Night-to-Day Translation via Illumination Degradation Disentanglement

Anonymous Author(s)
Affiliation
Address
email

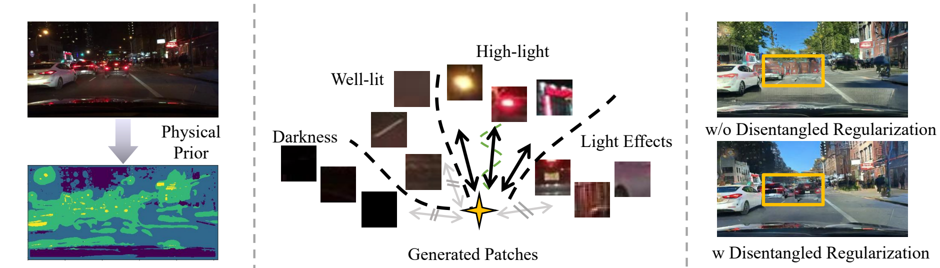
Abstract

Night-to-Day translation (Night2Day) aims to achieve day-like vision for nighttime scenes. However, processing night images with complex degradations remains a significant challenge under unpaired conditions. Previous methods that uniformly mitigate these degradations have proven inadequate in simultaneously restoring daytime domain information and preserving underlying semantics. In this paper, we propose **N2D3** (**N**ight-to-**D**ay via **D**egradation **D**isentanglement) to identify different degradation patterns in nighttime images. Specifically, our method comprises a degradation disentanglement module and a degradation-aware contrastive learning module. Firstly, we extract physical priors from a photometric model based on Kubelka-Munk theory. Then, guided by these physical priors, we design a disentanglement module to discriminate among different illumination degradation regions. Finally, we introduce the degradation-aware contrastive learning strategy to preserve semantic consistency across distinct degradation regions. Our method is evaluated on two public datasets, **demonstrating a significant improvement of 5.4 FID on BDD100K and 10.3 FID on Alderley**.

1 Introduction

Nighttime images often suffer from severe information loss, posing significant challenges to both human visual recognition and computer vision tasks including detection, segmentation, *etc.* [14]. In contrast, daylight images exhibit rich content and intricate details. Achieving day-like nighttime vision remains a primary objective in nighttime perception, sparking numerous pioneering works [30]. Night-to-Day image translation (Night2Day) offers a comprehensive solution to achieve day-like vision at night. The primary goal is to transform images from nighttime to daytime while maintaining their underlying semantic structure. However, achieving this goal is challenging. It requires to process complex degraded images using unpaired data, which raises additional difficulties compared to other image translation tasks.

Recently, explorations have been made in Night2Day. Early approaches, such as ToDayGAN, demonstrated the effectiveness of cycle-consistent learning in maintaining semantic structure [1]. Subsequent methods incorporated auxiliary structure regularization techniques, including perceptual loss and uncertainty regularization, to better preserve the original structure [33, 18]. Furthermore, some methods utilized daytime images with nearby GPS locations to aid in coarse structure regularization [26]. However, these methods often neglect the complex degradations at nighttime, applying structure regularization uniformly and resulting in severe artifacts. To address this issue, more recent approaches adopt auxiliary human annotations to maintain semantic consistency, such as segmentation maps and bounding boxes [16, 22]. Despite their potential, these methods are labor-intensive and challenging, especially since many nighttime scenes are beyond human cognition.



(a) Disentanglement Process (b) Disentangled Regularization (c) Visual Comparison

Figure 1: Illustration of our motivation. (a) The disentanglement process leverages physical priors. (b) The image patches are restored individually for each degradation type. (c) The proposed Disentangled Regularization improves the overall performance.

The critical limitation of the aforementioned methods is the disregard for complex degraded regions. Specifically, different regions in nighttime images possess varying characteristics, such as extreme darkness, well-lit regions, light effects, *etc*. Treating all these degraded regions equally could adversely impact the results. As illustrated in Figure 1, our key insight emphasizes that nighttime images suffer from various degradations, necessitating customizing restoration for different degradation types. Intuitively, we manage to disentangle nighttime images into patches according to the recognized degradation type and learn individual restoration patterns for them to enhance the overall performance.

Motivated by this point, we propose N2D3 (**N**ight to **D**ay via **D**egradation **D**isentanglement), which utilizes Generative Adversarial Networks (GANs) to bridge the domain gap between nighttime and daytime in a degradation-aware manner, as illustrated in Figure 2. There are two modules in N2D3, including physical-informed degradation disentanglement and degradation-aware contrastive learning, which are employed to preserve the semantic structure of nighttime images. In the disentanglement of nighttime degradation, a photometric model tailored to nighttime scenes is conducted to extract physical priors. Subsequently, the illuminance and physical priors are integrated to disentangle regions into darkness, well-lit, high-light, and light effects. Building on this, degradation-aware contrastive learning is designed to constrain the similarity of the source and generated images in different regions. It comprises disentanglement-guided sampling and reweighting strategies. The sampling strategy mines valuable anchors and hard negative examples, while the reweighting process assigns their weights. They enhance vanilla contrastive learning by prioritizing valuable patches with appropriate attention. Ultimately, our method yields highly faithful results that are visually pleasing and beneficial for downstream vision tasks including keypoint matching and semantic segmentation.

Our contributions are summarized as follows:

(1) We propose the N2D3 translation method based on the illumination degradation disentanglement module, which enables degradation-aware restoration of nighttime images.

(2) We present a novel degradation-aware contrastive learning module to preserve the semantic structure of generated results. The core design incorporates disentanglement-guided sampling and reweighting strategies, which greatly enhance the performance of vanilla contrastive learning.

(3) Experimental results on two public datasets underscore the significance of considering distinct degradation types in nighttime scenes. Our method achieves state-of-the-art performance in visual effects and downstream tasks.

2 Related Work

Unpaired Image-to-Image Translation. Unpaired image-to-image translation addresses the challenge of lacking paired data, providing an effective self-supervised learning strategy. To overcome the efficiency limitations of traditional cycle-consistency learning, Park *et al.*, first introduces contrastive learning to this domain, achieving efficient one-sided learning[20]. Following this work, several studies have improved the contrastive learning by generating hard negative examples [24], re-weighting positive-negative pairs [31], and selecting key samples [9]. Furthermore, other constraints, such as density [27] and path length [28], have been explored in unpaired image translation. However, all these works neglect physical priors in the nighttime, leading to suboptimal results in Night2Day.

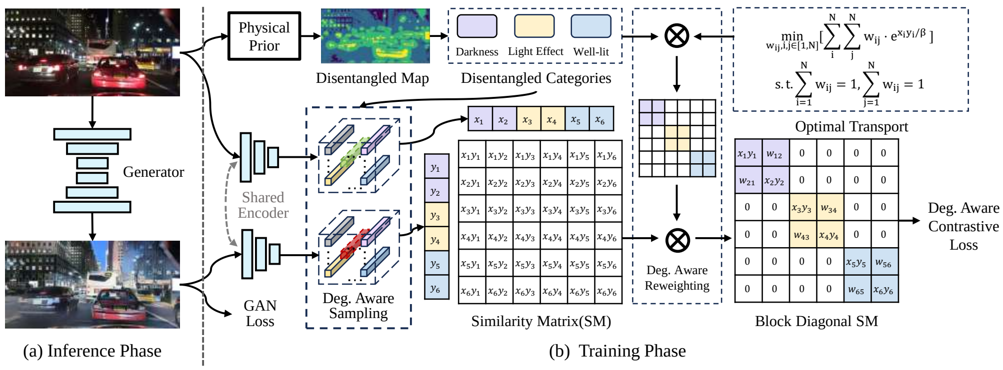


Figure 2: The overall architecture of the proposed N2D3 method. The training phase contains the physical prior informed degradation disentanglement module and degradation-aware contrastive learning module. They are utilized to optimize the ResNet-based generator which is the main part in the inference phase.

Nighttime Domain Translation. Domain translation techniques have been applied to address adverse nighttime conditions. An early contribution is made by Anoosheh *et al.*, which demonstrates the effectiveness of cycle-consistent learning in Night2Day[1]. Following this, many works incorporate different modules into cycle-consistent learning to enhance structural modeling capabilities. Zheng *et al.* incorporate a fork-shaped encoder to enhance visual perceptual quality[33]. AUGAN employs uncertainty estimation to mine useful features in nighttime images[18]. Fan *et al.* explore inter-frequency relation knowledge to streamline the Night2Day process[5]. Xia *et al.* utilize nearby GPS locations to form paired night and daytime images, providing weak supervision[26]. Some other studies incorporate human annotations to impose structural constraints, overlooking the practical difficulty of acquiring such annotations at nighttime with multiple degradations [11][16] [22]. To address the concerns of the aforementioned methods, the proposed N2D3 explores patch-wise contrastive learning with physical guidance, so as to achieve degradation-aware Night2Day. N2D3 is free of human annotations and offers comprehensive structural modeling to provide faithful translation results.

3 Methods

Given nighttime image $\mathbf{I}_{\mathcal{N}} \in \mathcal{N}$ and daytime image $\mathbf{I}_{\mathcal{D}} \in \mathcal{D}$, the goal of Night2Day is to translate images from nighttime to daytime while preserving content semantic consistency. This involves the construction of a mapping function $\mathcal{F}$ with parameters θ, which can be formulated as $\mathcal{F}_\theta : \mathbf{I}_{\mathcal{N}} \to \mathbf{I}_{\mathcal{D}}$. Our method N2D3 is illustrated in Figure 2. To train a generator for Night2Day, we employ GANs as the overall learning framework to bridge the domain gap between nighttime and daytime. Our core design, consisting of the degradation disentanglement module and the degradation-aware contrastive learning module, aims to preserve the structure from the source images and suppress artifacts.

In this section, we first introduce physical priors in the nighttime environment, and then describe the degradation disentanglement module and the degradation-aware contrastive learning module, respectively.

3.1 Physical Priors for Nighttime Environment

The illumination degradations at night are primarily categorized as darkness, well-lit regions, high-light regions, and light effects. As shown in Figure 3, well-lit represents the diffused reflectance under normal light, while the light effects denote phenomena such as flare, glow, and specular reflections. Intuitively, these regions can be disentangled through the analysis of illumination distribution. Among these degradation types, darkness and high-light are directly correlated with illuminance and can be effectively disentangled through illumination estimation.

As a common practice, we estimate the illuminance map L by utilizing the maximum RGB channel of image $\mathbf{I}_{\mathcal{N}}$ as $L = \max_{c \in R,G,B} \mathbf{I}_{\mathcal{N}}^c$. Then k-nearest neighbors [4] is employed to acquire three clusters representing darkness, well-lit, and high-light regions. These clusters are aggregated as masks M_d, M_n, M_h. However, the challenge arises with light effects that are mainly related to

Figure 3: The first row displays nighttime images, while the second row shows the corresponding degradation disentanglement results. The color progression from **blue**, **light blue**, **green** to **yellow** corresponds to the following regions: darkness, well-lit, light effects, and high-light, respectively.

the illumination. Light effects regions tend to intertwine with well-lit regions when using only the illumination map, as they often share similar illumination densities. To disentangle light effects from well-lit regions, we need to introduce additional physical priors.

To extract the physical priors for disentangling light effects, we develop a photometric model derived from Kubelka-Munk theory [17]. This model characterizes the spectrum of light E reflected from an object as follows:

$$E(\lambda, x) = e(\lambda, x)(1 - \rho_f(x))^2 R_\infty(\lambda, x) + e(\lambda, x)\rho_f(x), \tag{1}$$

here x represents the horizontal component for analysis, while the analysis of the vertical component y is the same as the horizontal component. λ corresponds to the wavelength of light. $e(\lambda, x)$ signifies the spectrum, representing the illumination density and color. ρ_f stands for the Fresnel reflectance coefficient. R_∞ is the material reflectivity function, formulated as follows at a specific location $x = x_0$:

$$R(\lambda) = a(\lambda) - \sqrt{a(\lambda)^2 - 1}, a(\lambda) = 1 + \frac{k(\lambda)}{s(\lambda)}, \tag{2}$$

where $k(\lambda)$ and $s(\lambda)$ denote the absorption and scattering coefficients, respectively. This formulation implies that for any local pixels, the material reflectivity is determined if the material is given. Assuming C is the material distribution function, which describes the material type varying across locations, the material reflectivity R_∞ can be formulated as:

$$R_\infty(\lambda, x) = R(\lambda)C(x). \tag{3}$$

Since the mixture of light effects and well-lit regions has been obtained previously, the core of disentangling light effects from well-lit regions lies in separating the illumination $e(\lambda, x)$ and reflectance components $R(\lambda)C(x)$. Note that the Fresnel reflectance coefficient $\rho_f(x)$ approaches 0 in reflectance-dominating well-lit regions, while $\rho_f(x)$ approaches 1 in illumination-dominating light effects regions. According to Equation (1), the photometric model for the mixture of light effects and well-lit regions is formulated as:

$$E(\lambda, x) = \begin{cases} e(\lambda, x), & \text{if } x \notin \Omega \\ e(\lambda, x)R(\lambda)C(x), & \text{if } x \in \Omega \end{cases}, \tag{4}$$

where Ω denotes the reflectance-dominating well-lit regions.

Subsequently, we observe that the following color invariant response to the regions with high color saturation, which is suitable to extract the illumination:

$$N_{\lambda^m x^n} = \frac{\partial^{m+n-1}}{\partial \lambda^{m-1} \partial x^n} \{ \frac{1}{E(\lambda, x)} \frac{\partial E(\lambda, x)}{\partial \lambda} \}, \tag{5}$$

This invariant has the following characteristics:

$$\begin{aligned} N_{\lambda^m x^n} &= \frac{\partial^{m+n-2}}{\partial \lambda^{m-1} \partial x^{n-1}} \frac{\partial}{\partial x} \left\{ \frac{1}{E(\lambda, x)} \frac{\partial E(\lambda, x)}{\partial \lambda} \right\} \\ &= \frac{\partial^{m+n-2}}{\partial \lambda^{m-1} \partial x^{n-1}} \frac{\partial}{\partial x} \left\{ \frac{1}{e(\lambda, x)} \frac{\partial e(\lambda, x)}{\partial \lambda} + \frac{1}{R(\lambda)C(x)} \frac{\partial R(\lambda)C(x)}{\partial \lambda} \right\} \\ &= \frac{\partial^{m+n-1}}{\partial \lambda^{m-1} \partial x^n} \left\{ \frac{1}{e(\lambda, x)} \frac{\partial e(\lambda, x)}{\partial \lambda} \right\}. \end{aligned} \tag{6}$$

Equation (5) to Equation (6) demonstrate that the invariant $N_{\lambda^m x^n}$ captures the features only related to illumination $e(\lambda, x)$. Consequently, we assert that $N_{\lambda^m x^n}$ functions as a light effects detector because light effects are mainly related to the illumination. It allows us to design the illumination disentanglement module based on this physical prior.

3.2 Degradation Disentanglement Module

In this subsection, we will elucidate how to incorporate the invariant for extracting light effects into the disentanglement in computation. As common practice, the following second and third-order components, both horizontally and vertically, are taken into account in the practical calculation of the final invariant, which is denoted as N:

$$N = \sqrt{N_{\lambda x}^2 + N_{\lambda\lambda x}^2 + N_{\lambda y}^2 + N_{\lambda\lambda y}^2}. \tag{7}$$

here $N_{\lambda x}$ and $N_{\lambda\lambda x}$ can be computed through $E(\lambda, x)$ by simplifying Equation (5). The calculation of $N_{\lambda y}$ and $N_{\lambda\lambda y}$ are the same. Specifically,

$$N_{\lambda x} = \frac{E_{\lambda x}E - E_{\lambda}E_x}{E^2}, N_{\lambda\lambda x} \quad = \frac{E_{\lambda\lambda x}E^2 - E_{\lambda\lambda}E_x E - 2E_{\lambda x}E_{\lambda}E + 2E_{\lambda}^2 E_x}{E^3}, \tag{8}$$

where E_x and E_λ denote the partial derivatives of x and λ.

To compute each component in the invariant N, we develop a computation scheme starting with the estimation of E and its partial derivatives E_λ and $E_{\lambda\lambda}$ using the Gaussian color model:

$$\begin{bmatrix} E(x,y) \\ E_{\lambda}(x,y) \\ E_{\lambda\lambda}(x,y) \end{bmatrix} = \begin{bmatrix} 0.06, & 0.63, & 0.27 \\ 0.3, & 0.04, & -0.35 \\ 0.34, & -0.6, & 0.17 \end{bmatrix} \begin{bmatrix} R(x,y) \\ G(x,y) \\ B(x,y) \end{bmatrix}, \tag{9}$$

where x, y are pixel locations of the image. Then, the spatial derivatives E_x and E_y are calculated by convolving E with Gaussian derivative kernel g and standard deviation σ:

$$E_x(x, y, \sigma) = \sum_{t \in \mathbf{Z}} E(t, y) \frac{\partial g(x - t, \sigma)}{\partial x}, \tag{10}$$

where t denotes the index of the horizontal component x and **Z** represents set of integers. The spatial derivatives for $E_{\lambda x}$ and $E_{\lambda\lambda x}$ are obtained by applying Equation (10) to E_λ and $E_{\lambda\lambda}$. Then invariant N can be obtained following Equation (8) and Equation (7).

To extract the light effects, ReLU and normalization functions are first applied to filter out minor disturbances. Then, by filtering invariant N with the well-lit mask M_n, we obtain the light effects from the well-lit regions. The operations above can be formulated as:

$$M_{le} = \mathrm{ReLU}(\frac{N - \mu(N)}{\sigma(N)}) \odot M_n, \tag{11}$$

while the well-lit mask are refined: $M_n \leftarrow M_n - M_{le}$.

With the initial disentanglement in Section 3.1, we obtain the final disentanglement: M_d, M_n, M_h and M_{le}. All the masks are stacked to obtain the disentanglement map. Through the employment of the aforementioned techniques and processes, we successfully achieve the disentanglement of various degradation regions.

3.3 Degradation-Aware Contrastive Learning

For unpaired image translation, contrastive learning has validated its effectiveness for the preservation of content. It targets to maximize the mutual information between patches in the same spatial location from the generated image and the source image as below:

$$\ell(v, v^+, v^-) = -\log \frac{\exp(v \cdot v^+/\tau)}{\exp(v \cdot v^+/\tau) + \sum_{n=1}^{Q} \exp(v \cdot v_n^-/\tau)}, \tag{12}$$

v is the anchor that denotes the patch from the generated image. The positive example v^+ corresponds to the source image patch with the same location as the anchor v. The negative examples v^- represent patches with locations distinct from that of the anchor v. Q denotes the total number of negative examples. In our work, the key insight of degradation-aware contrastive learning lies in two folds: (1) How to sample the anchor, positive, and negative examples. (2) How to manage the focus on different negative examples.

Degradation-Aware Sampling. In this paper, N2D3 selects the anchor, positive, and negative patches under the guidance of the disentanglement results. Initially, based on the disentanglement mask obtained in the Section 3.2, we compute the patch count for different degradation types, denoting as $K_s, s \in [1, 4]$. Then, within each degradation region, the anchors v are randomly selected from the patches of generated daytime images $I_{\mathcal{N}\to\mathcal{D}}$. The positive examples v^+ are sampled from the same locations with the anchors in the source nighttime images $I_{\mathcal{N}}$, and the negative examples v^- are randomly selected from other locations of $I_{\mathcal{N}}$. For each anchor, there is one corresponding positive example and K_s negative examples. Subsequently, the sample set with the same degradation type will be assigned weights and the contrastive loss will be computed in the following steps.

Degradation-Aware Reweighting. Despite the careful selection of anchor, positive, and negative examples, the importance of anchor-negative pairs still differs within the same degradation. A known principle of designing contrastive learning is that the hard anchor-negative pairs (*i.e.*, the pairs with high similarity) should assign higher attention. Thus, weighted contrastive learning can be formulated as:

$$\ell(v, v^+, v^-, w_n) = -\log \frac{\exp(v \cdot v^+/\tau)}{\exp(v \cdot v^+/\tau) + \sum_{n=1}^{Q} w_n \exp(v \cdot v_n^-/\tau)}, \tag{13}$$

w_n denotes the weight of the n-th anchor-negative pairs.

The contrastive objective is depicted in the *Similarity Matrix* in Figure 2. The patches in different regions are obviously easy examples. We suppress their weights to 0, which transforms the similarity matrix into a blocked diagonal matrix with $diag(A_1, \ldots, A_4)$. Within each degradation matrix $A_s, s \in [1, 4]$, a soft reweighting strategy is implemented. Specifically, for each anchor-negative pair, we apply optimal transport to yield an optimal transport plan, serving as a reweighting matrix associated with the disentangled results. It can adaptively optimize and avoid manual design. The reweight matrix for each degradation type is formulated as:

$$\begin{aligned} &\min_{w_{ij}, i,j\in[1,K_s]} [\sum_{i=1}^{K_s} \sum_{j=1, i\neq j}^{K_s} w_{ij} \cdot \exp\left(v_i \cdot v_j^-/\tau\right)], \\ &\sum_{i=1}^{K_s} w_{ij} = 1, \sum_{j=1}^{K_s} w_{ij} = 1, i, j \in [1, K_s], \end{aligned} \tag{14}$$

The aforementioned operations transform the contrastive objective to the *Block Diagonal Similarity Matrix* depicted in Figure 2. As a common practice, our degradation-aware contrastive loss is applied to the S layers of the CNN feature extractor, formulated as:

$$\mathcal{L}_{DegNCE}(\mathcal{F}) = \sum_{l=1}^{S} \ell(v, v^+, v^-, w_n). \tag{15}$$

3.4 Other Regularizations

As a common practice, GANs are employed to bridge the domain gap between daytime and nighttime. The adversarial loss is formulated as:

$$\begin{aligned} \mathcal{L}_{adv}(\mathcal{F}) &= ||D(\mathbf{I}_{\mathcal{N}\to\mathcal{D}}) - 1||_2^2, \\ \mathcal{L}_{adv}(D) &= ||D(\mathbf{I}_{\mathcal{D}}) - 1||_2^2 + ||D(\mathbf{I}_{\mathcal{N}\to\mathcal{D}})||_2^2, \end{aligned} \tag{16}$$

where D denotes the discriminator network. The final loss function is formatted as :

$$\begin{aligned} \mathcal{L}(\mathcal{F}) &= \mathcal{L}_{adv}(\mathcal{F}) + \mathcal{L}_{DegNCE}(\mathcal{F}), \\ \mathcal{L}(D) &= \mathcal{L}_{adv}(D). \end{aligned} \tag{17}$$

4 Experiments

4.1 Experimental Settings

Datasets. Experiments are conducted on the two public datasets BDD100K [29] and Alderley [19]. **Alderley** dataset consists of images captured along the same route twice: once on a sunny day and another time during a stormy rainy night. The nighttime images in this dataset are often blurry due to the rainy conditions, which makes Night2Day challenging. **BDD100K** dataset is a large-scale high-resolution autonomous driving dataset. It comprises 100,000 video clips under various conditions. For each video, a keyframe is selected and meticulously annotated with details. We reorganized this dataset based on its annotations, resulting in 27,971 night images for training and 3,929 night images for evaluation.

Evaluation Metric. Following common practice, we utilize the *Fréchet Inception Distance* (FID) scores [7] to assess whether the generated images align with the target distribution. This assessment helps determine if a model effectively transforms images from the night domain to the day domain. Additionally, we seek to understand the extent to which the generated daytime images maintain structural consistency compared to the original inputs. To measure this, we employ SIFT scores, mIoU scores and LPIPS distance [32].

DownStream Vision Task. Two downstream tasks are conducted. In the Alderley dataset, GPS annotations indicate the locations of two images, one in the nighttime and the other in the daytime, as the same. We calculate the number of SIFT-detected key points between the generated daytime images and their corresponding daytime images to measure if the two images represent the same location. The BDD100K dataset includes 329 night images with semantic annotations. We employ Deeplabv3 pretrained on the Cityscapes dataset as the semantic segmentation model [2], then perform inference on our generated daytime images without any additional training and compute the mIoU (mean Intersection over Union).

Table 1: The quantitative results on Alderley and BDD100k. $\downarrow$ means lower result is better. $\uparrow$ means higher is better.

Dataset		Alderley			BDD100k		
Methods		FID$\downarrow$	LPIPS$\downarrow$	SIFT$\uparrow$	FID$\downarrow$	LPIPS$\downarrow$	mIoU$\uparrow$
Original	Conf./Jour.	210	-	3.12	101	-	15.63
CycleGAN[34]	ICCV 2017	167	0.706	3.36	51.7	0.477	13.42
StarGAN[3]	CVPR 2018	117	-	3.28	68.3	-	-
ToDayGAN[1]	ICRA 2019	104	0.770	4.14	43.8	0.577	16.77
UGATIT[15]	ICLR 2020	170	-	2.51	72.2	-	-
CUT[20]	ECCV 2020	64.7	0.707	6.78	55.5	0.583	9.30
ForkGAN[33]	ECCV 2020	61.2	0.759	12.1	37.6	0.581	11.81
AUGAN[18]	BMVC 2021	65.2	-	-	38.6	-	-
MoNCE[31]	CVPR 2022	72.7	0.737	6.35	40.2	0.502	17.21
Decent[27]	NIPS 2022	76.5	0.768	6.31	40.3	0.582	10.49
Santa[28]	CVPR 2023	67.1	0.757	6.93	36.9	0.559	11.03
N2D-LPNet[5]	CVPR 2023	-	-	-	69.1	-	-
EnlightenGAN [13]	TIP 2021	209.8	-	2.00	103.5	-	16.10
Zero-DCE [6]	TPAMI 2022	246.4	-	4.34	90.5	-	15.90
DeLight [21]	ECCV 2022	222.9	-	3.07	113.8	-	14.48
LLformer [23]	AAAI 2023	275.6	-	7.62	123.1	-	15.28
WCDM [12]	ToG 2023	239.6	-	7.10	124.3	-	16.32
GSAD [8]	NIPS 2023	214.7	-	6.29	116.0	-	15.76
N2D3(Ours)	-	**50.9**	**0.650**	**16.62**	**31.5**	**0.466**	**21.58**

4.2 Results on Alderley

We first apply Night2Day on the Alderley dataset, a challenging collection of nighttime images captured on rainy nights. In Figure 4, we present a visual comparison of the results. CycleGAN [34] and CUT [20] manage to preserve the general structural information of the entire image but often lose many fine details. ToDayGAN [1], ForkGAN [33], Decent [27], and Santa [28] tend to miss important elements such as cars in their results.

In Table 1, thirteen translation methods and three enhancement methods are compared, considering both visual effects and keypoint matching metrics. Our method showcases **an improvement of 10.3**

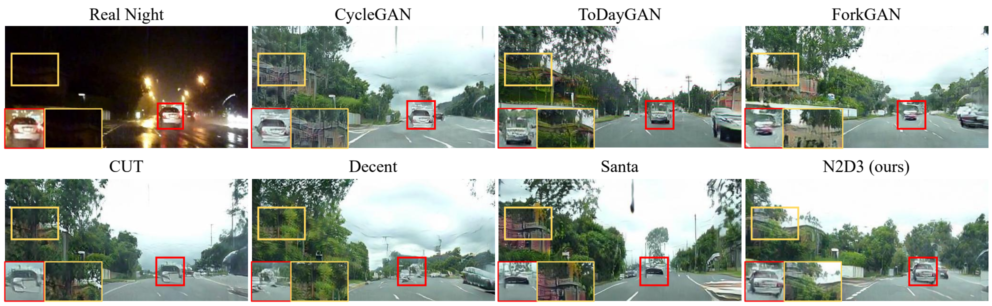


Figure 4: The qualitative comparison results on the Alderley dataset.

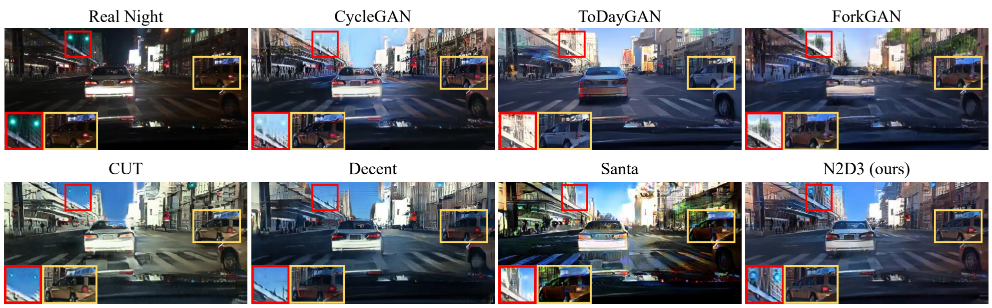


Figure 5: The qualitative comparison results on the BDD100K dataset.

in FID scores and 4.52 in SIFT scores compared to the previous state-of-the-art. This suggests that N2D3 successfully achieves photorealistic daytime image generation, underscoring its potential for robotic localization applications. The qualitative comparison results are demonstrated in Figure 4. In conclusion, N2D3 achieves top scores in both FID and LPIPS metrics, demonstrating its superiority in the Night2Day task. N2D3 excels in generating photorealistic daytime images while effectively preserving structures, even in challenging scenarios such as rainy nights in the Alderley.

4.3 Results on BDD100K

We conducted experiments on a larger-scale dataset, BDD100K, focusing on more general night scenes. The qualitative results can be found in Figure 5. CycleGAN, ToDayGAN, and CUT succeed in preserving the structure in well-lit regions. ForkGAN, Santa, and Decent demonstrate poor performance in such challenging scenes. Regretfully, none of them excel in handling light effects and exhibit weak performance in maintaining global structures. With a customized design specifically addressing light effects, our method successfully preserves the structure in all regions.

The quantitative results are presented in Table 1. As the scale of the dataset increases, all the compared methods show an improvement in their performance. Notably, N2D3 demonstrates the best performance with **a significant improvement of 5.4 in FID scores**, showcasing its ability to handle a broader range of nighttime scenes and establishing itself as the most advanced method in this domain.

We also investigate the potential of Night2Day in enhancing downstream vision tasks in nighttime environments using the BDD100K dataset. The quantitative results are summarized in Table 1. The enhancement methods demonstrate a slight improvement in segmentation results, while some image-to-image translation methods have a negative impact on performance. N2D3 exhibits the best performance in enhancing nighttime semantic segmentation with **a remarkable improvement of 5.95 in mIoU** compared to inferring the segmentation model directly on nighttime images.

In conclusion, N2D3 achieves top scores in both FID and LPIPS metrics, establishing itself as the most advanced method for the Night2Day task. It excels in generating photorealistic daytime images while preserving local and global structures. Moreover, the substantial improvement in nighttime semantic segmentation highlights its benefits for downstream tasks and its potential for wide-ranging applications.

(a) Ablation in FID

(b) Ablation in LPIPS

Figure 6: The quantitative results of ablation on the number of patches of the degradation-aware sampling.

Table 2: The quantitative results of ablation on the main component of degradation-aware contrastive learning. (a) denotes the degradation-aware sampling, and (b) denotes the degradation-aware reweighting. L and N denotes the invariant types.

Main Component		BDD100K		Alderley		
(a)	(b)	FID	LPIPS	FID	LPIPS	SIFT
✗	✗	55.5	0.583	64.7	0.707	6.78
✓	✗	36.9	0.495	56.6	0.698	16.52
✓	✓	31.5	0.466	50.9	0.650	16.62

Invariant Type		BDD100K		Alderley		
L	N	FID	LPIPS	FID	LPIPS	SIFT
✗	✗	55.5	0.583	64.7	0.707	6.78
✓	✗	49.1	0.592	62.9	0.726	9.83
✓	✓	31.5	0.466	50.9	0.650	16.62

4.4 Ablation Study

Ablation on the main component of degradation-aware contrastive learning. The core design of the degradation-aware contrastive learning module relies on two main components: (a) degradation-aware sampling, and (b) degradation-aware reweighting. As shown in Table 2, when degradation-aware sampling is exclusively activated, there is a noticeable decrease in FID on both datasets compared to the baseline (no components activated). Notably, the combination of degradation-aware sampling and reweighting achieves the lowest FID on both BDD100K and Alderley, indicating the effectiveness of degradation-aware sampling in conjunction with degradation-aware reweighting.

Ablation on the number of patches in the degradation-aware sampling. To explore the impact of the number of sampling patches in our method, we conduct an ablation study on the number of sampling patches with settings of 64, 128, 256, 512, and 1024 for degradation-aware sampling. The FID and LPIPS scores are evaluated, as shown in Figure 6. The optimal performance is achieved with 256 patches, and increasing the number of sampling patches beyond this point leads to a degradation in performance.

Ablation on the type of the invariant in disentanglement. To explore different invariants for obtaining degradation-disentangled prototypes, we conduct an ablation study on the type of invariant. As shown in Table 2, when L is enabled, the FID decreases from 55.5 to 49.1 on BDD100K and from 64.7 to 62.9 on Alderley. This suggests that incorporating illuminance maps helps in reducing the perceptual gap between generated and source nighttime images. When N is activated, there is a consistent improvement in FID on both datasets, indicating that considering physical priors invariant contributes to more realistic image generation. The combination of both illuminance map and physical prior invariant results in the lowest FID on both datasets, showcasing the complementary nature of these degradation types in improving contrastive learning.

5 Conclusion

This paper introduces a novel solution for the Night2Day image translation task, focusing on translating nighttime images to their corresponding daytime counterparts while preserving semantic consistency. To achieve this objective, the proposed method begins by disentangling the degradation presented in nighttime images, which is the key insight of our method. To achieve this, we contribute a degradation disentanglement module and a degradation-aware contrastive learning module. Our method outperforms the existing state-of-the-art, which shows the effectiveness of N2D3 and the superiority of the insight to disentangle the degradation.

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.

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[3] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.

[4] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[5] Zhentao Fan, Xianhao Wu, Xiang Chen, and Yufeng Li. Learning to see in nighttime driving scenes with inter-frequency priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4217–4224, 2023.

[6] 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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1780–1789, 2020.

[7] 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.

[8] Jinhui Hou, Zhiyu Zhu, Junhui Hou, Hui Liu, Huanqiang Zeng, and Hui Yuan. Global structure-aware diffusion process for low-light image enhancement. *Advances in Neural Information Processing Systems*, 36, 2024.

[9] Xueqi Hu, Xinyue Zhou, Qiusheng Huang, Zhengyi Shi, Li Sun, and Qingli Li. Qs-attn: Query-selected attention for contrastive learning in i2i translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18291–18300, 2022.

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei Efros. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[11] Somi Jeong, Youngjung Kim, Eungbean Lee, and Kwanghoon Sohn. Memory-guided unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6558–6567, 2021.

[12] Hai Jiang, Ao Luo, Haoqiang Fan, Songchen Han, and Shuaicheng Liu. Low-light image enhancement with wavelet-based diffusion models. *ACM Transactions on Graphics (TOG)*, 42(6):1–14, 2023.

[13] 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.

[14] Mikhail Kennerley, Jian-Gang Wang, Bharadwaj Veeravalli, and Robby T Tan. 2pcnet: Two-phase consistency training for day-to-night unsupervised domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11484–11493, 2023.

[15] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*, 2019.

[16] Soohyun Kim, Jongbeom Baek, Jihye Park, Gyeongnyeon Kim, and Seungryong Kim. Instaformer: Instance-aware image-to-image translation with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18321–18331, 2022.

[17] Paul Kubelka. Ein beitrag zur optik der farbanstriche (contribution to the optic of paint). *Zeitschrift fur technische Physik*, 12:593–601, 1931.

[18] 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.

[19] Michael J. Milford and Gordon. F. Wyeth. Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In *2012 IEEE International Conference on Robotics and Automation*, pages 1643–1649, 2012.

[20] Taesung Park, Alexei Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345, 2020.

[21] Aashish Sharma and Robby T Tan. Nighttime visibility enhancement by increasing the dynamic range and suppression of light effects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11977–11986, 2021.

[22] Seokbeom Song, Suhyeon Lee, Hongje Seong, Kyoungwon Min, and Euntai Kim. Shunit: Style harmonization for unpaired image-to-image translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):2292–2302, Jun. 2023.

[23] Tao Wang, Kaihao Zhang, Tianrun Shen, Wenhan Luo, Bjorn Stenger, and Tong Lu. Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2654–2662, 2023.

[24] Weilun Wang, Wengang Zhou, Jianmin Bao, Dong Chen, and Houqiang Li. Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14020–14029, 2021.

[25] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[26] Youya Xia, Josephine Monica, Wei-Lun Chao, Bharath Hariharan, Kilian Q Weinberger, and Mark Campbell. Image-to-image translation for autonomous driving from coarsely-aligned image pairs. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7756–7762. IEEE, 2023.

[27] Shaoan Xie, Qirong Ho, and Kun Zhang. Unsupervised image-to-image translation with density changing regularization. In *Advances in Neural Information Processing Systems*, 2022.

[28] Shaoan Xie, Yanwu Xu, Mingming Gong, and Kun Zhang. Unpaired image-to-image translation with shortest path regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10177–10187, June 2023.

[29] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

[30] Zhenjie Yu, Shuang Li, Yirui Shen, Chi Harold Liu, and Shuigen Wang. On the difficulty of unpaired infrared-to-visible video translation: Fine-grained content-rich patches transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1631–1640, June 2023.

[31] Fangneng Zhan, Jiahui Zhang, Yingchen Yu, Rongliang Wu, and Shijian Lu. Modulated contrast for versatile image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18280–18290, June 2022.

[32] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[33] Ziqiang Zheng, Yang Wu, Xinran Han, and Jianbo Shi. Forkgan: Seeing into the rainy night. In *European conference on computer vision*, pages 155–170. Springer, 2020.

[34] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

A Overview

This supplementary material is organized as follows. Appendix B provides additional details about the proof that the invariant $N_{\lambda^m x^n}$ is exclusively related to the illumination. Appendix C outlines the limitations and failure case of N2D3. Appendix D illustrates the implementation details, including N2D3 and other methods used in the experiments. Appendix E presents additional visualization results.

B More Proof Details

We provide a detailed proof process to demonstrate **how the invariant $N_{\lambda^m x^n}$ is exclusively related to the illumination and can function as the light effect detector.** First, consider the following equations, corresponding to Equation (5) in the main paper:

$$\begin{aligned} N_{\lambda^m x^n} &= \frac{\partial^{m+n-2}}{\partial\lambda^{m-1}\partial x^{n-1}}\frac{\partial}{\partial x}\{\frac{1}{E(\lambda,x)}\frac{\partial E(\lambda,x)}{\partial\lambda}\} \\ &= \frac{\partial^{m+n-2}}{\partial\lambda^{m-1}\partial x^{n-1}}\frac{\partial}{\partial x}\{\frac{1}{e(\lambda,x)}\frac{\partial e(\lambda,x)}{\partial\lambda} + \frac{1}{R(\lambda)C(x)}\frac{\partial R(\lambda)C(x)}{\partial\lambda}\}, \end{aligned} \tag{18}$$

by applying the additivity of linear differential operators, the first term represents the invariants only related to the illumination. The second term can be simplified by applying the chain rule as follows:

$$\begin{aligned} &\frac{\partial}{\partial x}\{\frac{1}{R(\lambda)C(x)}\frac{\partial R(\lambda)C(x)}{\partial\lambda}\} \\ &= \frac{1}{R(\lambda)^2C(x)^2}(\frac{\partial^2\{R(\lambda)C(x)\}}{\partial\lambda\partial x}\cdot R(\lambda)C(x) - \frac{\partial\{R(\lambda)C(x)\}}{\partial\lambda}\cdot\frac{\partial\{R(\lambda)C(x)\}}{\partial x}) \\ &= \frac{1}{R(\lambda)^2C(x)^2}(\frac{\partial R(\lambda)}{\partial\lambda}\frac{\partial C(x)}{\partial x}\cdot R(\lambda)C(x) - \frac{\partial R(\lambda)}{\partial\lambda}C(x)\cdot R(\lambda)\frac{\partial C(x)}{\partial x}) = 0. \end{aligned} \tag{19}$$

Finally, we conclude that the invariant $N_{\lambda^m x^n}$ is **exclusively related to the illumination** and can be formulated as follows:

$$\begin{aligned} N_{\lambda^m x^n} &= \frac{\partial^{m+n-2}}{\partial\lambda^{m-1}\partial x^{n-1}}\frac{\partial}{\partial x}\{\frac{1}{E(\lambda,x)}\frac{\partial E(\lambda,x)}{\partial\lambda}\} \\ &= \frac{\partial^{m+n-1}}{\partial\lambda^{m-1}\partial x^{n}}\{\frac{1}{e(\lambda,x)}\frac{\partial e(\lambda,x)}{\partial\lambda}\}. \end{aligned} \tag{20}$$

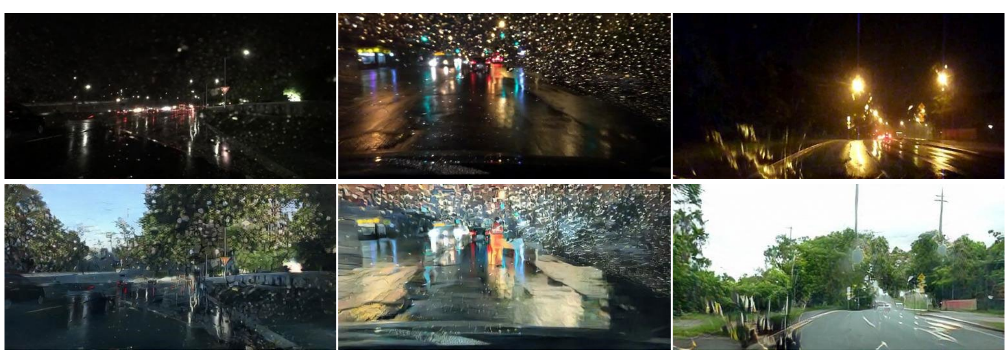


Figure 7: Failure Cases of N2D3: Our method struggles to handle various other types of degradation.

C Limitations and Failure Case

Despite the superior performance of N2D3 in Night2Day, it still exhibits certain limitations. On the one hand, this work focuses solely on addressing light degradation, while nighttime environments encompass various other types of degradation, including blur caused by rain, motion, and other

Figure 8: More disentanglement results. The first and third rows display nighttime images, while the second and fourth rows show the corresponding degradation disentanglement results. The color progression from **blue**, **light blue**, **green** to **yellow** corresponds to the following regions: darkness, well-lit, light effects, and high-light.

Figure 9: Qualitative comparison abalation results.

factors. Our method currently struggles to handle these situations effectively. On the other hand, the limitations of visible imaging in night vision arise from the scarcity of photos captured in low-light conditions, as illustrated by the failure cases presented inFigure 7. Future advancements in night vision will likely incorporate additional modalities, such as infrared images, radar, and other sensor data, to overcome these challenges and improve performance.

D Implementation Details

Training Details. We adopt the *resnet_9blocks*, a ResNetbased model with nine residual blocks, as the backbone for generator G. Additionally, we utilize the patch-wise discriminator D following PatchGAN[10]. To conduct degradation-aware contrastive learning on multiple layers, we extract features from 5 layers of the generator G encoder, as done in [20]. These layers include RGB pixels, the first and second downsampling convolution, and the first and fifth residual block. For the features of each layer, we apply a 2-layer MLP to acquire final 256-dimensional features. These features are then utilized in our degradation-aware contrastive learning.

All the comparison methods are reproduced using their released source code with default settings. Training procedures are consistent across all methods. All models are trained using the Adaptive Moment Estimation optimizer with an initial learning rate of 10^{-4}, a momentum of 0.9, and weight decay of 10^{-4}. For the BDD100K dataset, training consists of 10 epochs with the initial learning rate, followed by another 10 epochs with a decreased learning rate using the polynomial annealing procedure with a power of 0.9. On the Alderley dataset, given the limited training data compared to BDD100K, we extend the training to 20 epochs with the initial learning rate and an additional

Figure 10: More qualitative comparison results on the Alderley dataset.

20 epochs with the decayed learning rate. All the experiments are run on a single A100 GPU with 80GB of memory. Training our method with a smaller patch size and batch size on a device with less memory is feasible.

Evaluation Details. In the evaluation, we compute the *Fréchet Inception Distance* (FID) [7], Structural Similarity Index (SSIM) [25], and Learned Perceptual Image Patch Similarity (LPIPS) [32] scores on 256×512 images. Partial FID scores are provided by ForkGAN [33], and all SSIM and LPIPS scores are reproduced by us.

Semantic segmentation evaluation are conducted as follows. First, we use Deeplabv3 pretrained on the Cityscapes dataset as the semantic segmentation model [2]. The model is provided by `https://github.com/open-mmlab/mmsegmentation` with an R-18-D8 backbone and trained at a resolution of 512×1024. Second, we perform 512×1024 Night2Day translation to obtain the generation results. Finally, we infer the semantic segmentation on the generated daytime images.

E More Visualization Results

More Ablation Visualization Results. We provide ablation visualization results on both Alderley and BDD100K in Figure 9. The complete method is presented along with ablation studies on the invariant N and without degradation-aware reweighting. All the modules contribute to improving the ability to maintain semantic consistency.

More Disentanglement Results. We provide additional disentanglement results in Figure 8. Our disentanglement methods offer a comprehensive representation of different illumination degradation types in various nighttime scenes.

More Qualitative Comparison. We present more qualitative comparisons in Figure 10 and Figure 11 alongside other methods.Our method demonstrates visually pleasing results under various nighttime conditions.

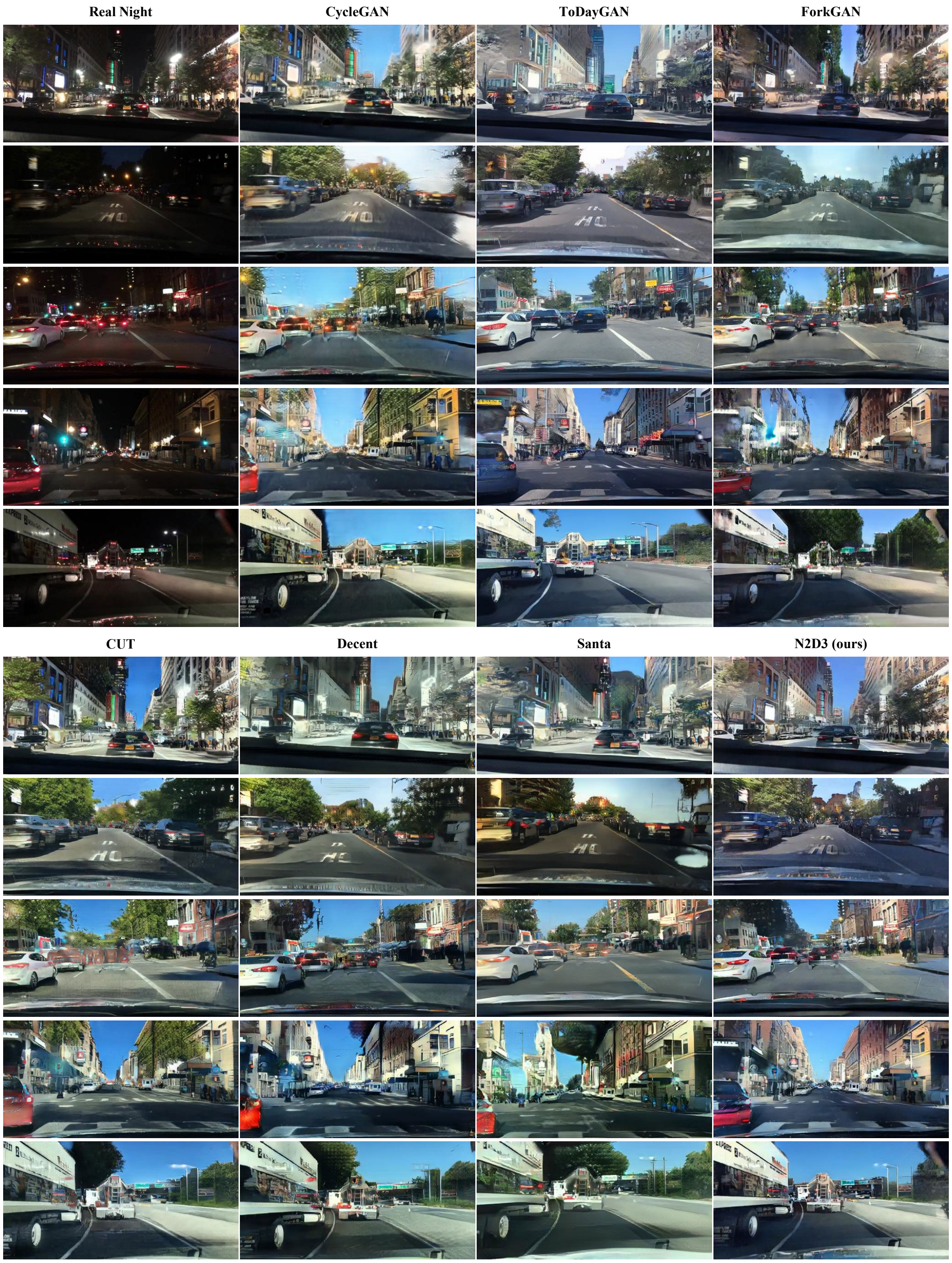


Figure 11: More qualitative comparison results on the BDD100K dataset.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We claim our main contribution as N2D3, which achieves SOTA performance by bridging the domain gap between nighttime and daytime in a degradation-aware manner.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss our limitation in degradations beyond light and low-light image scarcity in the appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide the full set of assumptions and complete proofs in both Section 3.1 and Appendix B .

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the information needed to reproduce the main experimental results is included in the Section 3 and Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Code will be released latter.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The training details and dataset information are provided in Section 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Error bars are not reported because it would be too computationally expensive. We report our results using a fixed random seed.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the compute resources in Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

   Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

   Answer: [Yes]

   Justification: The societal impacts are discussed in the manuscript and appendix.

   Guidelines:

   - The answer NA means that there is no societal impact of the work performed.
   - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our model does not have such risks, and all the datasets used in the experiments are open-source benchmarks in this field.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The code and data are properly credited, and the license and terms of use are explicitly mentioned and properly documented.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code introduced in the paper is well-documented, and the documentation is provided alongside it.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.