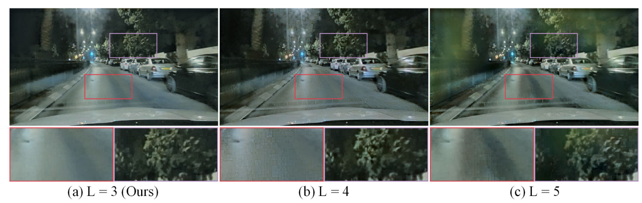

Figure 7. Visual comparison of different pyramid levels on BDD100K dataset.

color boxes, our results ($L = 3$) achieve better performance without distortion and artifacts in the translation results of other pyramid levels.

4.5. User Study

Our research goals are to assist autonomous driving, so that the related evaluation is the perceptual feeling with drivers. We designed a questionnaire online to make it objectively and easier for participants. The participants consisted of 50 professional drivers, who are willing to take part in the online survey. The survey included 20 questions about visual effect and image selection, such as "which picture do you prefer as a driving scene?", "which image is the most useful for driving at night?", "which image can guide drivers to accurately judge road conditions?" and so on. The results are shown in Fig. 7 (b). The proportions of the different parts of the pie chart represent the percentage of people who support that side. The survey helps us to define the degree of N2D translation and our approach has earned the largest number of votes in all above methods. The other methods do not perform well on this subjective task since there are visible structural distortions and artifacts of their results.

5. Conclusion

In this paper, we have introduced an effective image translator N2D-LPNet for exploring the benefits from taking advantage of inter-frequency prior as guidance in N2D task. Based on the lightweight Laplacian pyramid, the inter-frequency guiding strategy is adaptively introduced into the pyramid structure for progressive feature refinement. Attributed to the constraint from inter-frequency priors of images, our developed N2D-LPNet exhibits not only has fewer network parameters but also comparable translation performance. Furthermore, our method has potential values for high-level vision tasks in self-driving field.

Acknowledgements. This work was supported by Liaoning Provincial Applied Basic Research Project under Grant 2022JH2/101300247.

References

[1] Asha Anoosheh, Torsten Sattler, Radu Timofte, Marc Pollefeys, and Luc Van Gool. Night-to-day image translation for retrieval-based localization. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5958–5964. IEEE, 2019. 1, 4, 5

[2] Vinicius F Arruda, Thiago M Paixao, Rodrigo F Berriel, Alberto F De Souza, Claudine Badue, Nicu Sebe, and Thiago Oliveira-Santos. Cross-domain car detection using unsupervised image-to-image translation: From day to night. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 2, 3

[3] Xiang Chen, Yufeng Li, Caihua Kong, and Longgang Dai. Unpaired image dehazing with physical-guided restoration and depth-guided refinement. *IEEE Signal Processing Letters*, 29:587–591, 2022. 1

[4] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1033–1038. IEEE, 1999. 2

[5] Xueyang Fu, Borong Liang, Yue Huang, Xinghao Ding, and John Paisley. Lightweight pyramid networks for image deraining. *IEEE transactions on neural networks and learning systems*, 31(6):1794–1807, 2019. 2, 3

[6] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 2

[7] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340, 2001. 2

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[9] Xiaowei Hu, Xuemiao Xu, Yongjie Xiao, Hao Chen, Shengfeng He, Jing Qin, and Pheng-Ann Heng. Sinet: A scale-insensitive convolutional neural network for fast vehicle detection. *IEEE transactions on intelligent transportation systems*, 20(3):1010–1019, 2018. 2

[10] Zhuang Jia. Exploring inter-frequency guidance of image for lightweight gaussian denoising. *arXiv preprint arXiv:2112.11779*, 2021. 2

[11] 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, 5

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[13] Jeong-gi Kwak, Youngsaeng Jin, Yuanming Li, Dongsik Yoon, Donghyeon Kim, and Hanseok Ko. Adverse weather image translation with asymmetric and uncertainty-aware gan. *arXiv preprint arXiv:2112.04283*, 2021. 1, 2

[14] Mading Li, Jiaying Liu, Wenhan Yang, Xiaoyan Sun, and Zongming Guo. Structure-revealing low-light image enhancement via robust retinex model. *IEEE Transactions on Image Processing*, 27(6):2828–2841, 2018. 1

[15] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast image and video style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3809–3817, 2019. 2

[16] Jie Liang, Hui Zeng, and Lei Zhang. High-resolution photorealistic image translation in real-time: A laplacian pyramid translation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9392–9400, 2021. 2

[17] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017. 4, 5

[18] Fuya Luo, Yunhan Li, Guang Zeng, Peng Peng, Gang Wang, and Yongjie Li. Thermal infrared image colorization for nighttime driving scenes with top-down guided attention. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):15808–15823, 2022. 1

[19] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 4

[20] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020. 1, 4, 5

[21] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7374–7383, 2019. 5

[22] Pourya Shamsolmoali, Masoumeh Zareapoor, Huiyu Zhou, Ruili Wang, and Jie Yang. Road segmentation for remote sensing images using adversarial spatial pyramid networks. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6):4673–4688, 2020. 2

[23] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8174–8182, 2018. 1

[24] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 5

[25] Guanglei Yang, Hao Tang, Humphrey Shi, Mingli Ding, Nicu Sebe, Radu Timofte, Luc Van Gool, and Elisa Ricci. Global and local alignment networks for unpaired image-to-image translation. *arXiv preprint arXiv:2111.10346*, 2021. 1

[26] 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*, 2(5):6, 2018. 5

[27] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. *arXiv preprint arXiv:1705.02999*, 2017. 2

[28] Ziqiang Zheng, Yang Wu, Xinran Han, and Jianbo Shi. Forkgan: Seeing into the rainy night. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 155–170. Springer, 2020. 1

[29] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1, 2, 4, 5