Light the Night: A Multi-Condition Diffusion Framework for Unpaired Low-Light Enhancement in Autonomous Driving

Jinlong Li[1*], Baolu Li[1*], Zhengzhong Tu[2], Xinyu Liu[1], Qing Guo[3], Felix Juefei-Xu[4], Runsheng Xu[5†], Hongkai Yu[1†]
[1]Cleveland State University [2] University of Texas at Austin
[3]Centre for Frontier AI Research (CFAR), A*STAR [4]New York University
[5]University of California, Los Angeles

Abstract

Vision-centric perception systems for autonomous driving have gained considerable attention recently due to their cost-effectiveness and scalability, especially compared to LiDAR-based systems. However, these systems often struggle in low-light conditions, potentially compromising their performance and safety. To address this, our paper introduces LightDiff, a domain-tailored framework designed to enhance the low-light image quality for autonomous driving applications. Specifically, we employ a multi-condition controlled diffusion model. LightDiff works without any human-collected paired data, leveraging a dynamic data degradation process instead. It incorporates a novel multi-condition adapter that adaptively controls the input weights from different modalities, including depth maps, RGB images, and text captions, to effectively illuminate dark scenes while maintaining context consistency. Furthermore, to align the enhanced images with the detection model's knowledge, LightDiff employs perception-specific scores as rewards to guide the diffusion training process through reinforcement learning. Extensive experiments on the nuScenes datasets demonstrate that LightDiff can significantly improve the performance of several state-of-the-art 3D detectors in night-time conditions while achieving high visual quality scores, highlighting its potential to safeguard autonomous driving.

1. Introduction

Driving at night is challenging for humans, even more so for autonomous vehicles, as shown in Fig. 1. On March 18, 2018, a catastrophic incident highlighted this challenge when an Uber Advanced Technologies Group self-driving vehicle struck and killed a pedestrian in Arizona [37].

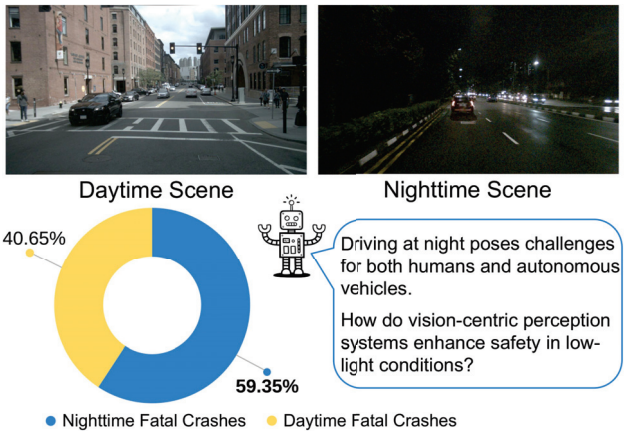


Figure 1. Nighttime driving scenarios pose a greater fatal threat than daytime. The fatal rate at night is much higher [4]. This paper aims to enhance nighttime images to improve the overall driving safety at night.

This incident, resulting from the vehicle's failure to detect the pedestrian under low-light conditions accurately, has brought the safety concerns of autonomous vehicles to the forefront, especially in such demanding environments. As vision-centric autonomous driving systems predominantly relying on camera sensors become more prevalent, addressing the safety implications of low-light conditions has become increasingly critical to ensure the overall safety of these vehicles.

One intuitive solution is to collect extensive night-time driving data. However, this approach is not only labor-intensive and costly, but it also risks impairing daytime model performance due to the differing image distributions between night and day. To navigate these challenges, we propose a Lighting Diffusion (LightDiff) model, a novel method that eliminates the need for manual data collection and maintains model performance during the daytime.

LightDiff aims to enhance low-light camera images, improving perception model performance. Utilizing a dy-

*Equal contribution. †Co-corresponding author, email address: xrxx3386@ucla.edu, h.yu19@csuohio.edu.

namic low-light-degradation process, LightDiff generates synthetic day-night image pairs from existing daytime data to train. We then employ Stable Diffusion [44] for its ability to produce high-quality visuals, effectively transforming night-time scenes into daytime equivalents. However, maintaining semantic consistency is critical in autonomous driving, a challenge with the original Stable Diffusion model. To overcome this, LightDiff incorporates multiple input modalities, such as estimated depth maps and camera image captions, coupled with a multi-condition adapter. This adapter intelligently determines the weighting of each input modality, ensuring the semantic integrity of the transformed images while keeping high visual quality. To guide the diffusion process not only to a direction that is visually brighter for humans but also for the perception model, we further finetune our LightDiff using reinforcement learning with perception-tailored domain knowledge in the loop. We conduct extensive experiments on the autonomous driving dataset nuScenes [7] and demonstrate that our LightDiff can significantly improve 3D vehicle detection Average Precision (AP) at nighttime by 4.2% and 4.6%, for two state-of-the-art models, BEVDepth [32] and BEVStereo [31], respectively. Our contributions are summarized as follows:

- We propose the Lighting Diffusion (LightDiff) model to enhance low-light camera images for autonomous driving, mitigating the need for extensive nighttime data collection and preserving daytime performance.
- We integrate multiple input modalities including depth maps and image captions with a proposed multi-condition adapter to ensure semantic integrity in image transformation while maintaining high visual quality. We employ a practical process that generates day-night image pairs from daytime data for efficient model training.
- We present a fine-tuning mechanism for LightDiff using reinforcement learning, incorporating perception-tailored domain knowledge (trustworthy LiDAR and statistical distribution consistency) to ensure that the diffusion process benefits both human visual perception and the perception model.
- Extensive experimentation with the nuScenes dataset demonstrates that LightDiff significantly improves 3D vehicle detection during the night and outperforms other generative models on multiple visual metrics.

2. Related Work

Dark Image Enhancement. Dark image enhancement aims to improve the visual quality and perceptibility of images suffering from dark conditions. It includes supervised methods [39, 43] that use paired datasets and unsupervised approaches [16, 28, 34, 35] that enhance images without such paired data. Some enhancement methods [8, 20–23] are developed to overcome the limitations in processing underexposed and/or overexposed regions in low-light conditions. There are some diffusion models for low-light image enhancement [11, 17, 49, 53], which explicitly integrate degradation prior and diffusion generative capability, but they require paired data in training.

Large Language Model in Vision. Vision & Language (VL) models [1, 3, 10, 12, 29, 30, 40] have shown obvious progress in computer vision. CLIP [40] acquires transferable visual concepts by natural language processing based supervision, learning knowledge from a large-scale dataset of image-caption pairs. Assisted by the language models, text/caption can be used to promote diverse computer vision tasks, such as CyCLIP [12] and unCLIP [41]. Because the VL models contain substantial visual and language understanding, they can be utilized to evaluate image quality [56]. This insight inspires us to leverage VL model related techniques for enhancing low-light images [12, 25, 41].

Diffusion-based Generative Models. The diffusion-based model [18] has achieved significant success in image synthesis through an iterative denoising process. Different diffusion-based methods have been developed for the text-to-image generation task [15, 45, 46, 55], with outstanding performance in computer vision. Unlike some diffusion based methods relying on text prompts like Dreambooth [45], the recent ControlNet [55] incorporates the spatial condition based control signals into the pre-trained text-to-image diffusion models. Using the strong Stable Diffusion [44] model as backbone which conducts the denoising process in the latent feature space, this paper makes efforts to enhance the dark visibility and address perception concerns to enhance the safety for driving at night.

3. Methodology

We aim to propose a general framework for low-light image enhancement that can benefit the perception in autonomous driving. To handle diverse driving-view scenarios, we exploit the strong generative prior imbued in the pre-trained Stable Diffusion model, which has been shown to deliver promising results for a variety of text-to-image and image-to-image tasks. To train the model, we built a versatile nighttime image generation pipeline that can simulate realistic low-light images to produce training data pairs (as detailed in Sec. 3.1). Then, we introduce our proposed (**LightDiff**) model in Sec. 3.2, a novel conditional generative model that can adaptively leverage various modalities of conditions—a low-light image, a depth map, and a text prompt—to predict the enhanced-light output. Fig. 2 depicts the entire pipeline of our proposed LightDiff architecture. To improve our model's task awareness, we introduce a reward policy that considers guidance from trustworthy LiDAR and statistical distribution consistency, further described in Sec. 3.3. Finally, we present a recurrent lighting inference strategy to further boost the results of our model during test time, which is explained in Sec. 3.4.

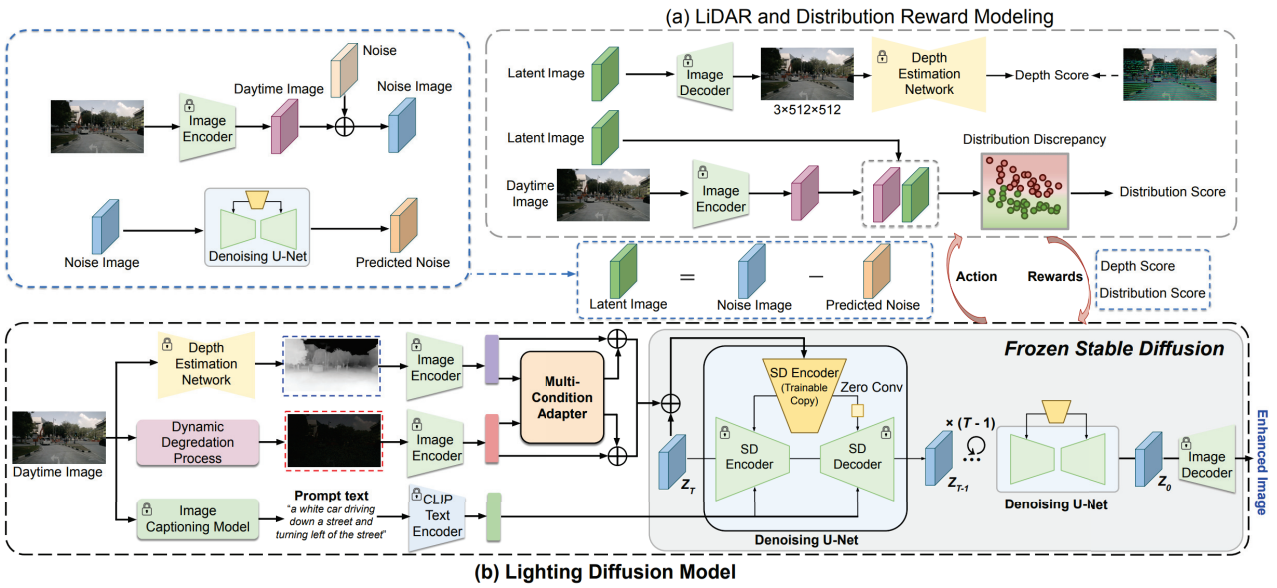


Figure 2. The architecture of our Lighting Diffusion model (LightDiff). During the training stage, a Training Data Generation pipeline enables the acquisition of triple-modality data without any human-collected paired data. Our LightDiff employs a Multi-Condition Adapter to dynamically weight multiple conditions, coupled with LiDAR and Distribution Reward Modeling (LDRM), allowing for perception-oriented control.

3.1. Training Data Generation

It is inherently challenging to collect nighttime-daytime paired images in dynamic driving scenarios. In response to this challenge and to introduce more controlled conditions, we build a novel training data generation pipeline. As illustrated in Fig. 3, this pipeline generates multi-modality paired data, including *1) instruction prompt*, *2) trustworthy depth map generated by LiDAR*, and *3) corresponding degraded dark light image*. Starting with a daytime image I_{day} as our target ground truth, we extract the text prompt by feeding it into a large image captioning model [5]. Meanwhile, we employ a pre-trained depth estimation network [42] to obtain the corresponding depth map. In common autonomous driving scenarios, where both LiDAR and camera sensors are provided, we project LiDAR point clouds onto the camera coordinate system as sparse points, which are then used as ground-truth supervision to train the depth estimation network. The pre-trained depth estimation network is frozen to be used for the training and testing of our lighting diffusion model. Unlike cameras, which are sensitive to illumination conditions, LiDAR maintains information consistency throughout both daytime and nighttime scenarios. Drawing inspiration from [9], we utilize a low-light-degradation transform T_{deg} to synthesize vivid dark-light images $T_{deg}(\mathrm{I}_{day})$, as depicted in Fig. 3. Specifically, we first transform the daytime image I_{day} into RAW data using the sRGB $\rightarrow$ RAW process [6]. Subsequently, we linearly attenuate the RAW image and introduce Shot and Read (S&R) Noise, as commonly found in camera imaging systems [6]. Finally, we apply the Image Signal Processing (ISP) pipeline to convert the low-light sensor measurement back to sRGB. The overall low-light-degradation transform T_{deg} can be simplified as:

$$T_{deg}(\mathrm{I}_{day}) = T_{ISP}(T_{sRGB \rightarrow RAW}(\mathrm{I}_{day}) + \mathrm{I}_{noise}), \quad (1)$$

which generates the degraded image I_{deg} similar to the dark nighttime image. We design a Dynamic Degradation Process adopting an online way with randomized combinations of parameter ranges of Eq. (1) to simulate the wider night driving scenes.

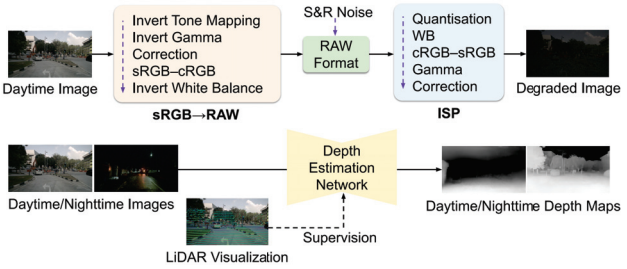


Figure 3. The pipeline of our Training Data Generation. The low-light-degradation transform [9] is exclusively implemented during the training stage. The trained depth estimation network will be frozen to be used for the training and testing stages of our lighting diffusion model.

3.2. Lighting Diffusion Model (LightDiff)

Our objective is to generate a pixel-level enhanced image which meticulously refines local textures and accurately reconstructs the global geometric outline of light details,

conditioned on the triplet of multi-modal input data generated using our data pipeline (Sec. 3.1). Unlike previous conditional generative models [44, 55], which only conditions on a *single* modality, such as depth map, canny edge, *etc*., our method recognizes and integrates the distinct contributions of each type of input modality towards the generation of the final output. Processed by the Image Encoder, the latent features from the degraded image I_{deg} and the depth map I_{dep}, denoted as $\mathrm{F}_{deg} \in \mathbb{R}^{H\times W\times C}$ and $\mathrm{F}_{dep} \in \mathbb{R}^{H\times W\times C}$ respectively, are fed into the proposed Multi-Condition Adapter (Sec. 3.2.2), which adaptively fuses multiple conditions based on the global contribution of each input modality. We adopt the ControlNet architecture [55] to learn the fused extra conditioning using a trainable copy of the UNet encoder, while leaving the backbone diffusion model frozen.

3.2.1 Preliminary: Stable Diffusion

We employ Stable Diffusion (SD), a large-scale text-to-image pre-trained latent diffusion model to achieve dark enhancement in dynamic driving scenarios. By definition, diffusion models generate data samples through a sequence of denoising steps that estimate the score of the data distribution. For improved efficiency and stabilized training, SD pretrains a variational autoencoder (VAE) [26] that compresses an image I into a latent z with encoder E and reconstructs it with decoder $\mathcal{D}$. Both the diffusion and denoising processes happen in the latent space. In the diffusion process, Gaussian noise with variance $\beta_t \in (0, 1)$ at time t is added to the encoded latent $z = E(I)$ to produce the noisy latent:

$$z_t = \sqrt{\bar{\alpha}_t} z + \sqrt{1-\bar{\alpha}_t}\chi, \tag{2}$$

where $\chi \sim \mathcal{N}(0, \mathbf{I})$, $\alpha_t = 1-\beta_t$, and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. When t is sufficiently large, the latent z_t approximates a standard Gaussian distribution. A network ϵ_θ is learned by predicting the noise ϵ conditioned on c_t (text prompts) at a randomly chosen time-step t. The optimization objective of the latent diffusion model is defined as:

$$\mathcal{L}_{LDM} = \mathbb{E}_{z,c_t,t,\epsilon}[||\epsilon - \epsilon_\theta(z_t, c_t, t)||_2^2], \tag{3}$$

where t is uniformly sampled and ϵ is sampled from the Gaussian distribution.

3.2.2 Multi-Condition Adapter

To discern the significance of different visual conditions, we introduce a novel multi-condition adapter, that is designed to dynamically weighs the conditions based on input data. Particularly, the latent features of the dark-light input $\mathrm{F}_{deg} \in \mathbb{R}^{H\times W\times C}$ with the paired depth map $\mathrm{F}_{dep} \in \mathbb{R}^{H\times W\times C}$ are concatenated as $\mathbf{F}_{(dep,deg)}$ and fed into a convolution layer. It is then reshaped to $\mathbb{R}^{2C\times(H\times W)}$ denoted as $\mathbf{F}^c_{(dep,deg)}$. A softmax layer is applied to the matrix multiplication of $\mathbf{F}^c_{(dep,deg)}$ and its transpose, obtaining the multi-condition weights $\mathbf{W} \in \mathbb{R}^{2C\times 2C}$:

$$w_{(dep,deg)} = \frac{\exp(\mathbf{F}^c_{dep} \cdot \mathbf{F}^c_{deg})}{\sum_c \exp(\mathbf{F}^c_{dep} \cdot \mathbf{F}^c_{deg})}, \tag{4}$$

where $w_{(dep,deg)}$ measures the impact of F_{deg} on F_{dep}. The transposed $\mathbf{W}$ is multiplied with $\mathbf{F}^c_{(dep,deg)}$, then reshaped to $\mathbb{R}^{2C\times H\times W}$. An element-wise sum operation with $\mathbf{F}^c_{(dep,deg)}$ yields the output $\mathrm{F}^{'}_{deg} \in \mathbb{R}^{2C\times H\times W}$:

$$\mathrm{F}^{'}_{deg} = \sum_c (w_{(dep,deg)}\mathbf{F}^c_{dep}) + \mathbf{F}^c_{deg}. \tag{5}$$

In the same way, we could obtain $\mathrm{F}^{'}_{dep} \in \mathbb{R}^{2C\times H\times W}$. The final output represents a weighted combination of all the conditions, capturing semantic dependencies between the multiple modalities. The multi-condition adapter is succinctly represented as:

$$\mathrm{F}^{'}_{deg}, \mathrm{F}^{'}_{dep} = \text{MC-Adaptor}(\mathrm{F}_{deg}, \mathrm{F}_{dep}). \tag{6}$$

3.2.3 Controlling the Stable Diffusion Model

Inspired by [55], we employ an additional conditioning network trained from scratch to encode additional condition information. We first use the encoder of Stable Diffusion's pre-trained VAE to map I_{deg} and I_{dep} into the latent space, obtaining the conditional latents F_{deg} and F_{dep}. The UNet denoiser in SD performs latent diffusion, which includes an encoder, a middle block, and a decoder. We create an additional copy of the UNet encoder (denoted in orange in Fig. 2(b)) to inject additional visual conditions. After being processed by the multi-condition adapter, the conditional latents $\mathrm{F}^{'}_{deg}$ and $\mathrm{F}^{'}_{dep}$ are concatenated with the randomly sampled noise z_t as inputs to the trainable copy of encoder. Their outputs are added back to the original UNet decoder, with a 1×1 convolutional layer (denoted as an orange rectangle in Fig. 2(b)) applied before the residual addition operation for each scale. During finetuning, the additional module and these 1×1 convolutional layers are optimized simultaneously. The entire network ϵ_θ learns to predict the noise z_n added to the noisy image z_t by minimizing the following latent diffusion objective:

$$\mathcal{L}_{\text{Lighting}} = \mathbb{E}_{z_t,c_t,c_d,t,\epsilon}[||\epsilon - \epsilon_\theta(z_t, c_t, c_d, t)||_2^2], \tag{7}$$

where c_d represents the condition combining the dark-light image and the depth map.

3.3. LiDAR and Distribution Reward Modeling

To achieve fine-grained task-oriented control, we introduce a reward policy that considers guidance from trustworthy

LiDAR and statistical distribution consistency during training our lighting diffusion model. We create a training schedule where the reward is applied only on predicted clean latent images z_I when the sampled time step t is less than a threshold τ. We leverage a frozen depth estimation network and apply a distribution-aware statistical consistency module to enforce distribution alignment. As shown in Fig. 2, the z_I is fed into the image decoder to generate the pixel-level image feature I_{pred} with the same shape as the original real daytime image. The depth estimation network predicts the depth map, whose misalignment metric (L_{Depth}) is the mean square error with the ground truth depth map by the trustworthy LiDAR point clouds.

To address the distribution gap between the enhanced lighting image and the real daytime image, we examine the relationship between statistical differences and feature distributions. Previous studies [19] have established a positive correlation between statistical differences and distribution disparities. As such, to minimize the discrepancy of feature distributions between z_I and z_{gt}, we introduce the distribution-aware statistical consistency module, utilizing the Maximum Mean Discrepancy (MMD) [14] distance (L_{MMD}) as a metric. Specifically, let $\mathcal{Z}_I = \{z_I^i\}$ and $\mathcal{Z}_{gt} = \{z_{gt}^i\}$ represent a set of enhanced lighting and real daytime features, respectively. The reward model takes the predicted clean latent image z_I as input and outputs two scalar rewards, namely depth and distribution scores. Following the Reinforcement Learning (RL) training strategy [27, 38], the agent, represented by the UNet denoiser ϵ_θ, presents a predicted clean latent image z_I and expects a response based on z_I. It takes the z_I and produces a reward determined by the reward model, thus concluding the episode. We minimize the following combined objective function in the RL training:

$$\begin{aligned}\mathcal{L}_{\text{obj}} = &\mathbb{E}_{z_t,c_t,c_d,t,\epsilon}\left[||\epsilon - \epsilon_\theta(z_t,c_t,c_d,t)||_2^2\right] \\ &+ \Phi^{RL}_{\mathbb{E}_{z_t,c_t,c_d,t,\epsilon}}\left(L_{\text{MMD}}(\mathcal{Z}_I,\mathcal{Z}_{gt}), L_{\text{Depth}}(z_I)\right),\end{aligned} \tag{8}$$

where $\Phi^{RL}_{\mathbb{E}_{z_t,c_t,c_d,t,\epsilon}}$ is the learned policy. This designed reward modeling will guide the training of our lighting diffusion model by leveraging the trustworthy LiDAR and statistical distribution consistency.

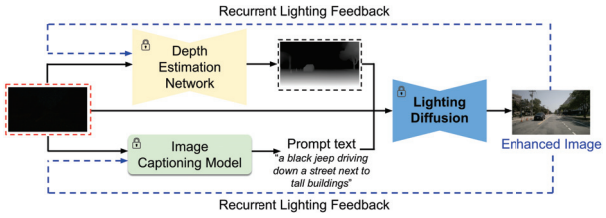


Figure 4. Illustration of the Recurrent Lighting Inference. It is designed to enhance the precision of generating text prompts and depth maps, thereby mitigating adverse effects on dark images.

3.4. Recurrent Lighting Inference

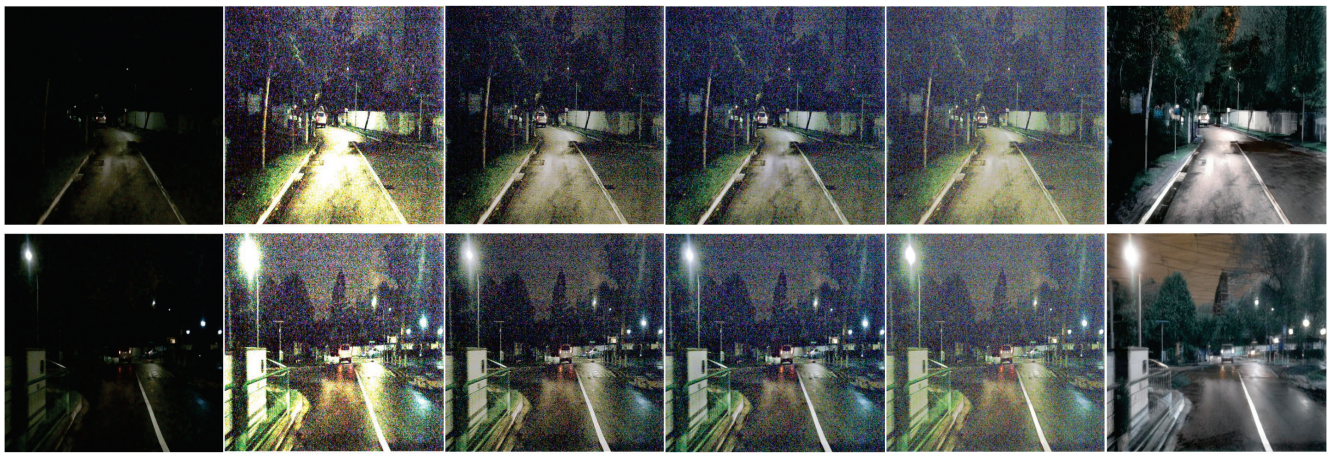
Real-night images, in contrast to clear daytime images, often suffer from low visibility and uneven light distribution. These conditions pose significant challenges for depth generation of the pre-trained depth estimation network, as well as the image captioning model. To address these issues, we implement an iterative feedback process that includes refining text prompts and tuning generated depth maps, as illustrated in Fig. 9. This process, executed in a loop with the depth estimation network, image captioning model, and lighting diffusion model remaining constant, aims to improve the accuracy of text prompts and refine depth maps for initial dark images, thereby enhancing the overall lighting results. Particularly, the procedure starts with feeding a real nighttime image into the depth estimation network and image captioning model to acquire an initial estimate of the corresponding text prompt and depth map. These inputs are then employed by the lighting diffusion model to produce an enhanced lighting image. Subsequently, we feed this initial enhanced image to replace the original nighttime image to further generate a refined text prompt and depth map, which are utilized as inputs for the next iteration. The cycle repeats until the similarity of the final generated images stabilizes, but in practice we found only two iterations are sufficient to generate a high-quality enhanced image.

4. Experiments

4.1. Experimental Setup

Datasets. To explore the low light enhancement for visual perception tasks on autonomous driving, we conduct experiments on the nuScenes dataset [7], which is one of the most popular autonomous driving datasets for multiple visual tasks. It consists of 700 scenes for training, 150 scenes for validation, and 150 scenes for testing. For each scene, it provides images with a resolution of $1,600 \times 900$ from 6 surrounding cameras (front, front left, front right, back, back left, back right) to cover the whole viewpoint, and a 360° LiDAR point cloud. Camera matrices including both intrinsic and extrinsic are provided, which establish a one-to-one correspondence between each 3D point and the 2D image plane. We select all 616 daytime scenes of the nuScenes training set containing total 24,745 camera front images as our training set. All 15 nighttime scenes in the nuScenes validation set containing total 602 camera front images are as our testing set.

Evaluation Metrics. We evaluate the low-light enhancement and 3D detection tasks in our experiments. For the quantitative assessment on low-light enhancement task, due to lack of paired day-night data in the real autonomous driving scenario, we employ nine no-reference image quality



(a) Night Images **(b)** RUAS-LOL [34] **(c)** SCI-difficult [35] **(d)** Zero-DCE++ [28] **(e)** URetinex-Net [50] **(f)** Ours

Figure 5. Visual comparison on the example nighttime images in the nuScenes validation set.

(a) Night Images **(b)** SCI-difficult [35] **(c)** Ours

Figure 6. Visualization of 3D detection results on the example nighttime images in the nuScenes validation set. We employ BEVDepth [32] as the 3D detector and visualize both the front view of camera and the Bird's-Eye-View.

evaluation (IQA) metrics, MUSIQ [24], NIQE [36], HyperIQA [47], ILNIQE [54], MANIQA [52], NIMA [48] and TReS [13]. In 3D perception task, we select the "Car" category as main object to report Average Precision (AP), along with Average Translation Error (ATE), Average Scale Error (ASE), Average Orientation Error (AOE) in experiments.

Training. We deploy the Training Data Generation method as described in Sec. 3.1 on the nuScenes daytime training set to obtain the triple modality paired data: 1) an instruction prompt, 2) a trustworthy depth map with LiDAR point cloud projection, and 3) a degraded dark image. We implement our LightDiff for 100 epochs with a batch size of 4 on a single NVIDIA RTX A6000 GPU. We utilize Adam as the optimizer with the learning rate of 1×10^{-5}. Following the setting [44, 55], we resize the input images and condition maps to 512 and adapt the pre-trained SD model [44] with the version of 2.1. To obtain the accurate depth map in the inference stage, we train a pre-trained Depth Estimation Network [42] based on the daytime and nighttime images and corresponding LiDAR point cloud projection of nuScenes training dataset.

Inference. Given the night images of the nuScenes validation set, different from the training stage utilizing the real LiDAR point clouds projection to help construct the estimated depth maps, we generate depth maps by a pre-trained depth estimation network. In addition, we apply our proposed Recurrent Lighting Inference (ReLI) to optimize their corresponding text prompts and depth maps.

Comparison Methods. In our experiments, we compare the generation quality, and 3D detection performance of our proposed LightDiff with other existing representative dark image enhancement related methods. We evaluate our approach by comparing it with prominent methods including supervised enhancement methods like Afifi et al. [2], URetinex-Net [50], SNR-Aware-LOLv1 [51], unsupervised enhancement methods like EnlightenGAN [22], CLIP-LIT [33], Zero-DCE++ [28], and diffusion-based methods like ShadowDiffusion [17], ExposureDiffusion [49]. Some methods have released their pretrained models on different datasets. In order to substantiate the exceptional perfor-

Table 1. Quantitative comparison of image quality on the nuScenes nighttime validation set. The best and second performance are marked in **red** and **blue**.

Type	Methods	MUSIQ↑	HyperIQA↑	MANIQA↑	NIMA↑	TReS↑	ILNIQE↓	NIQE↓
	Night Image	16.295	0.229	0.034	3.197	20.753	35.627	3.700
Supervised	Afifi et al. [2]	13.562	0.207	0.045	3.715	19.731	32.207	8.388
	URetinex-Net [50]	24.086	0.244	0.065	**4.503**	53.571	38.972	4.460
	SNR-Aware-LOLv1 [51]	13.591	0.229	0.032	3.947	22.893	**30.119**	8.898
	SNR-Aware-LOLv2real [51]	13.591	0.230	0.032	3.947	23.025	**30.119**	8.898
	SNR-Aware-LOLv2synthetic [51]	14.528	0.219	0.026	3.414	12.571	38.652	8.967
	SNR-Aware-nuScene [51]	16.928	0.241	0.056	3.767	24.013	35.087	7.664
	ExposureDiffusion [49]	20.365	0.204	0.026	3.977	10.795	50.562	5.020
	ShadowDiffusion [17]	**40.695**	**0.446**	0.078	3.808	**67.086**	57.984	5.607
Unsupervised	Zero-DCE [16]	24.562	0.250	0.064	4.360	52.988	35.804	4.316
	Zero-DCE++ [28]	20.276	0.242	0.059	4.375	49.553	36.025	4.232
	RUAS-LOL [34]	24.587	0.257	0.056	4.311	43.836	52.761	6.031
	RUAS-MIT5K [34]	14.965	0.234	0.041	4.046	34.945	46.425	4.506
	RUAS-DarkFace [34]	20.277	0.248	0.067	4.382	45.488	55.351	5.595
	SCI-easy [35]	15.240	0.250	0.044	4.004	33.928	31.679	3.975
	SCI-medium [35]	15.513	0.241	0.056	4.374	50.316	36.317	4.423
	SCI-difficult [35]	34.718	0.260	**0.081**	4.370	58.037	34.583	5.050
	EnlightenGAN [22]	20.686	0.242	0.070	4.383	42.829	40.080	4.307
	LESNet [23]	19.410	0.205	0.032	3.477	14.453	32.784	7.905
	CLIP-LIT [33]	23.805	0.229	0.064	4.402	49.557	42.560	4.701
Unsupervised (retrained)	SCI [35]	14.781	0.238	0.044	3.909	34.819	34.220	4.118
	EnlightenGAN [22]	16.334	0.239	0.035	3.309	24.654	33.294	**3.397**
	CLIP-LIT [33]	16.288	0.229	0.033	3.206	20.766	35.681	3.703
	Ours	**51.674**	**0.407**	**0.086**	**4.594**	**58.622**	**20.250**	**3.516**

Table 2. 3D detection comparison on the nuScenes nighttime validation set. Both BEVDepth [32] and BEVStereo [31] are trained using the nuScenes daytime training set. The best and second performance are marked in **red** and **blue**. * indicates that it has been retrained on the nuScenes training set.

Methods	BEVDepth [32]				BEVStereo [31]			
	AP↑	ATE↓	ASE↓	AOE↓	AP↑	ATE↓	ASE↓	AOE↓
Night Image	0.134	0.787	0.195	**0.957**	0.124	0.746	**0.205**	0.714
SCI-diffcult [35]	0.067	0.828	0.187	1.071	0.032	0.764	0.239	0.774
Zero-DCE++ [28]	0.089	0.826	0.197	1.029	0.077	0.780	0.224	0.787
URetinex-Net [50]	0.053	0.831	0.184	1.114	0.035	0.782	0.243	0.803
ExposureDiffusion [49]	0.040	0.829	**0.179**	1.092	0.035	0.796	0.244	0.769
ShadowDiffusion [17]	0.072	0.851	0.184	1.242	0.082	0.789	0.224	0.718
SNR-Aware* [51]	0.089	0.817	0.193	1.088	0.072	0.773	0.251	0.742
SCI* [35]	0.062	0.829	0.181	1.133	0.034	0.783	0.235	0.801
EnlightenGAN* [22]	**0.138**	**0.786**	0.193	**0.948**	**0.128**	**0.743**	**0.204**	**0.680**
CLIP_LIT* [33]	0.131	0.791	0.199	0.972	0.121	0.753	0.211	0.739
Ours	**0.176**	**0.774**	**0.180**	1.108	**0.170**	**0.690**	0.210	**0.620**

mance of our approach impartially, we engage in a comparative analysis with these pretrained models. Moreover, we furnish the performance evaluations of retrained unsupervised methods, executed on the identical training set as our methodology, thereby contributing to a more comprehensive validation.

4.2. Comparison Results

Visual comparison. We present visual comparisons of some samples from the nuScenes nighttime validation set in Fig. 5. Our method consistently produces visually pleasing results with improved color and eliminated noise. Moreover, our method excels in handling challenging dark regions, restoring clear texture details and satisfactory luminance without introducing any noise, while other methods may either fail to address such dark regions or produce unsatisfactory results with visible noise. Specifically, we can see that compared to RUAS-LOL [34] and SCI-difficult [35], our method produces results without over-exposure or under-exposure. Our results exhibit better color contrast and input-output consistency in global regions.

Figure 7. Visual showcase of our LightDiff with and without the Multi-Condition Adapter. The input for ControlNet [55] remains consistent, comprising the same text prompt and depth map. Multi-Condition Adapter makes better color contrast and richer details during enhancement.

Quantitative Comparison. It is unattainable to collect nighttime-daytime paired images in real dynamic driving scenarios, currently we rely on several non-reference image quality evaluation (IQA) metric to evaluate the quantitative results. The quantitative comparison on the nuScenes nighttime validation set is presented in Table 1. Our method achieve the best performance in the four no-reference IQA metrics when compared to other methods, demonstrating the satisfactory image quality of our results.

3D perception Comparison and Visualization. For 3D perception task, we only enhance the front camera view of the nuScene nighttime validation set, while other five

camera-view are keep the original darkness. We utilize two 3D perception state-of-the-art methods BEVDepth [32] and BEVStereo [31] trained on the nuScenes dayitme training set, which is more effective to collect and annotate in the real-world driving scenario to evaluate the car detection with our effect of generated qualify on perception performance. We show the quantitative comparison of 3D perception performance on the nuScenes nighttime validation set in Table 2. Compared to results on the original nighttime images, by applied our enhanced images, the BEVDepth and BEVStereo can achieve 17.6% AP and 17.0% AP, respectivley, which have the improvement of 4.2% AP and 4.6% AP. Without any extra training requirement, our proposed method can improve the perception performance for current models by directly applied our generated enhanced images. But some comparison enhancement methods like SCI [35] and Zero-DCE++ [28], show the negative effect on 3D perception performance to make a performance drop. We visualize some 3D detection results on front camera-view and Bird's-Eye-View (BEV) in Fig. 6. Our proposed LightDiff not only help the driver see more clear in the darkness, but also help the deep learning perception detect more accurately in the challenging real-world dark conditions.

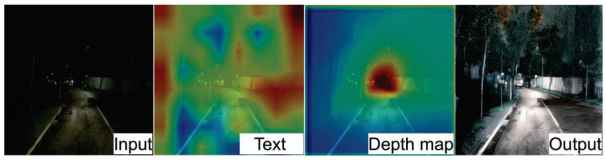


Figure 8. The examples of attention map for different modality inputs.

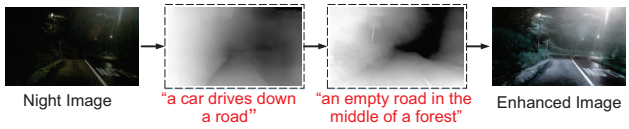


Figure 9. Illustration of enhanced multi-modality generation through Recurrent Lighting Inference (ReLI). Improved accuracy in text prompts and depth map prediction are achieved by invoking the ReLI once.

4.3. Ablation Study

To validate the effectiveness of our proposed components, we provide the quantitative comparison on 3D perception and dark enhancement tasks in Table 4. The visual comparison results of Fig. 7 show the effectiveness of discerning the significance of different visual conditions. The heatmaps in Fig. 8 illustrate the correlation of each image pixel with the two different modality inputs. The Table 3 unequivocally demonstrates the beneficial impact of each modality input within our LightDiff. We present the effectiveness of Recurrent Lighting Inference (ReLI), which can optimize the accuracy of multi-modality generation effectively in Fig. 9.

Table 3. Ablation study for multi-modality inputs on the nuScenes nighttime validation set. $*$ indicates that it has been retrained on the nuScenes training set.

Modality Input	3D Detection		Dark Enhancement			
	AP↑	ASE↓	MUSIQ↑	TReS↑	ILNIQE↓	NIQE↓
Image-only	0.119	0.190	40.103	30.472	36.789	5.717
w/o. Text	0.146	0.191	42.380	31.367	35.761	5.650
w/o. Depth Map	0.139	0.188	43.400	33.504	32.974	5.419
Image-only*	0.132	0.182	44.766	49.304	23.471	4.353
w/o. Text*	0.165	0.183	47.700	36.723	30.064	5.144
w/o. Depth Map*	0.122	**0.174**	47.286	43.340	21.655	**3.358**
Ours	**0.176**	0.180	**51.674**	**58.622**	**20.250**	3.516

Table 4. Ablation study for each proposed component on the nuScenes nighttime validation set.

	3D Detection		Dark Enhancement			
	AP↑	ASE↓	MUSIQ↑	TReS↑	ILNIQE↓	NIQE↓
w/o. LDRM	0.151	0.197	41.387	34.056	25.353	3.569
w/o. MC-Adapter	0.152	0.183	47.370	52.342	21.666	3.841
w/o. ReLI	0.163	0.184	48.277	54.462	22.423	4.085
Ours	**0.176**	**0.180**	**51.674**	**58.622**	**20.250**	**3.516**

It indicates that our LightDiff can produce better color contrast and richer details with our Multi-Conditional Adapter. The result in Table 4 clearly justifies the positive effect of each proposed component of our LightDiff.

5. Conclusions

This paper introduces LightDiff, a domain-tailored framework designed to enhance the low-light image quality for autonomous driving applications, mitigating the challenges faced by vision-centric perception systems. By leveraging a dynamic data degradation process, a multi-condition adapter for diverse input modalities, and perception-specific score guided reward modeling using reinforcement learning, LightDiff significantly enhances the image quality and 3D vehicle detection in nighttime on the nuScenes dataset. This innovation not only eliminates the need for extensive nighttime data but also ensures semantic integrity in image transformation, demonstrating its potential to enhance safety and reliability in autonomous driving scenarios. Without realistic paired day-night images, synthesizing dark driving images with vehicle lights is quite difficult, limiting the research in this field. Future research can be focused on a better collection or generation of the high-quality training data.

Acknowledgments: This work is supported in part by NSF 2215388, and the National Research Foundation, Singapore, and DSO National Laboratories under the AI Singapore Programme (No: AISG2-GC-2023-008).

References

[1] Rabab Abdelfattah, Qing Guo, Xiaoguang Li, Xiaofeng Wang, and Song Wang. Cdul: Clip-driven unsupervised learning for multi-label image classification. In *IEEE/CVF International Conference on Computer Vision*, pages 1348–1357, 2023. 2

[2] Mahmoud Afifi, Konstantinos G Derpanis, Bjorn Ommer, and Michael S Brown. Learning multi-scale photo exposure correction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9157–9167, 2021. 6, 7

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2

[4] Imran Ashraf, Soojung Hur, Muhammad Shafiq, and Yongwan Park. Catastrophic factors involved in road accidents: Underlying causes and descriptive analysis. *PLoS one*, 14 (10):e0223473, 2019. 1

[5] William Berrios, Gautam Mittal, Tristan Thrush, Douwe Kiela, and Amanpreet Singh. Towards language models that can see: Computer vision through the lens of natural language. *arXiv:2306.16410*, 2023. 3

[6] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11036–11045, 2019. 3

[7] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 2, 5

[8] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4):2049–2062, 2018. 2

[9] Ziteng Cui, Guo-Jun Qi, Lin Gu, Shaodi You, Zenghui Zhang, and Tatsuya Harada. Multitask aet with orthogonal tangent regularity for dark object detection. In *IEEE/CVF International Conference on Computer Vision*, pages 2553–2562, 2021. 3

[10] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022. 2

[11] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9935–9946, 2023. 2

[12] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719, 2022. 2

[13] S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1220–1230, 2022. 6

[14] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25): 723–773, 2012. 5

[15] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 2

[16] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1780–1789, 2020. 2, 7

[17] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14049–14058, 2023. 2, 6, 7

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[19] Yang Hou, Qing Guo, Yihao Huang, Xiaofei Xie, Lei Ma, and Jianjun Zhao. Evading deepfake detectors via adversarial statistical consistency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12271–12280, 2023. 5

[20] Junjie Hu, Xiyue Guo, Junfeng Chen, Guanqi Liang, Fuqin Deng, and Tin Lun Lam. A two-stage unsupervised approach for low light image enhancement. *IEEE Robotics and Automation Letters*, 6(4):8363–8370, 2021. 2

[21] Jie Huang, Yajing Liu, Feng Zhao, Keyu Yan, Jinghao Zhang, Yukun Huang, Man Zhou, and Zhiwei Xiong. Deep fourier-based exposure correction network with spatial-frequency interaction. In *European Conference on Computer Vision*, pages 163–180. Springer, 2022.

[22] 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. 6, 7

[23] Yeying Jin, Wenhan Yang, and Robby T Tan. Unsupervised night image enhancement: When layer decomposition meets light-effects suppression. In *European Conference on Computer Vision*, pages 404–421. Springer, 2022. 2, 7

[24] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021. 6

[25] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 2

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114*, 2013. 4

[27] Ngan Le, Vidhiwar Singh Rathour, Kashu Yamazaki, Khoa Luu, and Marios Savvides. Deep reinforcement learning in computer vision: a comprehensive survey. *Artificial Intelligence Review*, pages 1–87, 2022. 5

[28] Chongyi Li, Chunle Guo, and Chen Change Loy. Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4225–4238, 2021. 2, 6, 7, 8

[29] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705, 2021. 2

[30] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2

[31] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with dynamic temporal stereo. *arXiv:2209.10248*, 2022. 2, 7, 8

[32] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *arXiv:2206.10092*, 2022. 2, 6, 7, 8

[33] Zhexin Liang, Chongyi Li, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Iterative prompt learning for unsupervised backlit image enhancement. In *IEEE/CVF International Conference on Computer Vision*, pages 8094–8103, 2023. 6, 7

[34] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10561–10570, 2021. 2, 6, 7

[35] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5637–5646, 2022. 2, 6, 7, 8

[36] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012. 6

[37] National Transportation Safety Board. Inadequate safety culture contributed to uber automated test vehicle crash. https://www.ntsb.gov/news/press-releases/Pages/NR20191119c.aspx, 2019. Accessed: 2023-11-16. 1

[38] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022. 5

[39] Jongchan Park, Joon-Young Lee, Donggeun Yoo, and In So Kweon. Distort-and-recover: Color enhancement using deep reinforcement learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5928–5936, 2018. 2

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[41] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*, 1(2):3, 2022. 2

[42] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2020. 3, 6

[43] Wenqi Ren, Sifei Liu, Lin Ma, Qianqian Xu, Xiangyu Xu, Xiaochun Cao, Junping Du, and Ming-Hsuan Yang. Low-light image enhancement via a deep hybrid network. *IEEE Transactions on Image Processing*, 28(9):4364–4375, 2019. 2

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 4, 6

[45] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2

[46] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[47] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3667–3676, 2020. 6

[48] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8): 3998–4011, 2018. 6

[49] Yufei Wang, Yi Yu, Wenhan Yang, Lanqing Guo, Lap-Pui Chau, Alex C Kot, and Bihan Wen. Exposurediffusion: Learning to expose for low-light image enhancement. In *IEEE/CVF International Conference on Computer Vision*, pages 12438–12448, 2023. 2, 6, 7

[50] Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. Uretinex-net: Retinex-based

deep unfolding network for low-light image enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5901–5910, 2022. 6, 7

[51] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Snr-aware low-light image enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17714–17724, 2022. 6, 7

[52] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022. 6

[53] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In *IEEE/CVF International Conference on Computer Vision*, pages 12302–12311, 2023. 2

[54] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015. 6

[55] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 4, 6, 7

[56] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14071–14081, 2023. 2