Purifying Low-light Images via Near-Infrared Enlightened Image

Renjie Wan, Boxin Shi, Wenhan Yang, Bihan Wen, Ling-Yu Duan, Alex C. Kot, *Fellow, IEEE*

***Abstract*—Cameras usually produce low-quality images under low-light conditions. Though many methods have been proposed to enhance the visibility of low-light images, they are mainly designed for illumination correction and less capable of suppressing the artifacts. In this paper, we propose to enhance the visibility and suppress artifacts by purifying low-light images under the guidance of the NIR enlightened image captured by using the near-infrared light as compensation. Specifically, we introduce a disentanglement framework to disentangle the structure and color components from the NIR enlightened and RGB images, respectively. Correspondingly, we introduce a new dataset with the RGB and NIR enlightened images for training and evaluation purposes. The experimental results show that our proposed method achieves promising results.**

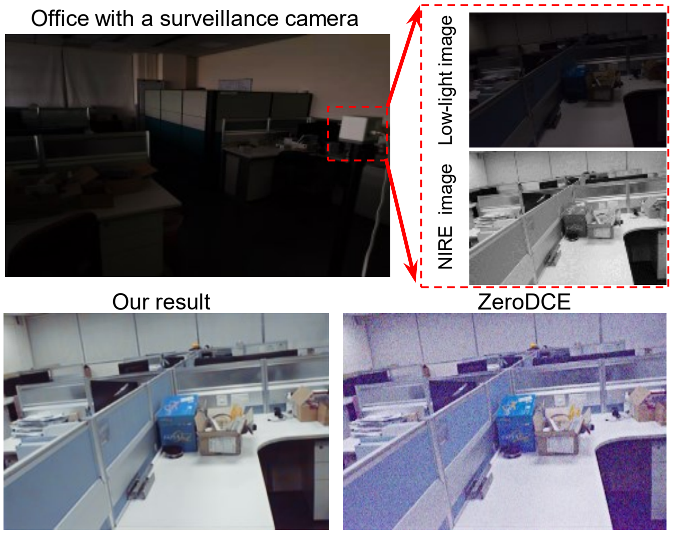

Figure 1: Given a low-light image captured under the visible mode and the near-infrared enlightened image captured under the grayscale/night-vision mode both by a surveillance camera (the region labeled by the red box), our method enhances the low-light images with better visibility and quality than ZeroDCE [1].

I. Introduction

THE wide usage of camera sensors has made photography to be a ubiquitous part of the human experience. However, due to the size limitation of some devices (*e.g.*, mobile phones and surveillance cameras), the aperture size built into these devices is restricted, which limits the amount of light received by camera sensors and leads to artifacts. As a result, most commercial cameras can only produce low-light images dominated by noise and artifacts for low-light scenes (Figure 1). Thus, purifying low-light images to produce an image with high visibility and fewer artifacts becomes a meaningful task.

By assuming the reflectance component as the well-exposed image, most Retinex-based methods [7], [8] have already been able to correct the illumination well using different learning strategies. However, the reflectance component is far from a well-exposed image for most off-the-shelf devices. For example, since the surveillance image always needs compression before uploading to the cloud or downloading to mobile devices, they largely suffer from compression artifacts for low-light images, which cannot faithfully preserve the structure information. Without considering various degradation during the image formation or transmission process, existing image enhancement methods (*e.g.*, ZeroDCE [1] in Figure 1) cannot get clean results and may even amplify artifacts during the brightness correction process. Supervised low-light image enhancement methods can suppress the artifacts [52]. However, the strict requirement for the paired ground truth limits their practicability in some changing environments. Though the unpaired and unsupervised fashion like CycleGAN [30], EnlightenGAN [31], or DPE [67] can alleviate the strict requirement for training pairs in supervised methods, the learning-based strategy alone without physical guidance is not able to suppress artifacts [32].

Instead of solely relying on the paired ground truth with normal brightness or the learning strategy, we propose to utilize the NIR enlightened (NIRE) image with the information from visible and NIR bands as guidance. Due to the invisibility to human eyes and effectiveness in enlightening the environment, near-infrared light has been utilized by different devices to compensate for visible light. With more light in the near-infrared spectrum, the NIRE image effectively suppresses artifacts and provides reliable guidance for the whole image enhancement process [34]. For example, recently proposed methods [55], [35] recover the information of the visible band by extracting them from the NIRE image. However, since the visible and NIR information is highly mixed during the formation process, it may be difficult to accurately obtain the color information of the visible band for some cases.

Instead of extracting information of the visible band from NIRE images like previous methods [55], [35], we propose to purify the low-light images in a weakly-supervised manner via the disentanglement of color, artifact, and structure com-

Renjie Wan is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong.

Alex C. Kot and Bihan Wen are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

Wenhan Yang is with the Peng Cheng Laboratory, China.

Boxin Shi and Ling-Yu Duan are with the National Engineering Laboratory of Video Technology, Department of Computer Science and Technology, Peking University, China.

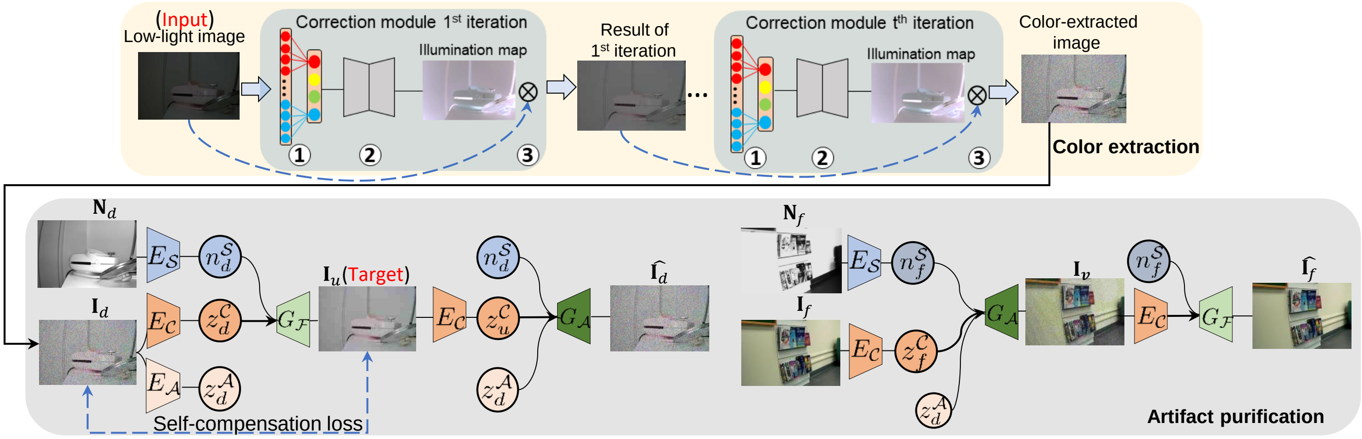

Figure 2: The framework of our proposed approach. With the NIRE image as guidance, we adopt the disentanglement framework with the self-compensation loss for the low-light image artifact purification. To facilitate the self-compensation, the color extraction module extracts the color information from low-light conditions by iteratively correcting the image illumination. $\mathbf{I}_d$ denotes the color-extracted image and $\mathbf{I}_f$ denotes the artifact-free image for unpaired training. $\mathbf{N}_d$ and $\mathbf{N}_f$ denotes their corresponding NIRE images, respectively. For the color extraction stage, ①, ②, and ③ denote the feature estimation layer, illumination output layer, and the extraction layer, respectively.$E_{\mathcal{S}}$, $E_{\mathcal{C}}$, and $E_{\mathcal{A}}$ denotes the encoder for structure, color, and artifacts, respectively. $n_d^{\mathcal{S}}$, $z_d^{\mathcal{C}}$, $z_d^{\mathcal{A}}$, $z_u^{\mathcal{C}}$, $n_f^{\mathcal{S}}$, and $z_f^{\mathcal{C}}$ denote the disentangled latent factors. More details about the network structure and latent factors can be found in Figure 6 and Section IV-B.

ponents from low-light images and NIRE images, respectively. Since the latent factors related to the color, structure, and artifact can be highly entangled and mixed in the examples from the real world, the learned representations are consequently prone to mistakenly preserve the confounding of the factors [33], leading to the color and structure inconsistency for estimated images. We further use self-compensation constraints to avoid interference from highly entangled latent factors and achieve more accurate color and structure preservation. Besides, without using any specifically designed cameras or settings, we obtain the low-light and NIRE images using commercial off-the-shelf devices to explore the influence from practical scenarios.

Our whole framework is shown in Figure 2. Since the low-light conditions hide the color information, instead of directly building the self-compensation between the estimated image (Target image in Figure 2) and the low-light image, we build the self-compensation loss based on the color-extracted image with the exposed color information from low-light conditions to preserve the color consistency better. Then, at the artifact purification stage, We further assume that a color-extracted image consists of an artifact component and a color component, while the NIRE image contains a structure component. From Figure 2, by employing the encoder for the color ($E_{\mathcal{C}}$), artifact ($E_{\mathcal{A}}$), and structure ($E_{\mathcal{S}}$), we disentangle these three components from the corresponding images and then achieve the image purification via the fusion of color and structure component. At last, we propose a dataset with the NIRE and low-light images for evaluation and training purposes.

Our major contributions can be concluded as follows:

- A disentanglement framework to purify the low-light images with the NIRE image as the guidance.
- A self-compensation loss with the color extraction module to mutually complement NIRE and visible image for artifact suppression and color consistency.
- A hybrid dataset with images from visible domain and NIRE domain for training and evaluation purposes.

II. Related Work

A. Low-light image enhancement

The enhancement for underexposed images has been studied for more than decades. Guo *et al.* [7] developed a structure-aware smoothing model to estimate the illumination map. Wang *et al.* [61] utilized the deep networks to more effectively estimate the illumination map. Recently, Wei *et al.* [8] combined the classical retinex theory with the deep learning technique to enhance the low-light images effectively. Chen *et al.* [67] proposed to correct the brightness using unpaired learning. A recent method [31] proposed an unsupervised method based on the classical CycleGAN model [30]. Inspired by the traditional non-learning-based methods, some unsupervised low-light image enhancement methods have been proposed recently to solve this problem. For example, the method proposed by [37] enhanced the visibility of the night scenes by leveraging advantages from the bright channel priors. Li *et al.* [1] corrected the illumination by iteratively updating the illumination map. The above methods are effective in correcting the image illumination but less capable of suppressing the artifacts. Recently, some methods are also proposed to suppress the image artifacts during the low-light image enhancement process. For example, Chen *et al.* [10] proposed to operate directly on raw sensor data and replace much of the traditional image processing pipeline. The method proposed in [29] and [21] utilized a restoration module and adjustment module to suppress artifacts and correct the illumination simultaneously. Lore *et al.* [36] proposed a method based on the denoising autoencoder to do the denoising and low-light image enhancement jointly. Recently, Yang *et al.* [52] also proposed a semi-supervised method to handle the artifacts that exist in the low-light images.

B. Near-infrared guided computational photography

The near-infrared information has been widely employed by different computational photography tasks. For example, Krishnan and Fergus [38] proposed to use gradient in both the NIR and Ultra Violet (UV) bands to improve the performances of visible image denoising. Zhuo *et al.* [39] applied the weighted least squares smoothing method to the visible band and transferred details from the NIR band. The method proposed in [34] tried to use the joint bilateral filtering to decompose the visible image into a large-scale image and a detail image. The detail image is then restored and recombined with the large-scale image to get the final result. Shen *et al.* [40] proposed a cross-field method for the image restoration based on both the visible and NIR information. Wang *et al.* [18] introduced a NIR image-guided deep networks for color image denoising. Lyu [35] introduced to extract visible information from the mixed multi-spectrum images to restore image contents. Besides, Li *et al.* [15] introduced an algorithm to fuse the visible and near-infrared images by considering their different reflection and scattering characteristics. By using the image sequence as the input, Wu *et al.* [55] introduced a multi-task deep network with state-synchronization modules to better utilize texture and chrominance information for this problem. Recently, Duan *et al.* [17] utilized the multi-scale edge-preserving decomposition and multiple saliency features for infrared and visible image fusion.

Besides the image restoration, the near-infrared information has also been adopted by other image processing tasks. For example, the method proposed by [41] solves the image intrinsic decomposition problem under the guidance of NIR images. Most above methods consider illumination correction or artifact suppression only. We propose to correct the image illumination and artifact suppression simultaneously under the guidance of NIRE images. The method proposed in [29] and [21] utilized a restoration module and adjustment module to suppress artifacts and correct the illumination simultaneously. Lore *et al.* [36] proposed a method based on the denoising autoencoder to do the denoising and low-light image enhancement jointly.

C. Perceptual quality assessment

How to evaluate the quality of the images after the enhancement is also a pivotal issue for low-light image enhancement or even almost all image restoration tasks. Currently, most methods mainly rely on PSNR and SSIM for evaluation. The development of deep learning also introduces some novel error metrics based on deep learning features (*e.g.*, LPIPS [62]). However, all those error metrics rely on ground truth/reference. In some situations, the lack of reference images makes the evaluation difficult. A new error metric is proposed to address such difficulties for low-light image enhancement by comparing the low-light images with their enhanced counterparts [57], which makes substantial progress in addressing the no-reference evaluation for low-light image enhancement. Moreover, they further introduce an important IQA framework specifically for low-light image enhancement problems, which set a standard for the subsequent IQA frameworks in this area [56]. Besides those pivotal error metrics specifically designed for low-light image enhancement, several methods have been developed in the past several years to provide more robust perceptual quality assessment. The pioneers in perceptual quality assessment also proposed to use free energy for the quality assessment [63]. Besides the error metric for images, some methods also proposed a blind quality evaluator for UGC videos [64], [65]. Recently, more methodshave been introduced to evaluate the quality of audio-visual signal [66]. More detailed surveys can be found in the following papers [58], [59].

III. Dataset collection

Since we use a data-driven approach to solve this problem, an appropriate dataset becomes necessary to learn the purifying process. Previous low-light image enhancement methods proposed to capture an image set with the low-light image $\mathbf{I}$ and its corresponding ground truth $\mathbf{R}$ under low/normal-light conditions (*e.g.*, LOL dataset [8]), respectively. This is a reasonable way to obtain pairs of normal/low-light images by adjusting the light amount received by camera sensors. However, since the ground truth image with normal light is not available for training in our problem, the low/normal-light image pair $(\mathbf{I}, \mathbf{R})$ is not applicable for our purpose.

We instead introduce the low-light image set $(\mathbf{I}, \mathbf{N})$ and the non-paired reference image set $(\mathbf{I}_f, \mathbf{N}_f)$ for training, where $\mathbf{I}$ and $\mathbf{I}_f$ denote low-light images to be enhanced, and the non-paired reference images, respectively, and $\mathbf{N}$ and $\mathbf{N}_f$ denote the corresponding NIRE images. We use a two-step setting to capture images. The low-light image $\mathbf{I}$ is first captured under the low-light conditions similar to the previous dataset [8]. Then, its NIRE image $\mathbf{N}$ is captured by turning on the near-infrared light emitter and switching the shooting mode to night-vision mode. This two-stage process can be easily implemented for the cameras with the internal or external near-infrared emitter. For the unpaired reference image sets $(\mathbf{I}_f, \mathbf{N}_f)$, we first capture $\mathbf{I}_f$ in the normal-light conditions using visible mode and then capture $\mathbf{N}_f$ using the same mode for $\mathbf{N}$.

We capture images using the off-the-shelf surveillance cameras (Wyze cam V2 and Anker Eufy Indoor Cam 2K) in the wild. Then, to investigate the performances in a more controlled scenario, we use a digital camera with the night vision function (Ordro V12) to capture images in different scenarios. We do not find specific details about NIR wavelength of the three cameras, while most IR illuminators employ 850nm for their settings [4]. Based on our experiments, the NIR wavelength of Wyze cam V2, Anker Eufy Indoor Cam 2K, Ordro V2 is less than 900nm. In general, the spectral sensitivity of the three surveillance cameras with silicon senors used in our experiments is from 300nm to 950nm [5].

The images captured by the two devices are with distinct properties. As shown in Figure 4, the images captured by digital cameras are mainly corrupted by the thermal noise, and the structure information is relatively better preserved. However, since surveillance cameras usually compress images before transferring images to the cloud and their aperture size

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3232206

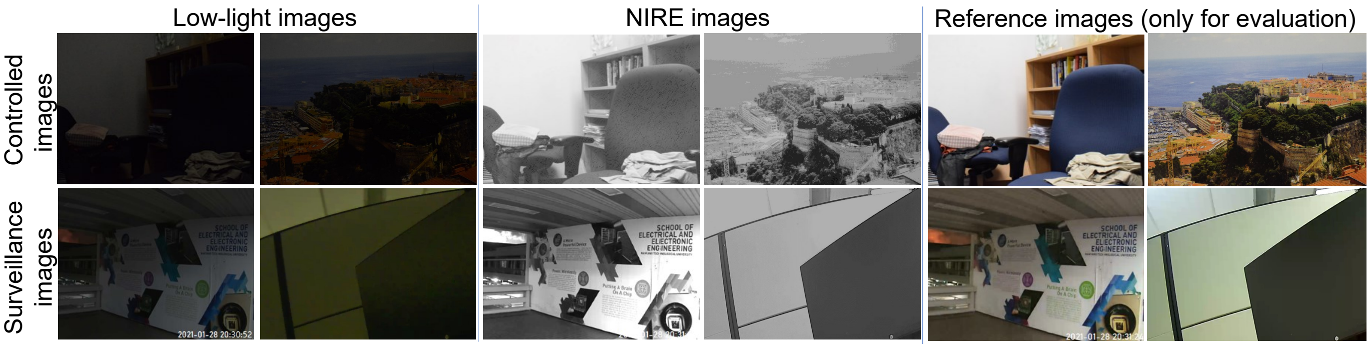

Figure 3: Examples of the low-light images, the NIRE images, and the reference image only used for evaluation.

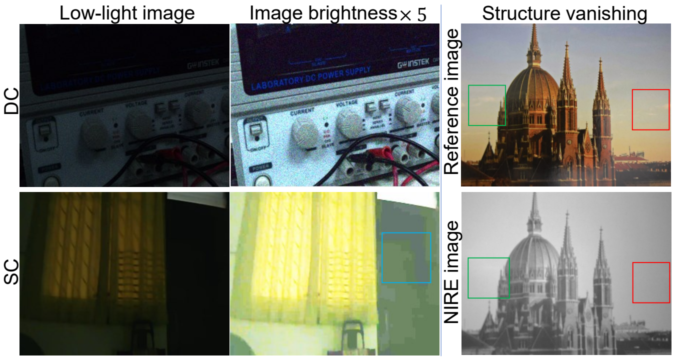

Figure 4: (Left) The low-light images captured by the digital camera (DC) and the surveillance camera (SC). For visualization purposes, we multiply the images in the left column by 5 in the right column. The green box labels the regions with compress artifacts. (Right) Examples of the reference image, and the NIRE image with structure vanishing problem labeled by the green and red boxes.

is limited, compression artifacts corrupt the images captured by surveillance cameras, and the structure information are usually lost during the transmission process. Such differences increase the diversity and pose unique challenges for our dataset. We manually make the near-infrared light distribution uniformly to avoid the bright-spot phenomenon.

A. *Training dataset*

The images captured by the two devices are with distinct properties. As shown in Figure 4, the images captured by digital cameras are mainly corrupted by the thermal noise, and the structure information is relatively better preserved. However, since surveillance cameras usually compress images before transferring images to the cloud and their aperture size is limited, compression artifacts corrupt the images captured by surveillance cameras, and the structure information is usually lost during the transmission process. By flipping, rotating, and cropping these images, our training dataset contains 1200 low-light image sets $(\mathbf{I}, \mathbf{N})$ and 1200 reference image sets $(\mathbf{I}_f, \mathbf{N}_f)$ all from the real world.

B. *Evaluation dataset*

To evaluate the performances of the proposed method, besides the two steps used to capture the training dataset, as shown in the rightmost part of Figure 3, we further increase the light intensity (*e.g.*, by turning on the light) to obtain the corresponding reference image with normal light for the evaluation. Our evaluation dataset contains 100 image triplets with 300 images in total. Among the evaluation dataset, 60 triplets are from the controlled scenes, and 40 triplets are from the surveillance scenes. The surveillance cameras are all controlled remotely as in their real working conditions, which also avoids the potential misalignment between low-light images and NIRE images.

C. *Structure vanishing problem*

Since some real-world materials exhibit slightly different reflective properties under NIR and visible bands [41], the NIRE images may face the structure vanishing problem [6], where some materials become invisible under the NIR band. Since some visible light remains being captured by the camera for our NIRE images, the structure vanishing problem is not widely observed in our dataset. However, as shown in Figure 4, if the NIR light dominates the spectrum received by the camera under highly dark situations, the structure vanishing problem may affect our NIRE images. We also propose a self-compensation loss to address this issue in Section IV-B.

IV. Proposed Method

In this section, we describe the design methodology of the proposed method and the implementation details. As shown in Figure 2, under the guidance of the NIRE image, we first extract the color information by correcting the image illumination in an unsupervised way and then employ a disentanglement framework with the self-compensation loss for image purification.

A. *Color extraction*

Due to to setting of our approach, we directly leverage advantages from previous unsupervised methods [1], [7] to build the color extraction module by correcting the image illumination. A classical way to correct the image illumination without the ground truth is to solve the following formulation [7]:

$$\min_{\mathbf{L}} \|\hat{\mathbf{L}} - \mathbf{L}\|_F^2 + P(\mathbf{L}), \tag{1}$$

where $\mathbf{L}$ denotes the corrected illumination map, $\hat{\mathbf{L}}$ denotes the illumination map initially extracted from the low-light image $\mathbf{I}$, and $P(\mathbf{L})$ denotes the regularization prior on $\mathbf{L}$. In general, Equation 1 can be solved in an iterative manner [42], [43] and the initial estimation of the illumination map can be approximated by the following equation [7]:

$$\hat{\mathbf{L}}(x) = \max_{c \in \{R,G,B\}} \mathbf{I}^c(x). \tag{2}$$

Inspired by the formulation in Equation 1, Equation 2 and the iterative scheme used in [1], we train an extraction module $\mathcal{F}$ to learn the mapping from the input image to its corresponding illumination map. Then, similar to the iterative optimization strategy for Equation 1, by unfolding $\mathcal{F}$ for T times, the illumination map can be iteratively updated as $\mathbf{L}^t = \mathcal{F}(\mathbf{I}^{t-1})$, where $\mathbf{I}^{t-1}$ denotes color-extracted image obtained at $(t-1)$-th iteration and $\mathbf{L}^t$ denotes the illumination map obtained in the t-th iteration.

As shown in Figure 2, the extraction module can be divided into three layers: 1) the feature estimation layer f_{est} to initially extract the illumination related features; 2) an illumination output layer f_{out} to generate the corrected illumination map; 3) an extraction layer f_{extract} to extract the color information based on the corrected illumination map, which can be calculated as follows:

$$\begin{aligned} &\mathbf{z}^t = f_{\text{est}}(\mathbf{I}^{t-1}), \\ &\mathbf{L}^t = f_{\text{out}}(\mathbf{z}^t), \\ &\mathbf{I}^t = f_{\text{extract}}(\mathbf{I}^{t-1}, \mathbf{L}^t). \end{aligned} \tag{3}$$

In Equation 3, $\mathbf{z}^t$ denotes the illumination features extracted at the t-th iteration, and $\mathbf{L}^t$ denotes the illumination map obtained at the t-th iteration. As shown in Figure 2, the parameters of f_{est} and f_{out} are shared across each stage. We employ U-Net [44] with BatchNorm [46] as the backbone for the illumination output layer f_{out}. For f_{extract} in Equation 3, we directly employ the pointwise multiplication used in [47] to get the color-extracted image as

$$f_{\text{extract}}(\mathbf{I}^{t-1}, \mathbf{L}^t) = \mathbf{I}^{t-1} \circ \mathbf{L}^t. \tag{4}$$

Specifically, from Equation 2, since the illumination features are more related to pixels with larger values [24], instead of the ReLU activation layer used by many methods [30], we embed the maxout network into the feature estimation layer f_{est} for non-linear mapping as $F^i(x) = \max_{j \in [1,k]} g^{i,j}(x)$, where $g(x)$ denotes affine feature transformation, $F^i(x)$ denotes features after the maxout mapping, i and j denotes feature positions. $F^i(x)$ generates a new feature map by taking a pixel-wise maximization operation over k affine feature maps. The maxout unit maps each of kN-dimensional vectors into N-dimensianl one by extracting the vectors with maximum values related to the illumination components. In this place, we empirically set $k = 4$.

For the estimated illumination map at each iteration, we employ the total variation loss to preserve the monotonicity relations between neighboring pixel as follows:

$$\mathcal{L}_{\text{tv}} = \sum_{i,j} |\nabla \mathbf{L}_{i+1,j} - \nabla \mathbf{L}_{i,j}| + |\nabla \mathbf{L}_{i,j+1} - \nabla \mathbf{L}_{i,j}|, \tag{5}$$

where i and j denote the pixel positions and ∇ represents the gradient operations.

We further adopt the color constancy loss proposed in [1] to correct the potential color deviations in the estimated image and build the relations among the three corrected channels as follows:

$$\mathcal{L}_{\text{col}} = \sum_{\forall (p,q) \in \varepsilon} \left(\mathbf{I}^p - \mathbf{I}^q\right)^2, \varepsilon = \{(R,G),(R,B),(G,B)\}, \tag{6}$$

where $\mathbf{I}^p$ denotes the average intensity value of p channel in the corrected image, (p,q) represents a pair of channels.

By combining the terms in Equation 5 and Equation 6, the loss functions for the illumination correction stage can be concluded as follows:

$$\mathcal{L}_{\text{extraction}} = \alpha_c \mathcal{L}_{\text{tv}} + \beta_c \mathcal{L}_{\text{col}}, \tag{7}$$

where α_c and β_c are the weighting coefficients to balance the two terms. An example of the color-extracted image is shown in Figure 5, where the color information has been extracted from low-light conditions.

B. Artifact purification

As shown in Figure 5, the color-extracted image generated at the first stage is still with noise and artifacts. Due to the lack of ground truth for training, we utilize the NIRE image as guidance to purify image artifacts. Because of the light compensation from the near-infrared band, the monochrome NIRE image is with fewer artifacts and better preserves the information related to structure and shape [41]. Instead of separating the visible and NIR information mixed in the NIRE image like previous methods [35], we only extract the artifact-free information related to structure and shape. If the structure information can be well extracted from the NIRE image, it may contribute to the artifact purification process by combining it with the color information from the color-extracted image. Based on this assumption, we propose using the disentanglement framework to obtain the artifact-free image by disentangling the color and structure components from the color-extracted image and its NIRE image, respectively. For simplicity, the color-extracted image obtained in Section IV-A is denoted as the artifact-remained image $\mathbf{I}_d$ in this section.

As shown in Figure 2 and Figure 6, the branch for the artifact-remained images $\mathbf{I}_d$ contains a pair of artifact-remained image encoder as $\left\{E_{\mathcal{C}} : \mathbf{I}_d \rightarrow z_d^{\mathcal{C}}, E_{\mathcal{A}} : \mathbf{I}_d \rightarrow z_d^{\mathcal{A}}\right\}$ to encode the artifact-remained image $\mathbf{I}_d$ into color space $\mathcal{C}$ and artifact space $\mathcal{A}$, respectively, and its corresponding structure encoder $\left\{E_{\mathcal{S}} : \mathbf{N}_d \rightarrow n_d^{\mathcal{S}}\right\}$ to disentangle the structure information from the NIRE image. If the disentanglement is well addressed, the encoded color components should contain no information related to the artifact while preserving the color information and the encoded structure space should also only contain the structure information of the image. Then, the decoder $G_{\mathcal{F}}$ can reconstruct a clean image $\mathbf{I}_u$ conditioned only on the color and structure components as $\left\{G_{\mathcal{F}} : z_d^{\mathcal{C}} \times n_d^{\mathcal{S}} \rightarrow \mathbf{I}_u\right\}$.

Correspondingly, as shown in Figure 2 and Figure 6, the branch for the reference image $\mathbf{I}_f$ contains a pair of artifact-free image encoder $\{E_{\mathcal{C}} : \mathbf{I}_f \rightarrow z_f^{\mathcal{C}}\}$, its NIRE image encoder

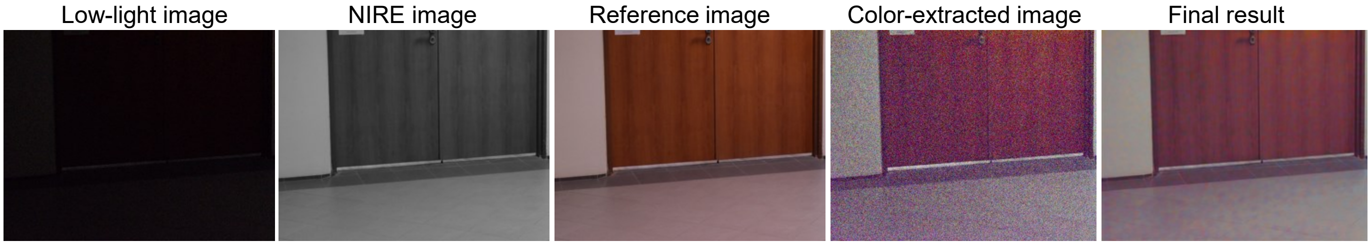

Figure 5: Examples of the low-light image, the NIRE image, the reference image, the color-extracted image from the first stage, and the final result estimated by the second stage.

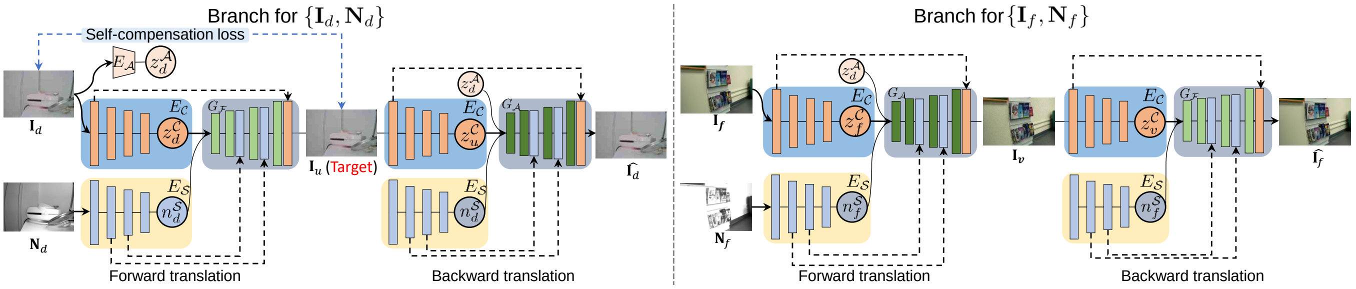

Figure 6: The branch for $\{\mathbf{I}_d, \mathbf{N}_d\}$ and $\{\mathbf{I}_f, \mathbf{N}_f\}$. $\mathbf{I}_u$ is the target clean image. $\hat{\mathbf{I}_d}$ and $\hat{\mathbf{I}_f}$ is the reconstructed $\mathbf{I}_d$ and $\mathbf{I}_f$ for the cycle consistency, respectively.

$\{E_\mathcal{S} : \mathbf{N}_f \rightarrow n_f^\mathcal{S}\}$, and the corresponding decoder $\{G_\mathcal{A} : z_f^\mathcal{C} \times z_d^\mathcal{A} \times n_f^\mathcal{S} \rightarrow \mathbf{I}_v\}$ to generate an artifact-remained image $\mathbf{I}_v$ conditioned on the color, structure, and artifacts components.

1) Forward translation: Given two unpaired image sets from the artifact-remained image and reference image domains as: $\{\mathbf{I}_d \in \mathcal{I}_d, \mathbf{N}_d \in \mathcal{N}_d\}$ and $\{\mathbf{I}_f \in \mathcal{I}_f, \mathbf{N}_f \in \mathcal{N}_f\}$, we encode them into $\{z_d^\mathcal{C}, z_d^\mathcal{A}, n_d^\mathcal{S}\}$ and $\{z_f^\mathcal{C}, n_f^\mathcal{S}\}$, respectively. We then perform the first translation by encoding each representation to generate $\{\mathbf{I}_u, \mathbf{I}_v\}$ as follows:

$$\mathbf{I}_u = G_\mathcal{F}\left(z_d^\mathcal{C}, n_d^\mathcal{S}\right), \quad \mathbf{I}_v = G_\mathcal{A}\left(z_f^\mathcal{C}, n_f^\mathcal{S}, z_d^\mathcal{A}\right), \tag{8}$$

where $\mathbf{I}_u \in \mathcal{I}_f$ denotes the target clean image and $\mathbf{I}_v \in \mathcal{I}_d$ denotes the estimated artifact-remained image.

2) Backward translation: We then encode $\mathbf{I}_u$ and $\mathbf{I}_v$ into $\{z_u^\mathcal{C}\}$ and $\{z_v^\mathcal{C}\}$, and perform the second translation as follows:

$$\hat{\mathbf{I}_f} = G_\mathcal{F}\left(z_v^\mathcal{C}, n_f^\mathcal{S}\right), \quad \hat{\mathbf{I}_d} = G_\mathcal{A}\left(z_u^\mathcal{C}, n_d^\mathcal{S}, z_d^\mathcal{A}\right), \tag{9}$$

where $\hat{\mathbf{I}_f}$ and $\hat{\mathbf{I}_d}$ denote the reconstructed $\mathbf{I}_f$ and $\mathbf{I}_d$, respectively. Specifically, the artifact component $z_d^\mathcal{A}$ is disentangled from the artifact-remained image and shared among the forward and backward translation to ensure that the generated artifact-remained images $\mathbf{I}_v$ and $\hat{\mathbf{I}_d}$ are with consistent artifacts.

After the two translation stages, the cycle-consistency loss for this stage can be represented as follows:

$$\mathcal{L}_{\text{cycle}} = \|\hat{\mathbf{I}_f} - \mathbf{I}_f\|_1 + \|\hat{\mathbf{I}_d} - \mathbf{I}_d\|_1. \tag{10}$$

As shown in Figure 6, the color encoder and the structure encoder share the similar network architecture just with different skip connections to the decoder network, which helps the decoder network better preserve the color and structure information.

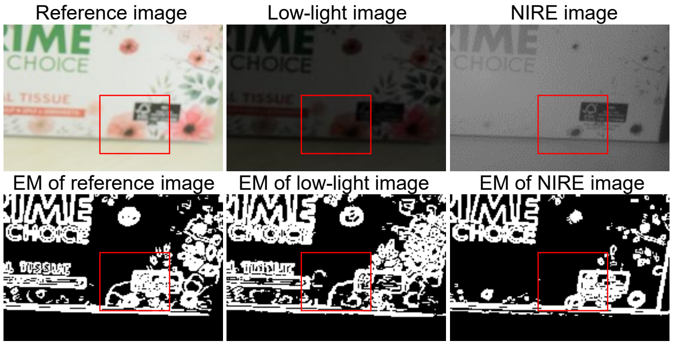

Figure 7: Examples of the reference image, low-light image, NIRE image, and their corresponding Edge Maps (EM). The red boxes label the regions with the structure vanishing problem and all images are from the evaluation dataset.

3) Self-compensation loss: Though the NIRE image provides more reliable guidance for the whole enhancement process, it may cause a negative influence on the final estimated results. The first problem is the color shift problem caused by NIRE images. Since the latent factors related to structure, color, and artifact information are highly entangled in some cases, the residual grayscale information from the NIRE image may interfere with the color information from the artifact-remained image, leading to the color shift problem for final estimated results. Besides, some essential structures may also not be preserved in the NIRE image due to the structure vanishing problem, which degrades the structure consistency of the final result.

We propose a self-compensation loss for color to complement the disentanglement framework by considering the color and structure consistency. Our self-compensation loss for color penalizes the errors between the target image $\mathbf{I}_u$ and the color-

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3232206

extracted image $\mathbf{I}_d$ as follows:

$$\mathcal{L}_{\text{ccl}}(\mathbf{I}_u, \mathbf{I}_d) = \sum_i \rho\left(|\mathbf{I}_u(i) - \mathbf{I}_d(i)|\right). \tag{11}$$

Equation 11 requires the estimated result $\mathbf{I}_u$ not to wildly deviate from the color extracted image $\mathbf{I}_d$ [16] and also stabilizes the consistency in the color space. In Equation 11, $\mathbf{I}_u(i)$ and $\mathbf{I}_d(i)$ denote their pixel values in position i. ρ is the robust function used to eject part of the noise from $\mathbf{I}_d$ [16] and it is defined as

$$\rho(x) = |x|^{\alpha}, 0 < \alpha < 1. \tag{12}$$

We set $\alpha = 0.8$ empirically in our experiments.

We introduce a self-compensation loss for the structure to handle the structure vanishing problem by measuring the salient edge differences between the estimated artifact-free image $\mathbf{I}_u$ and its corresponding low-light image $\mathbf{I}$. As shown in Figure 7, though the low-light image $\mathbf{I}$ is corrupted by artifacts, its salient edges are still consistent with its corresponding artifact-free version. Since the image gradient may be enlarged during the purification process, instead of measuring their pixel-wise difference, we propose to maximize their correlation by minimizing the following loss:

$$\mathcal{L}_{\text{scl}}(\mathbf{I}_u, \mathbf{I}) = -\sum_{n=1}^{N} \left\| \tanh\left(\lambda_u |\nabla \mathbf{I}_u|^{\downarrow n}\right) \odot \tanh\left(\lambda_{\mathbf{I}} |\nabla \mathbf{I}|^{\downarrow n}\right) \right\|_F, \tag{13}$$

where λ_u and $\lambda_{\mathbf{I}}$ are normalization factors, $\|\cdot\|_F$ is the Frobenius norm, $\odot$ denotes element-wise multiplication, and n is the image downsampling factor. Since the gradient information present different properties under different scales [22], we set $n = 3$ to downsample the gradient map for 3 times. Similar to the settings in [19], we set $\lambda_u = \sqrt{\frac{\|\nabla \mathbf{I}\|_F}{\|\nabla \mathbf{I}_u\|_F}}$ and $\lambda_{\mathbf{I}} = \sqrt{\frac{\|\nabla \mathbf{I}_u\|_F}{\|\nabla \mathbf{I}\|_F}}$.

Besides, we also impose the adversarial loss $\mathcal{L}_{\text{adv}}$ [51] for the estimated clean image $\mathbf{I}_u$ and artifact-remained image $\mathbf{I}_v$ to make them similar to the real images. Our discriminator network takes an input image with a size of 224×288 and has 6 strided convolutional layers followed by the ReLU activation function. In the last layer, we use the sigmoid function to generate the final result.

By combining the loss functions in Equation 10, Equation 13, and the adversarial loss, the loss functions for the second stage can be concluded as follows:

$$\mathcal{L}_{\text{purp}} = \alpha_e \mathcal{L}_{\text{cycle}} + \omega_e \mathcal{L}_{\text{ccl}} + \gamma_e \mathcal{L}_{\text{scl}} + \delta_e \mathcal{L}_{\text{adv}}, \tag{14}$$

where α_e, ω_e, γ_e, and δ_e are the weighting coefficients and $\mathcal{L}_{\text{adv}}$ denotes the adversarial losses for $\mathbf{I}_v$ and $\mathbf{I}_u$.

C. Implementation details

We have implemented our method by using PyTorch. The whole training process of our network can be divided into two stages. In the first stage, we first train the illumination correction network for five epochs. In the second stage, we connect the illumination correction network with the suppression network. We then train the whole network to convergence.

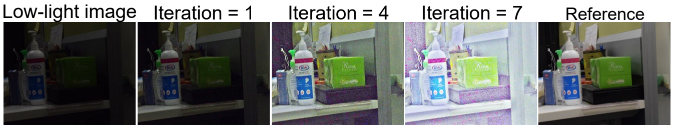

Figure 8: The progressive refinement stage with iteration number T equals to 1, 4 (our setting), and 7, respectively.

The learning rate for the first and second stages is all set to 1×10^{-2}. The weighting coefficient in Equation 7 and Equation 14 are empirically set as: $\alpha_c = 2$, $\beta_c = 1$, $\lambda_c = 0.005$, $\alpha_e = 10$, $\delta_e = 1$, $\omega_e = 1$, and $\gamma_e = 0.5$. At the first stage, the iteration time T is set to 4. From the results shown in Figure 8, since our first stage is a progressive refinement progress, this iteration time T is empirically set to guarantee effective color extraction.

V. Experiments

Due to the lack of ground truth for training, we choose several weakly-supervised methods for low-light image enhancement as the baseline for comparisons: LIME [7], EnlightenGAN [31], and ZeroDCE++ [23]. Besides, we also compare with RetinexNet [8], KinD++ [21], and DRPB [52], three supervised methods. For the supervised method, we employ the training samples with degradations and artifacts from VE-LoL [45] to finetune them, which can improve their capability to handle artifacts in our dataset. We also compare with CycleGAN [30] to investigate the performances. CycleGAN [30], EnlightenGAN [31], and ZeroDCE++ [23] are all trained on our dataset by only using the low-light images as the input. We also compare with ScaleMap [16], a method specified designed for NIR/RGB fusion tasks, to evaluate the effectiveness of our proposed method. Besides, to better evaluate the effectiveness of NIRE images in a relatively simple framework, we train an additional CycleGAN model by considering the NIRE image as another guidance. Since CycleGAN [30] does not have any branches for the NIR feature extraction or embedding, we directly concatenate the NIRE image and the low-light image as 4-channel tensor as the input for CyleGAN [30]. The input channel number of CyleGAN [30] is also changed to 4. The comparison related to this part can be found in Table I and Section V-B.

Besides the low-light image enhancement methods, we also compare with grayscale/infrared image colorization methods, including CIC16 [54], and IDC17 [60]. Based on their settings, we directly use the NIRE images as their input. Some image restoration methods [40], [39], [35], [34] based on the near-infrared information are not involved in the comparisons, since they have different settings or do not release their codes.

In addition to the classical PSNR and SSIM error metrics, we adopt LPIPS [62] as the error metric. The lower LPIPS values indicate better performances. It measures perceptual image similarity using a pre-trained deep network. We also employ a newly proposed error metric NLIEE [57] to evaluate the performance. By directly comparing the enhanced results against its low-light counterparts, this method proposes a reasonable and convenient way to evaluate the performance. Besides, the recently proposed LIEQA [56] also makes a

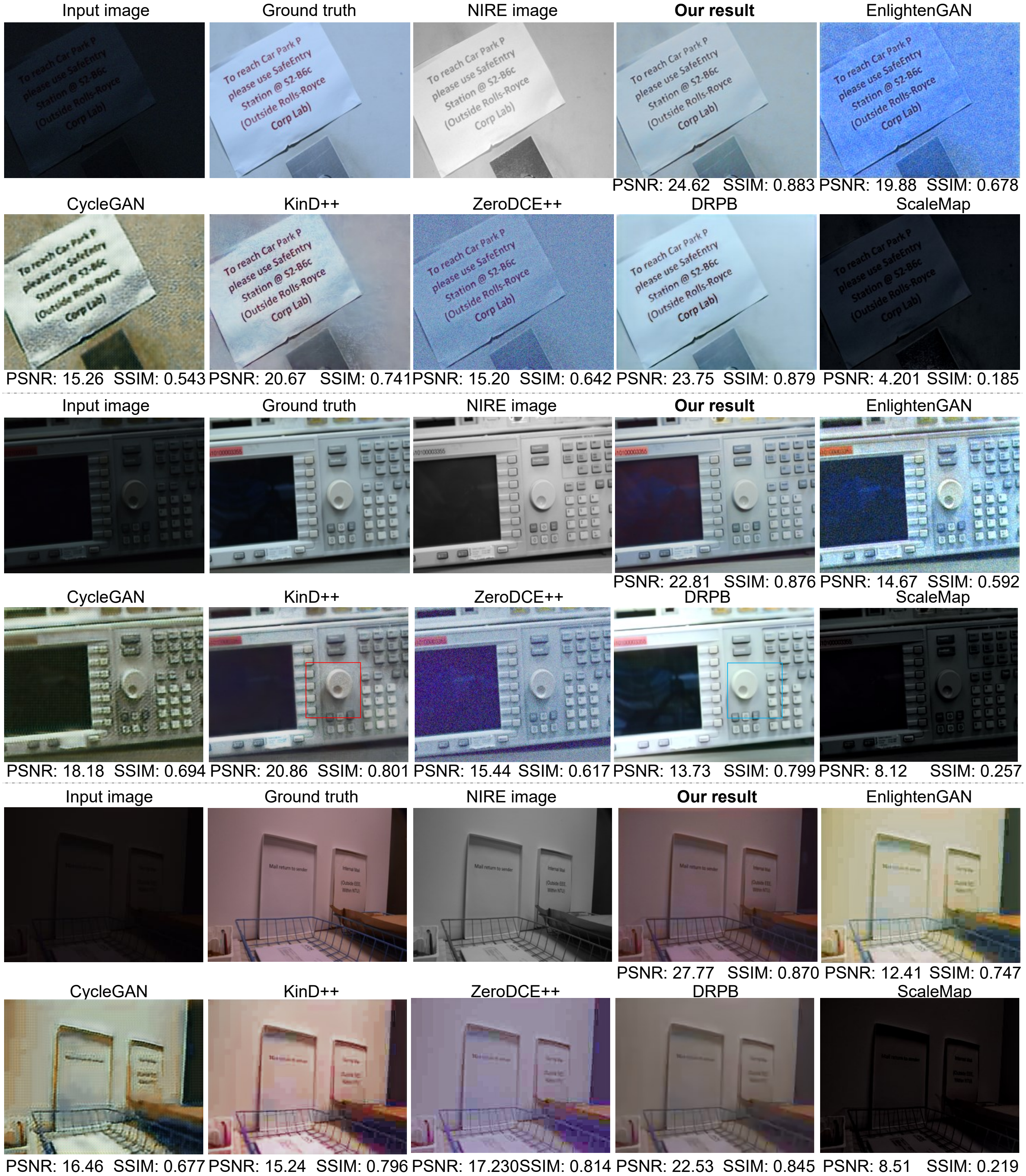

Figure 9: Three examples captured by the digital cameras with result obtained by our method, EnlightenGAN [31], CycleGAN [30], KinD++ [21], ZeroDCE++ [23], DRPB [52], and ScaleMap [16]. The SSIM and PSNR values are shown below each image. The red and blue boxes denote the regions with over-smooth effects and other artifacts.

substantial progress in evaluating the performance of low-light image enhancement, which sets a standard for the follow-up IQA methods. Though LIEQA [56] is not publicly available, the model used in NLIEE [57] is trained on the database used in LIEQA [56]. We, therefore, directly use NLIEE [57] for our evaluation.

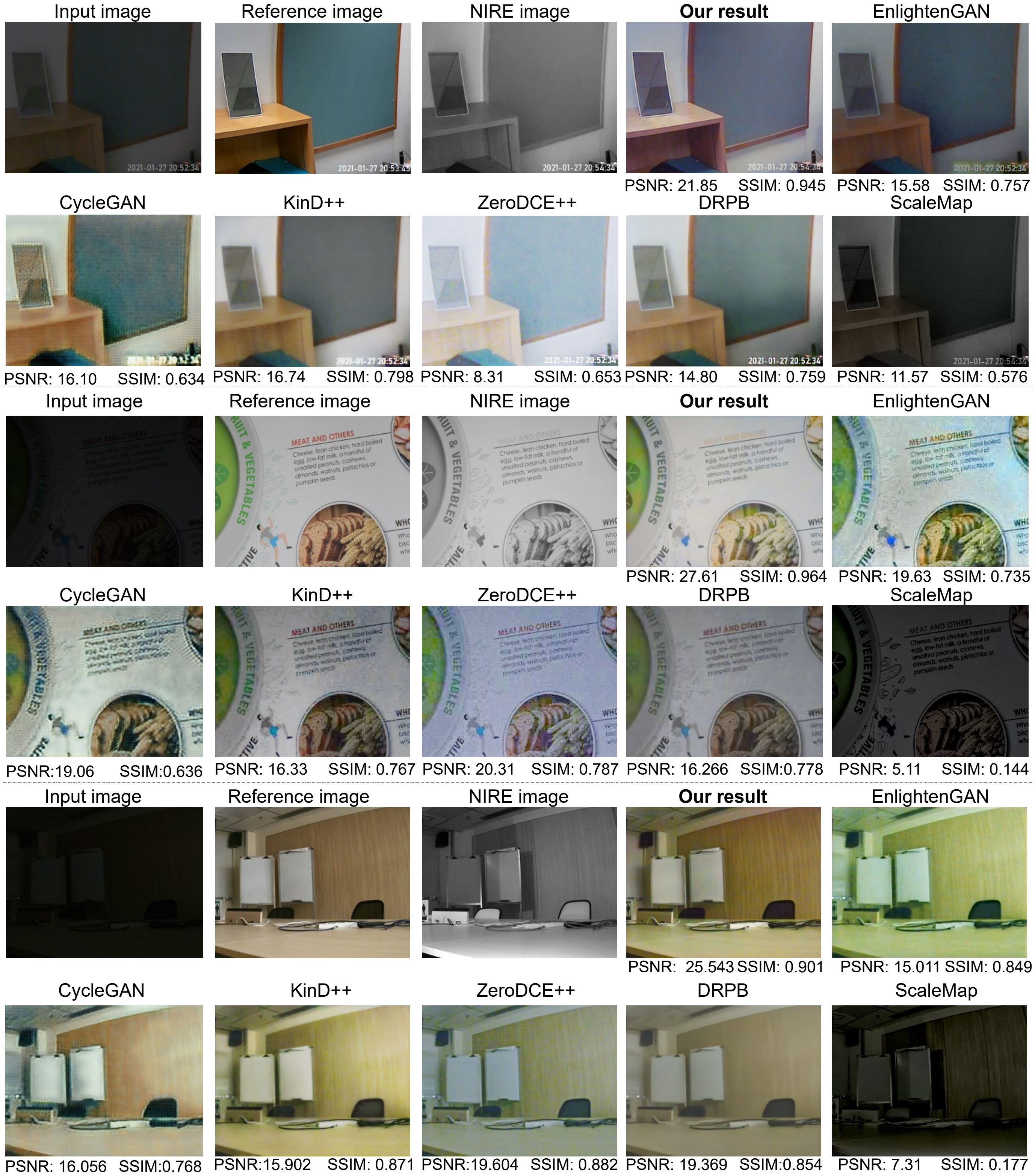

Figure 10: Three examples recorded by the surveillance cameras with results obtained by our method, EnlightenGAN [31], CycleGAN [30], KinD++ [29], ZeroDCE++ [23], DRPB [52], and ScaleMap [16]. The SSIM and PSNR values are shown below each image.

A. *Qualitative evaluations*

The qualitative comparisons are shown in Figure 9 and more examples can be found in our supplementary material. Our proposed method not only enhances the visibility of the low-light images but also better suppresses the artifacts and preserves the image details. Though KinD++ [21] and DRPB [52] also suppress artifacts effectively, it introduces new artifacts to some results or causes over-smooth effects (the

Table I: Quantitative evaluation results using three error metrics, compared with CycleGAN [30], CycleGAN with NIRE, RetinexNet [8], KinD++ [21], ZeroDCE++ [23], EnlightenGAN [31], LIME [7], DRPB [52], ScaleMap [16], CIC16 [54], and IDC17 [60]. The lower LPIPS value indicates better performances (↓). The higher values indicate better performance for other three error metrics (↑).

	SSIM↑	PSNR↑	LPIPS↓	NLIEE↑
Ours	**0.867**	**19.630**	**0.131**	**12.614**
CycleGAN [30]	0.605	13.405	0.370	2.972
CycleGAN with NIRE	0.793	16.731	0.253	3.112
RetinexNet [8]	0.573	13.924	0.357	2.145
KinD++ [21]	0.809	16.331	0.167	4.475
ZeroDCE++ [23]	0.708	15.640	0.261	3.301
EnlightenGAN [31]	0.784	15.270	0.318	1.116
LIME [7]	0.641	16.116	0.291	1.654
DRPB [52]	0.842	18.224	0.233	5.913
ScaleMap [16]	0.272	6.347	0.258	3.168
CIC16 [54]	0.832	16.774	0.191	6.354
IDC17 [60]	0.822	15.773	0.172	6.185

regions labeled by the red box in Figure 9), which also leads to the lower quantitative values. The other low-light image enahancement methods (*e.g.*, LIME [7], EnlightenGAN [31], and ZeroDCE++ [23]) can correct the illumination, but less capable of the artifacts suppression, which influences the visual quality of the final results.

We further show the results on surveillance images in Figure 10. Since the surveillance camera cannot accurately record the color information under low-light conditions, it becomes difficult to faithfully recover the color information. Even the results of our method still show color bias (the first example in Figure 10). Besides, in contrast to the images captured by the digital camera, the surveillance images are mainly corrupted by the compression artifacts and blurring effects. Our method still shows better ability in preserving the structure and color consistency. The results estimated by previous methods (*e.g.*, CycleGAN [30] and DRPB [52]) are still with noticeable compression artifacts. Though ScaleMap [16] is proposed for RGB/NIR fusion, it cannot extract the color information like our methods. From the results shown in Figure 9 and Figure 10, the results obtained by ScaleMap [16] are still with dark appearance.

Since the grayscale/infrared image colorization methods (CIC16 [54] and IDC17 [60]) can well preserve the structure consistency by utilizing the NIRE image, they all achieve acceptable SSIM values for the examples in Figure 9 and Figure 10. However, without the color representations disentangled from the color-extracted image, their estimated results show obvious color bias and lower PSNR values.

B. Quantitative evaluation

The quantitative results in Table I reconfirm the observations in Figure 9. Our method achieves the best scores among all other methods. The higher SSIM values indicate that our method recovers images with better quality. The smaller LPIPS values indicate that our proposed method indeed generates images with a better perceptual similarity. KinD++ [21] achieves the second-best results. However, due to the new artifacts in the final estimated image, KinD++ [21] still cannot outperform our proposed method. Since the other low-light image enhancement methods cannot suppress the artifacts effectively, their error metric values cannot outperform our methods and KinD++ [21].

Table II: Quantitative evaluations for the model without the color extraction module (CEM), the model without the self-compensation loss for color (CCL) and the self-compensation loss for structure (SCL).

	SSIM↑	PSNR↑	LPIPS↓	NLIEE↑
Complete model	**0.867**	**19.630**	**0.131**	**12.614**
W/o CEM	0.739	13.936	0.223	5.664
W/o CCL	0.829	15.282	0.215	7.064
W/o SCL	0.843	18.891	0.146	7.510

The SSIM values of CIC16 [54] and IDC17 [60] are better than other low-light image enhancement methods. However, as discussed before, this is mainly because SSIM only calculates the structural similarity (SSIM) index for grayscale images. It cannot fairly reflect the color bias problem of the final result. The lower PSNR and LPIPS values indicate that the image colorization methods cannot accurately estimate the final results.

The results obtained by CycleGAN with NIRE also show the NIRE image's effectiveness, which improves the performance of original CycleGAN. From results shown in Table II, NIRE images help CycleGAN [30] achieve better results. Besides, from the examples shown in Figure 13, the result of CycleGAN with NIRE image (the third column of Figure 13) can better suppress the artifacts that its counterpart shown in the second column of Figure 13.

At last, the better NLIEE [56], [57] error metrics values also confirm that our method achieves better results than other methods. By directly comparing the enhanced results against its low-light counterparts, it provides another reasonable and convenient way to evaluate the performance.

C. Ablation study

1) Two-stage vs. One-stage: By directly utilizing the low-light image and NIRE image as the input, the one-stage framework only with artifact purification module is also a solution to this problem. However, since our method relies on self-compensation loss to compensate for the color and structure information, the second stage alone cannot effectively correct the image illumination. As shown in Figure 11, the results without the first stage appear darker than the result obtained by the full model. The quantitative values in Figure 11 also prove the effectiveness of the two-stage framework.

If we further remove the self-compensation loss, the illumination information can be better recovered. However, the color consistency cannot be accurately preserved without self-compensation constraints. More details about the effectiveness of the self-compensation loss can be found in Section V-C3.

2) Effectiveness of NIRE: We then remove the structure encoder for the NIRE image. From Figure 11, without the NIRE encoder, the artifacts cannot be well suppressed in the final results. The quantitative values in Table II also become

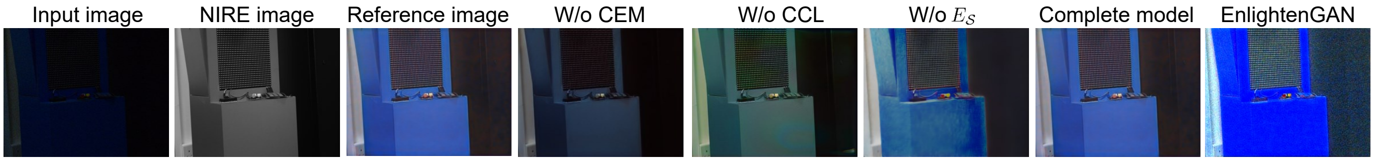

Figure 11: Examples from our model without the color extraction module (CEM), our model without the color compensation loss (CCL), our model without the structure encoder ($E_{\mathcal{S}}$) for the NIRE image, our complete model, and EnlightenGAN [31].

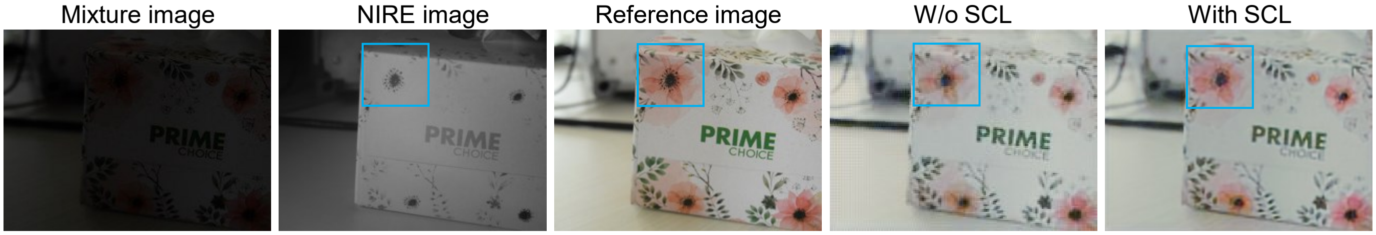

Figure 12: Examples from our model without the structure compensation loss (SCL) and the complete model with SCL. The blue box denotes the regions with the structure vanishing problem in NIRE image, reference images, and results obtained without SCL and with SCL, respectively.

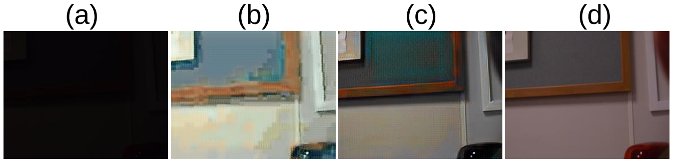

Figure 13: From left to right: (a) low-light image, (b) result of CycleGAN [30] without NIRE image, (c) result of CycleGAN with NIRE image, and (d) the reference image.

Table III: The quantitative comparisons for results obtain under different iteration number. We set $T = 1, 3, 4, 7$, respectively.

	SSIM↑	PSNR↑	LPIPS↓	NLIEE↑
$T = 1$	0.739	13.936	0.223	5.664
$T = 3$	0.857	17.530	0.139	10.504
$T = 4$	**0.867**	**19.630**	**0.131**	**12.614**
$T = 7$	0.847	16.567	0.155	9.650

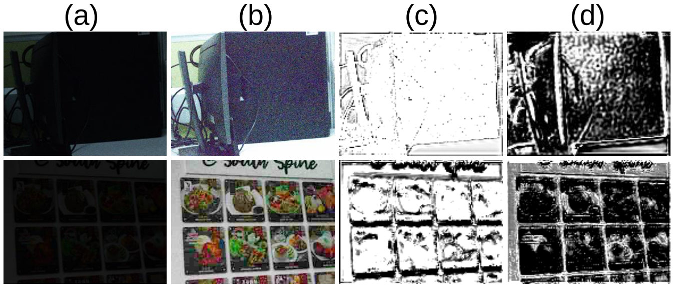

Figure 14: Examples of (a) low-light images, (b) their corresponding color-extracted image, and (c) the gradient response of the structure encoder $E_{\mathcal{C}}$ and (d) the color encoder $E_{\mathcal{S}}$.

similar to the values obtained by EnlightenGAN [31]. However, this experiments show that our method still has the ability to improve the illumination of low-light images even without the guidance of NIRE images. When the guidance from NIRE images are removed, our method can be regarded as regular low-light image enhancement methods. If incorporating NIRE images, our method is assumed to have better results.

3) Effectiveness of self-compensation loss: We then remove the self-compensation loss to evaluate its effectiveness. We first remove the structure compensation loss in Equation 13 and then the color compensation loss in Equation 11. The examples in Figure 12 show that some regions without structure compensation loss become invisible in the final estimated results since these details do not exist in the NIRE images. However, since such texture vanishing problem only occupies a small amount in our near-infrared images, the improvement brought by the loss functions is also not very significant.

From the results shown in Figure 12, the color compensation loss plays a vital role during the purifying process. Without the color compensation loss, the color information cannot be well embedded into the final estimated results and also degrades the performances of the final results.

D. Analysis to the structure and color encoders

We provide further analysis for the structure ($E_{\mathcal{S}}$) and color ($E_{\mathcal{C}}$) encoders. From the gradient response of the two encoders in Figure 14, they indeed play different roles during the enhancement process. The gradient response of $E_{\mathcal{S}}$ mainly focuses on the edge regions with more meaningful structure information, while the gradient response of $E_{\mathcal{C}}$ focuses on the global regions.

E. Analysis to the iteration numbers

Our method relies on a color extraction module with progressive refinement to extract color information from low-light images. The iteration number for such progressive refinement is currently set to 4. We further perform quantitative evaluations to validate the effectiveness of such settings. As evidenced by the corresponding results displayed in Table III, when T is set to 4, our method can achieve the highest error metric values about other settings, which supports its rationale.

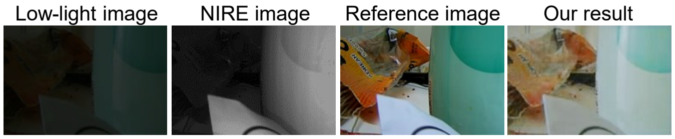

Figure 15: An example with the nonuniform near-infrared light intensity.

VI. Conclusion

We propose to purify the low-light images via the disentanglement of NIRE image. Taking the low-light image as the input, the color extraction module first extracts the color information. Then, the artifact purifying module disentangles the color and structure information from the color-extracted and NIRE images to suppress the artifacts. The results show that our method achieves promising results.

A. Limitations

In spite of the promising results, our method still has several limitations that need to be addressed. First, due to the energy attenuation of the near-infrared light, some areas covered by the near-infrared light may cast nonuniform light intensity, which also influences the results of our proposed method (*e.g.*, the example shown in Figure 15). A more powerful near-infrared light emitter with more uniform illumination distribution can partially address this problem. Then, since the color gamut of the low-light images acquired under extremely dark situations may be distorted, the extracted color information may not be accurate enough for the next stage, which finally deteriorate the color consistency between our result and the reference image. We will address these issues in our future work by employing more sophisticated experimental devices and considering a more effective restoration model. Furthermore, due to the difficulty of capturing the data in the proposed capturing setup, our dataset does not represent the entire truth of the real world. We will explore different techniques in the future to expand our dataset, such as updating the capturing setup or synthesizing images with better consideration of physical properties.

References

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

[2] Georg Petschnigg, Richard Szeliski, Maneesh Agrawala, Michael Cohen, Hugues Hoppe, and Kentaro Toyama, "Digital photography with flash and no-flash image pairs," *ACM Transactions on graphics (TOG)*, 2004.

[3] Shen, Xiaoyong and Yan, Qiong and Xu, Li and Ma, Lizhuang and Jia, Jiaya, "Multispectral joint image restoration via optimizing a scale map," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[4] AXTON, "https://axtontech.com/infrared-850nm-vs-940nm-wavelength/," *AXTON*, 2019.

[5] Feifan Lv, Yinqiang Zheng, Bohan Zhang, Feng Lu, "Turn a Silicon Camera Into an InGaAs Camera," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognitionn*, 2019.

[6] Jian Wang, Tianfan Xue, Jonathan T Barron, and Jiawen Chen, "Stereoscopic dark flash for low-light photography," in *Proceedings of International Conference on Computational Photography*, 2019.

[7] Xiaojie Guo, Yu Li, and Haibin Ling, "Lime: Low-light image enhancement via illumination map estimation," *IEEE Transactions on image processing (TIP)*, 2016.

[8] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu, "Deep retinex decomposition for low-light enhancement," *British Machine Vision Association*, 2018.

[9] Zhang, Yonghua and Guo, Xiaojie and Ma, Jiayi and Liu, Wei and Zhang, Jiawan, "Beyond Brightening Low-light Images," *International Journal of Computer Vision*, 2021.

[10] Chen, Chen and Chen, Qifeng and Xu, Jia and Koltun, Vladlen, "Learning to see in the dark," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[11] Ma, Jiayi and Chen, Chen and Li, Chang and Huang, Jun, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Information Fusion*, 2016.

[12] Chen, Yu-Sheng and Wang, Yu-Ching and Kao, Man-Hsin and Chuang, Yung-Yu, "Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[13] Gharbi, Michaël and Chen, Jiawen and Barron, Jonathan T and Hasinoff, Samuel W and Durand, Frédo, "Deep bilateral learning for real-time image enhancement," *ACM Transactions on Graphics (TOG)*, 2017.

[14] Hu, Yuanming and He, Hao and Xu, Chenxi and Wang, Baoyuan and Lin, Stephen, "Exposure: A white-box photo post-processing framework," *ACM Transactions on Graphics (TOG)*, 2018.

[15] Li, Zhuo and Hu, Hai-Miao and Zhang, Wei and Pu, Shiliang and Li, Bo, "Spectrum characteristics preserved visible and near-infrared image fusion algorithm," *IEEE Transactions on Multimedia*, 2020.

[16] Shen, Xiaoyong and Yan, Qiong and Xu, Li and Ma, Lizhuang and Jia, Jiaya, "Multispectral joint image restoration via optimizing a scale map," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[17] Duan, Chaowei and Wang, Zhisheng and Xing, Changda and Lu, Shanshan, "Infrared and visible image fusion using multi-scale edge-preserving decomposition and multiple saliency features," *Optik*, 2021.

[18] Wang, Xuehui and Dai, Feng and Ma, Yike and Guo, Junbo and Zhao, Qiang and Zhang, Yongdong, "Near-infrared image guided neural networks for color image denoising," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

[19] Zhang, Xuaner and Ng, Ren and Chen, Qifeng, "Single image reflection separation with perceptual losses," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[20] Yang, Xin and Xu, Ke and Song, Yibing and Zhang, Qiang and Wei, Xiaopeng and Lau, Rynson WH, "Image correction via deep reciprocating HDR transformation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[21] Zhang, Yonghua and Guo, Xiaojie and Ma, Jiayi and Liu, Wei and Zhang, Jiawan, "Beyond Brightening Low-light Images," *International Journal of Computer Vision*, 2021.

[22] Wan, Renjie and Shi, Boxin and Hwee, Tan Ah and Kot, Alex C, "Depth of field guided reflection removal," *IEEE International Conference on Image Processing (ICIP)*, 2016.

[23] Li, Chongyi and Guo, Chunle and Loy, Chen Change, "Learning to Enhance Low-Light Image via Zero-Reference Deep Curve Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[24] Edwin H Land, "The retinex theory of color vision," *Scientific american*, 1977.

[25] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia, "Underexposed photo enhancement using deep illumination estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[26] Leonardo Galteri, Lorenzo Seidenari, Marco Bertini, and Alberto Del Bimbo, "Deep generative adversarial compression artifact removal," in *Proceedings of International Conference on Computer Vision*, 2017.

[27] Pei Li, Loreto Prieto, Domingo Mery, and Patrick Flynn, "Face recognition in low quality images: a survey," *arXiv preprint arXiv:1805.11519*, 2018.

[28] Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau, "Learning to restore low-light images via decomposition-and-enhancement," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[29] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proceedings of ACM Multimedia*, 2019.

[30] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of International Conference on Computer Vision*, 2017.

[31] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "Enlightengan: Deep light enhancement without paired supervision," *IEEE Transactions on Image Processing*, 2021.

[32] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin, "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2018.

[33] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu, "Learning disentangled representations for recommendation," *Advances in Neural Information Processing Systems*, 2019.

[34] Sosuke Matsui, Takahiro Okabe, Mihoko Shimano, and Yoichi Sato, "Image enhancement of low-light scenes with near-infrared flash images," *Information and Media Technologies*, 2011.

[35] Feifan Lv, Yinqiang Zheng, Yicheng Li, and Feng Lu, "An integrated enhancement solution for 24-hour colorful imaging," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[36] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar, "Llnet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition*, 2017.

[37] Hunsang Lee, Kwanghoon Sohn, and Dongbo Min, "Unsupervised low-light image enhancement using bright channel prior," *IEEE Signal Processing Letters*, 2020.

[38] D. Krishnan and R. Fergus, "Dark flash photography," *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 2009.

[39] Shaojie Zhuo, Xiaopeng Zhang, Xiaoping Miao, and Terence Sim, "Enhancing low light images using near infrared flash images," in *Proceedings of International Conference on Image Processing*, 2010.

[40] Qiong Yan, Xiaoyong Shen, Li Xu, Shaojie Zhuo, Xiaopeng Zhang, Liang Shen, and Jiaya Jia, "Cross-field joint image restoration via scale map," in *Proceedings of International Conference on Computer Vision*, 2013.

[41] Ziang Cheng, Yinqiang Zheng, Shaodi You, and Imari Sato, "Non-local intrinsic decomposition with near-infrared priors," in *Proceedings of International Conference on Computer Vision*, 2019.

[42] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, Wen Gao, and Alex C Kot, "Region-aware reflection removal with unified content and gradient priors," *IEEE Transactions on Image Processing*, 2018.

[43] Linbin Yu, Miao Zhang, and Chris Ding, "An efficient algorithm for l 1-norm principal component analysis," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 2012.

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proceedings of International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, 2015.

[45] Liu, Jiaying and Xu, Dejia and Yang, Wenhan and Fan, Minhao and Huang, Haofeng, 'Benchmarking low-light image enhancement and beyond," in *International Journal of Computer Vision*, Springer, 2021.

[46] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of International Conference om Machine Learning*, 2015.

[47] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding, "A weighted variational model for simultaneous reflectance and illumination estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[48] Martin Arjovsky, Soumith Chintala, and Léon Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[49] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky, "Deep image prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9446–9454.

[50] Zhilin Zheng and Li Sun, "Disentangling latent space for vae by label relevant/irrelevant dimensions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[51] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of International Conference on Computer Vision*, 2017.

[52] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu, "From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3063–3072.

[53] Amanda Berg, Jorgen Ahlberg, and Michael Felsberg, "Generating visible spectrum images from thermal infrared," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018.

[54] Richard Zhang, Phillip Isola, and Alexei A Efros, "Colorful image colorization," in *Proceedings of European Conference on Computer Vision*. Springer, 2016.

[55] Wu, Guangming and Zheng, Yinqiang and Guo, Zhiling and Cai, Zekun and Shi, Xiaodan and Ding, Xin and Huang, Yifei and Guo, Yimin and Shibasaki, Ryosuke, "Learn to Recover Visible Color for Video Surveillance in a Day," in *Proceedings of European Conference on Computer Vision*. Springer, 2020.

[56] Zhai, Guangtao and Sun, Wei and Min, Xiongkuo and Zhou, Jiantao, "Perceptual quality assessment of low-light image enhancement," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2021.

[57] Zhang, Zicheng and Sun, Wei and Min, Xiongkuo and Zhu, Wenhan and Wang, Tao and Lu, Wei and Zhai, Guangtao, "A no-reference evaluation metric for low-light image enhancement," *Proceedings of IEEE International Conference on Multimedia and Expo*, 2021.

[58] Zhai, Guangtao, and Xiongkuo Min, "Perceptual image quality assessment: a survey," *Science China Information Sciences*, 2020.

[59] Min, Xiongkuo, Ke Gu, Guangtao Zhai, Xiaokang Yang, Wenjun Zhang, Patrick Le Callet, and Chang Wen Chen, "Screen content quality assessment: overview, benchmark, and beyond," *ACM Computing Surveys (CSUR)*, 2021.

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

[61] Wang, Ruixing and Zhang, Qing and Fu, Chi-Wing and Shen, Xiaoyong and Zheng, Wei-Shi and Jia, Jiaya, "Underexposed photo enhancement using deep illumination estimation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[62] 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*, 2018.

[63] Ke Gu, Guangtao Zhai, Xiaokang Yang, Wenjun Zhang, "Using Free Energy Principle For Blind Image Quality Assessment," in *IEEE Transactions on Multimedia*, 2014.

[64] Bowen Li, Weixia Zhang, Meng Tian, Jiu Jiang, Guangtao Zhai, Xianpei Wang, "Learning a Blind Quality Evaluator for UGC Videos in Perceptually Relevant Domains," in *Proceedings of IEEE International Conference on Multimedia and Expo*, 2022.

[65] Bowen Li, Weixia Zhang, Meng Tian, Guangtao Zhai, Xianpei Wang, "Blindly Assess Quality of In-the-Wild Videos via Quality-aware Pre-training and Motion Perception," in *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[66] Min, Xiongkuo, Guangtao Zhai, Jiantao Zhou, Mylene CQ Farias, and Alan Conrad Bovik, "Study of subjective and objective quality assessment of audio-visual signals," in *IEEE Transactions on Image Processing*, 2020.

[67] Chen, Yu-Sheng and Wang, Yu-Ching and Kao, Man-Hsin and Chuang, Yung-Yu, "Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.