NIR-assisted Video Enhancement via Unpaired 24-hour Data

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

Abstract

Low-light video enhancement in the visible (VIS) range is important yet technically challenging, and it is likely to become more tractable by introducing near-infrared (NIR) information for assistance, which in turn arouses a new challenge on how to obtain appropriate multispectral data for model training. In this paper, we defend the feasibility and superiority of NIR-assisted low-light video enhancement results by using unpaired 24-hour data for the first time, which significantly eases data collection and improves generalization performance on in-the-wild data. By accounting for different physical characteristics between unpaired daytime and nighttime videos, we first propose to turn daytime NIR & VIS into "nighttime mode". Specifically, we design a heuristic yet physics-inspired relighting algorithm to produce realistic pseudo nighttime NIR, and use a resampling strategy followed by a noiseGAN for nighttime VIS conversion. We further devise a temporal-aware network for video enhancement that extracts and fuses bi-directional temporal streams and is trained using real daytime videos and pseudo nighttime videos. We capture multi-spectral data using a co-axial camera and contribute Fulltime Multi-Spectral Video Dataset (FMSVD), the first dataset including aligned 24-hour NIR & VIS videos. Compared to alternative methods, we achieve significantly improved video quality as well as generalization ability on in-the-wild data in terms of both evaluation metrics and visual judgment. Codes and Data Available: https://github.com/MyNiuuu/NVEU.

1. Introduction

Visually pleasing videos under well-illuminated conditions are essential for human perception as well as high-level computer vision tasks. In practice, however, many videos are captured under sub-optimal conditions due to environmental constraints, leading to poor visibility, structural degradation, and unpredictable noise interference.

So far, a large number of algorithms have been proposed to enhance images/videos in the visible (VIS) range. Supervised methods [41, 59, 62, 5, 61, 4, 70] provide remarkable performance by denoising and enhancing the low-light inputs. However, in-the-wild image pairs for supervised training are laborsome to obtain, which basically needs different camera settings for low-light/normal image pairs of completely static objects. Capturing *video* pairs for moving objects is more involved, which requires complex systems, such as the coaxial imaging system with ND filters [24] and repeatable mechatronic motion tracks [58], making the process harder and less practical. Existing unsupervised methods [18, 25, 44] need no image pairs for training, but their capability in tackling noise tends to be undermined.

In addition to the severely degraded VIS image, it is

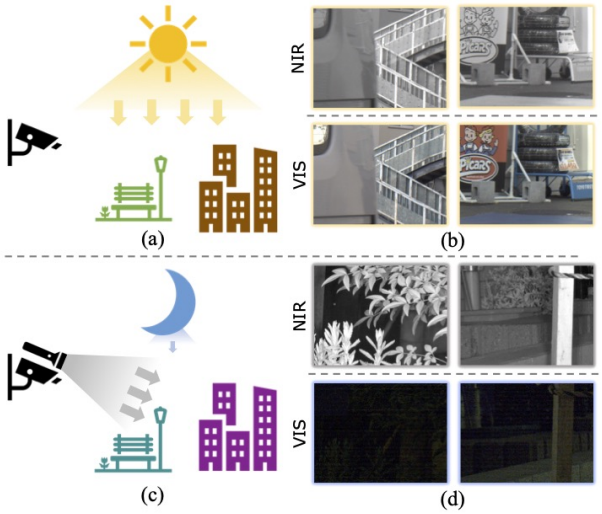


Figure 1: **Light source and distribution differences between daytime and nighttime.** During daytime, both NIR and VIS come from sunlight and skylight, which share similar spatial distribution (a), and the VIS and NIR images are bright and more uniform (b). During nighttime, the camera is equipped with NIR LEDs in a co-located setting (c), thus the intensity distribution of nighttime NIR images depends heavily on object distance and surface direction. The VIS images are much darker due to weak illuminants like moonlight and manmade lamps (d).

*Corresponding author

sometimes convenient to obtain another bright NIR image, by enabling auxiliary illuminants. Thus, VIS-NIR fusion [51, 26, 12] has become a promising method for low-light imaging. Compared to pure VIS-based methods, rich and detailed information is introduced from corresponding NIR images. Several supervised learning based methods [26, 67] have been proposed to enhance VIS images by fusing photographs of extra wavelengths.

Unquestionably, adding assistance images of additional wavelengths is likely to robustify the enhancement task and provide superior performance, but it also makes data collection more challenging. Given the practices of data collection in the VIS range, it is obvious that capturing realistic paired data in VIS-NIR domain is extremely difficult, if not infeasible. Existing methods including DVN [26] and DRF [67] simply synthesize training pairs from clean images or use static images to ease data collection, which yet inevitably leads to generality issues on real data.

In this paper, we propose to consider a paradigm that only requires unpaired real data, which is much easier to collect since we do not have to assure the same scenes for low-light and normal images/videos. As a result, we can easily capture large-scale real data (even in *video* form) for training and testing. The complex lighting and noise distributions covered in the dataset also intuitively assure better generality on in-the-wild data. For camera settings, we consider the most practical monocular systems which can be achieved through either co-axial systems or filters.

Apparently, by using a synchronized co-axial camera, it is possible to take aligned VIS and NIR videos in daytime and nighttime, respectively. Nevertheless, the differences in light source and brightness level lead to obvious domain gaps between daytime and nighttime VIS/NIR images (Fig. 1). Specifically, daytime VIS and NIR images are bright, and they share the same illumination distribution implied by the sunlight and skylight. In contrast, nighttime VIS frames suffer from poor visibility, structural degradation, and unpredictable noise interference due to low photon counts. Although nighttime NIR frames are free from those issues because of auxiliary NIR illuminants, they still have wide domain gaps from daytime NIR in terms of light distribution, considering that the auxiliary illuminants (like LEDs) are usually equipped around the camera lens in a nearly co-located setting, as in most security cameras.

Based on these 24-hour data, we propose the first NIR-assisted low-light video enhancement paradigm using unpaired videos, which can be divided into two stages (Fig. 2):

1) *Day-to-night video synthesis.* Given the different characteristics between daytime and nighttime videos, we first propose to turn daytime videos into "nighttime mode". For NIR day-to-night synthesis, we design a novel relighting (RL) algorithm. The algorithm takes a daytime NIR n_{day} together with an inferred depth map as input, and outputs its nighttime version ($\hat{n}_{night}$) by approximating the effects of co-located illumination. During this process, it considers the difference in light distribution between daytime and nighttime by formulating two factors: physical distance and surface angle to the camera. For VIS day-to-night synthesis, each daytime VIS frame v_{day} is first processed by a resampling trick (RST). We then simply leverage existing noise GAN techniques [6, 66, 23] to add realistic pseudo noise on each frame. The noise GAN is first trained on real nighttime VIS to learn the camera noise pattern.

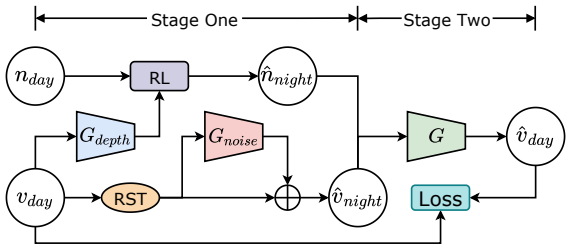


Figure 2: **The two-stage framework of our method.** RL and RST represent relighting algorithm and resampling trick, respectively.

2) *Improved VIS-NIR fusion with pseudo data pairs.* We design and optimize an enhancement network G using real daytime VIS and pseudo nighttime NIR & VIS. The network extracts and fuses the feature of three continuous frames and outputs the enhancement result for the middle frame.

We demonstrate the effectiveness of our model on in-the-wild video datasets collected by our camera system, showing significantly improved video quality compared to existing alternative methods. We also show the superior generalization ability through testing on another third-party dataset. Our contribution can be summarized as follows:

- For the first time we show that through physics-aware modifications on 24-hour unpaired data, an NIR-assisted low-light enhancement model can be trained with superior performance and generalization ability.
- A heuristic yet effective relighting method for realistic NIR day-to-night synthesis by modeling the distance and surface angle to the camera.
- A temporal-aware video enhancement network, which is trained on real daytime VIS and pseudo nighttime NIR & VIS synthesized by our method.
- Fulltime Multi-Spectral Video Dataset (FMSVD), the first dataset including in-the-wild aligned NIR and VIS videos during both daytime and nighttime.

2. Related Work

Low-light Enhancement. Traditional enhancement methods were mostly based on histogram manipulation [8, 22, 33, 53] or Retinex theory [32, 27, 60, 15, 19, 36]. In recent years, many learning-based methods have been proposed

Figure 3: **Samples from our dataset.** From top to down: nighttime NIR, nighttime VIS, daytime NIR, daytime VIS.

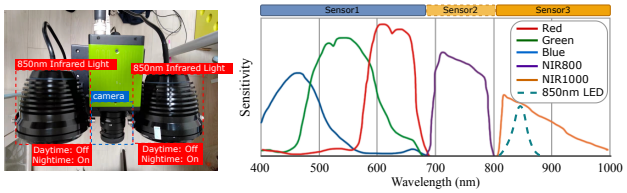


Figure 4: Left: camera system to collect FMSVD. Right: spectrum response of camera and curve of 850 nm LED.

Table 1: Configuration of FMSVD.

	Daytime	Nighttime	Total
Scenes	29	35	64
Frames	18381	22982	41363
Frame/Sce.	633.83	656.63	646.30
Format	240×320, PNG		
Settings	FPS: 24, Exposure: 2ms		
Camera	JAI FS-3200T-10GE-NNC		

and attracted increasingly wide interest [41, 59, 62, 18, 25]. Supervised settings have also been extensively explored for enhancing RAW images [5, 61] and videos [4, 24, 70, 58]. These methods often rely on paired datasets from numerical simulation [63, 72] or approximate capture [24, 58], which inevitably leads to generalization issues on challenging real data. Unsupervised methods [25, 18, 44] were further proposed to simplify data collection, but usually fail to resolve obvious noise caused by low-light conditions.

NIR and RGB Image Fusion. Traditional image fusion methods were often based on spatial transformation techniques such as wavelet transform [34], contourlet transform [9], and edge-preserving filter-based transform [43]. In recent years, deep learning have attracted great attention in this field [35, 68, 69, 56]. Xiong *et al*. [67] proposed a new flash technique for low-light imaging which uses deep-red light for assistance. Jin *et al*. [26] proposed to fuse NIR images into low-light RGB images and synthesized data for the supervised enhancement of single images.

NIR-to-RGB Translation. NIR-to-RGB Translation [38, 54, 57] aims to colorize NIR images into RGB images. Limmer *et al*. [38] first trained a deep multi-scale convolutional neural network that performs direct and integrated transfer between NIR and RGB pixels. To deal with unpaired data, Nyberg *et al*. [48] and Mehri *et al*. [45] learned the mapping with an unsupervised Generative Adversarial Network (GAN) [16] based on CycleGAN [75]. Wu *et al*. [64] proposed a supervised method for NIR2RGB video translation, yet the training data is captured in daytime, and the gap in illumination distribution between nighttime artificial LEDs and daytime illuminants still exists.

VIS-NIR Datasets. Various camera systems have been designed to capture VIS-NIR image pairs for further analysis and applications [14, 17, 55, 13, 31]. There are mainly three types of hardware: 1) cameras with IR-cut filters that switch between VIS and NIR [64, 42]. 2) single-chip sensor that is sensitive to NIR and VIS in different parts of the filter array [47]. 3) coaxial cameras that capture multi-spectral photographs in one shot, which is utilized to build our dataset. Sadeghipoor *et al*. [50] contributed a dataset including 50 VIS-NIR image pairs. Brown *et al*. [3] further proposed MSIFT containing 477 VIS-NIR images. Lv *et al*. [42] built a dataset with 714 aligned VIS-NIR images. These datasets are relatively small and limited to static scenes. VSIAD [64] is a large-scale dataset that contains NIR & VIS videos, but precludes low-light videos during nighttime. DVD [26] is the first VIS-NIR single-image dataset for static low-light scenes, which is not available yet.

3. FMSVD

We introduce Fulltime Multi-Spectral Video Dataset (FMSVD), the first dataset that includes aligned in-the-wild NIR and VIS videos during both daytime and nighttime.

3.1. Hardware Configuration

The camera system used to collect data is illustrated in Fig. 4. We choose a multi-sensor camera JAI FS-3200T-10GE-NNC together with two 850 nm infrared lights. As shown in Fig. 4, the camera is equipped with 3 CMOS image sensor: Sensor1 with response in 400 nm $\sim$ 700 nm range, Sensor2 in 700 nm $\sim$ 800 nm range, and Sensor3 in 800 nm $\sim$ 1000 nm range. We only adopt Sensor1 to cap-

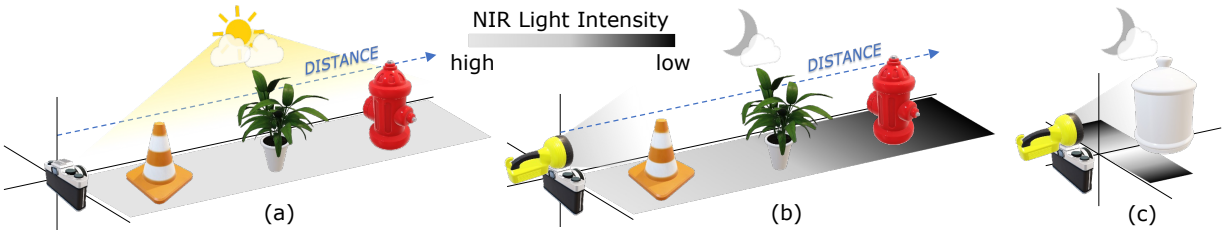


Figure 5: **Illustration for daytime and nighttime NIR light distribution.** (a) In daytime, the sunlight and skylight are infinite, and scene objects almost share the same intensity regardless of distance towards the camera. (b) In nighttime, because of the near-field characteristics of NIR LEDs, the light intensity decreases as the distance from the camera grows. (c) In nighttime, the local area perpendicular to the camera reflects more photons to the sensor and is therefore brighter, due to the co-located setting of camera and light source.

ture VIS frames and Senor3 to capture NIR frames. During nighttime, we use two $850\,\mathrm{nm}$ infrared LED lights for illumination, since this wavelength is widely used in industry.

3.2. Collected Data

The configuration of our dataset is listed in Tab. 1. Each frame includes a VIS image and its corresponding NIR image. The NIR and VIS frames are temporally and spatially aligned since the camera is a coaxial multi-sensor camera. There are a total of 64 scenes (41363 frames) in our dataset. Note that there is no paired video between daytime and nighttime because scenes are dynamic. The FPS of each video is set to 24 with an exposure time of 2ms. All images are stored in PNG format after the ISP process through Rawpy, with a resolution of 240 $\times$ 320. Fig. 3 presents some data samples from our collected dataset. More visual samples are in supplementary materials.

4. Method

4.1. Day-to-night Video Synthesis

In stage one, we turn daytime NIR & VIS into "nighttime mode". For NIR, we consider the difference in light distribution via the proposed relighting algorithm. For VIS, we consider the difference in brightness and noise interference.

4.1.1 Day-to-night NIR Synthesis.

Motivation. During daytime, the light of NIR images comes from sunlight and skylight, which provides enough photons for high-quality videography (Fig. 5(a)). During nighttime, though the camera is equipped with co-located NIR LEDs to assure enough illumination, the light distribution is apparently different from the infinite illuminants of daytime. First, the light intensity decreases as the distance between the object and the camera grows due to the near-field characteristics of NIR LEDs (Fig. 5(b)). Second, given the same distance, the local surface perpendicular to the camera reflects more photons to the sensor and is therefore brighter, because of the co-located setting of camera and light source (Fig. 5(c)). Typical daytime and nighttime NIR samples can be found in Fig. 3. Following these principles, we design the *relighting algorithm* that simulates pseudo nighttime NIR images from real daytime NIR by redistributing the light intensity value of each pixel.

Theoretically, a rigorous physics-based relighting process can be modeled as follows: 1) Obtaining reflectance component through Intrinsic Image Decomposition [37, 7, 39], which is a prerequisite for the following two steps. 2) Adjusting the light intensity of each pixel in reflectance according to the inverse-square law of near-field point source. 3) Calculating shading under the nighttime illuminant according to a reflectance model [1, 30]. However, existing intrinsic decomposition methods mainly focus on indoor images, and outdoor scenes are extremely hard to perform even with hyper-spectral data [71].

Without estimating the exact reflectance, modeling the light intensity decay is rather a practical process. Therefore, instead of strictly following physical rules, we empirically design a heuristic algorithm to model distance and surface angle effect, as will be introduced in the following parts. We also formulate and analyze the rigorous physics-based relighting process in the supplementary material .

Relighting algorithm. To start with, we predict a depth map using a monocular depth estimation network G_{depth}. We use MonoViT [73] as our depth estimation network because of its state-of-the-art performance and generalization ability. The network takes a daytime VIS frame v_{day} as input, and outputs a depth map $\mathcal{D} \in \mathbb{R}^{H \times W}$ for further usage:

$$\mathcal{D} = G_{depth}\left(v_{day}\right), \tag{1}$$

where bigger value in each position of $\mathcal{D}$ represents closer distance to the camera. We then normalize $\mathcal{D}$ to $(0, 1)$ and modulate the corresponding daytime NIR frame n_{day} as:

$$\hat{\mathcal{D}}(i,j) = \frac{\mathcal{D}(i,j) - MIN}{MAX - MIN}, \tag{2}$$

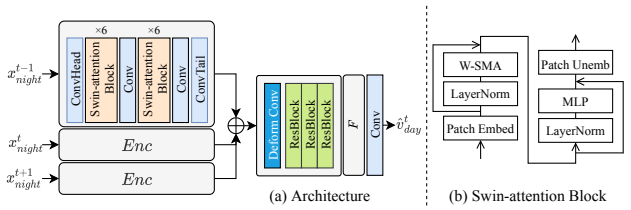
$$n_{day}^{dis} = n_{day} \odot \hat{\mathcal{D}}, \tag{3}$$

where $i \in \{1, 2, ..., H\}$, $j \in \{1, 2, ..., W\}$, and $\odot$ represents the Hadamard product. We set MAX and MIN to

Algorithm 1 Calculating Perpendicular Scale Map

Input: Depth Map $\mathcal{D}$ of size $H \times W$, patch size k.
Output: Perpendicular Scale Map $\mathcal{P}$

$\mathcal{P} \leftarrow$ zero matrix of size $H \times W$
for $i \leftarrow 1$ to H **do**
  for $j \leftarrow 1$ to W **do**
    $\mathcal{A} \leftarrow k \times k$ subarea centered at (i, j).
    $d_h \leftarrow \frac{\partial}{\partial x}\mathcal{A}$, $d_v \leftarrow \frac{\partial}{\partial y}\mathcal{A}$.
    $\mathcal{P}(i, j) \leftarrow \frac{d_h + d_v}{k \times k}$.
  end for
end for
$\mathcal{P} \leftarrow 1 - \frac{\mathcal{P} - \min(\mathcal{P})}{\max(\mathcal{P}) - \min(\mathcal{P})}$.
return Perpendicular Scale Map $\mathcal{P}$



(a) Architecture (b) Swin-attention Block

Figure 6: (a) The architecture of our enhancement model. (b) Structure of Swin-attention Block.

the maximum and minimum value of all depth maps in each scene instead of each frame to preserve temporal consistency. Based on the original NIR frame, the value of each pixel in n_{day}^{dis} fades as the distance from the camera grows. This process roughly reflects the near-field characteristic of nighttime auxiliary illuminant, yet in a linear decay model, rather than quadratic in the ordinary sense.

To further consider the surface angle effect, we calculate the perpendicular scale map $\mathcal{P}$ according to $\hat{\mathcal{D}}$, and introduce it as a modulated residual:

$$\hat{n}_{night} = n_{day}^{dis} + \lambda \cdot n_{day}^{dis} \odot \mathcal{P}, \tag{4}$$

where λ is the hyperparameter for tuning the weight of $\mathcal{P}$. The pseudo code for calculating $\mathcal{P}$ is shown in Alg.1. The idea is to walk through every pixel in $\hat{\mathcal{D}}$ and calculate the average differential (both horizontal and vertical) of the $k \times k$-sized subarea centered on it. A smaller average differential indicates that the subarea is more closed to be perpendicular to the camera and vice versa.

4.1.2 Day-to-night VIS Synthesis

In contrast to the gap of illumination distribution between daytime and nighttime NIR, the core gap between daytime and nighttime VIS lies in the light intensity and noise level. We first model the brightness difference through the resampling trick, then leverage a noise GAN to match the noise distribution of real nighttime VIS, following existing unsupervised denoising fashions [6, 66, 23].

Resampling trick. The idea of resampling trick derives from the data normalization used in many computer vision tasks. One of the most famous examples is to normalize each image from ImageNet [11] with the three-channel mean of $[0.485, 0.456, 0.406]$, and standard variation of $[0.229, 0.224, 0.225]$. After normalization, the data distribution approximately becomes $\mathcal{N}(0, 1)$. Similar to that, we first calculate the three-channel mean and standard variation of daytime VIS and nighttime VIS respectively, then normalize the daytime VIS according to its own mean and standard variation:

$$\check{v}_{day}^{c} = \frac{v_{day}^{c} - \mu_{day}^{c}}{\sigma_{day}^{c}}, \tag{5}$$

where $c \in \{R, G, B\}$ stands for color channels. μ_{day}^{c} and σ_{day}^{c} represent the mean and standard variation of channel c in daytime VIS, respectively. After that, we re-sample (denormalize) each image according to the mean and standard variation of *nighttime* VIS to obtain the pseudo output:

$$\tilde{v}_{night}^{c} = \check{v}_{day}^{c} \cdot \sigma_{night}^{c} + \mu_{night}^{c}. \tag{6}$$

Noise GAN. Inspired by recent unsupervised image denoising algorithms [6, 66, 21, 23], we leverage a noise GAN to mimic real noise patterns of nighttime VIS. The Generator G_{noise} takes $\tilde{v}_{night}$ as input and predicts a noise residual $\hat{s}$. $\hat{s}$ is then directly added to $\tilde{v}_{night}$ and forms the final output:

$$\hat{s} = G_{noise}(\tilde{v}_{night}), \tag{7}$$

$$\hat{v}_{night} = \tilde{v}_{night} + \hat{s}. \tag{8}$$

The architecture and training procedures of G_{noise} are simply based on [23], and details can be found in supplementary materials.

4.2. Video Enhancement Model

Now that we can generate realistic pseudo nighttime videos, our next step is to train a network that performs video enhancement given nighttime NIR and VIS.

Network Architecture. Fig. 6 presents the architecture of our enhancement network G, which operates on three continuous frames to consider temporal consistency. Specifically, G takes x_{night}^{t-1}, x_{night}^{t}, and x_{night}^{t+1} as input, where x_{night}^{T} is the concatenation of $\hat{n}_{night}^{T}$ and $\hat{v}_{night}^{T}$:

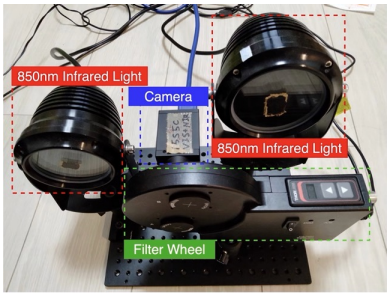
$$x_{night}^{T} = \text{concat}[\hat{n}_{night}^{T}, \hat{v}_{night}^{T}]. \tag{9}$$

The input of each timestamp is first encoded by the feature encoder Enc, which consists of several convolution layers and Swin-attention Blocks [40]. Since deep features from adjacent frames may not be spatially consistent with the present frame, it is beneficial to use deformable convolution which is able to dynamically adjust the receptive

Table 2: **Quantitative results against alternative methods.** ✓ (✗) for 'Data Pair' means the method needs (does not need) data pairs. ✓ (✗) for 'NIR' means the method takes (doesn't take) NIR as input. The best and second best results are in **red** and **blue**.

Data Pair	NIR	Methods	FMSVD			Third-Party		
			PI↓	NIQE↓	HSE↑	PI↓	NIQE↓	HSE↑
✗	✗	EnG	9.440	17.005	**3.16**	8.486	14.098	**3.50**
		Z-DCE	9.332	16.740	3.00	8.185	13.621	3.16
		SCI	9.888	17.756	2.67	8.666	13.778	3.00
	✓	Ours	**4.786**	**5.469**	**4.50**	**2.819**	**3.666**	**4.33**
✓	✗	Jiang *et al.*	5.022	**5.935**	1.83	4.595	5.319	2.00
		URetinex	8.200	14.484	2.83	7.488	11.993	2.33
		UTVNet	7.039	12.205	2.83	6.806	11.429	2.50
	✓	DRF	**4.989**	6.123	2.16	**3.437**	**3.995**	2.50
		DVN	8.282	9.998	1.83	6.959	7.080	1.00

Figure 7: **Camera to capture third-party dataset,** with two 850 nm NIR LEDs and a filter wheel to switch between VIS (400 nm to 700 nm) and NIR (> 800 nm).



field to handle various geometric transformations and spatial misalignment. Thus, we devise two successive fusing layers F to fuse encoded features of different timestamps via a Deformable Convolutional Block [10, 76] followed by three Residual Blocks [20]. The more detailed structure of G can be found in supplementary materials.

Loss Function. The perceptual loss [28] has been widely used in image reconstruction tasks due to its ability to recover details and preclude over-smooth results compared to pixel-wise losses. Based on it, we design a simple saturation loss to produce more perceptually satisfying results as well as powerful modeling capabilities for wider scenarios:

$$\mathcal{L}_{sat} = \sum_{i=1}^{I} \left\| \psi_i \left(\hat{v}_{day}^t \right) - \psi_i \left(f_{sat}(v_{day}^t) \right) \right\|_1, \qquad (10)$$

where $\hat{v}_{day}^t$ is the output of G, and v_{day}^t is the corresponding ground truth from daytime VIS. ψ_i denotes the activation map at the i-th layer of the pre-trained VGG-19 network [52]. Particularly, we chose 5 layers including $relu_{1-1}$, $relu_{2-1}$, $relu_{3-1}$, $relu_{4-1}$, and $relu_{5-1}$ from the VGG-19 network. f_{sat} is a non-linear function that modifies the saturation level of v_{day}. Specifically, we first turn v_{day} into HLS color space (h_{day}), transforming the saturation channel h_{day}^s according to the preset non-linear curve:

$$f_{sat}(h_{day}^s) = (-(1 - \frac{h_{day}^s}{255})^{\phi} + 1) \times 255, \qquad (11)$$

then turn it back to RGB color space.

4.3. Implementation Details.

We implement the training part of our model with Pytorch [49]. We use Adam optimizer [29] with $\beta_1 = 0.5, \beta_2 = 0.999$, and randomly crop the input images to 64×64. To train the video enhancement network G, we set batch size to 4 and learning rate to 2e-4. We use Adam optimizer with $\beta_1 = 0.5, \beta_2 = 0.999$ to optimize G, and ϕ for the saturation loss is set to 2. λ is set to 0.1, and k is set to 13 for the relighting algorithm. The total training iterations is 10,000. We randomly split 10 scenes from nighttime dataset for testing. The training procedure is performed on 4 NVIDIA GeForce RTX 3090.

5. Experiments

5.1. Settings

Methods. We evaluate and compare our method with 8 state-of-the-art methods for low-light image/video enhancement. EnGAN [25], Zero-DCE [18], and SCI [44] are unsupervised enhancement methods that take single VIS image as input. Jiang *et al.* [24], URetinex [65], and UTVNet [74] are supervised methods that take multiple or single VIS image as input. DRF [67] and DVN [26] are supervised methods that fuse additional spectral images with VIS.

Benchmarks. Since there is no publicly available multi-spectral benchmark containing low-light and NIR input, we thus conduct experiments mainly on our FMSVD. To further test the generalization ability of all the methods, we build a third-party dataset, which is collected at night with a different camera from FMSVD (FLIR GS3-U3-15S5C with the IR-cut filter removed). The camera system is illustrated in Fig. 7. Different from FMSVD, the VIS and NIR frames are obtained by switching on a VIS-range filter (400 nm ∼ 700 nm) and a NIR-range filter (larger than 800 nm), respectively. The dataset includes 41 static low-light scenes, each of which contains 10 continuous aligned VIS-NIR frames. All frames are stored in PNG format, with a resolution of 512 × 688. For all compared methods, we use the official codes and checkpoints, if they are available. Note that none of the models is re-trained or finetuned, unless explicitly indicated. Also, our model is trained on part of FMSVD, without any finetuning on the third party data.

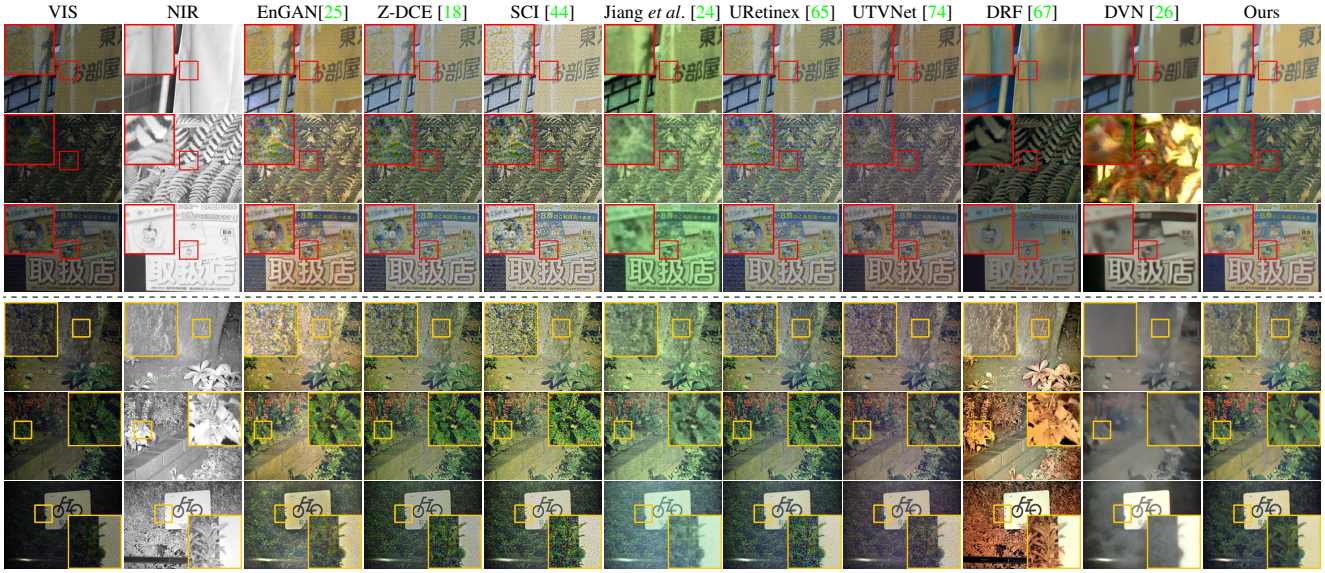


Figure 8: **Qualitative results** on FMSVD (above dash) and the third-party dataset (below dash). Zoom in for a clear view.

5.2. Main Results

Quantitative results. Comparing the visual quality of images/videos without reference is quite subjective. Here, standard metrics PI (Perceptual Index) [2] and NIQE [46] are adopted in quantitative experiments. Considering that these metrics can only evaluate the quality of *single images* from certain aspects, we also conduct a Human Subjective Evaluation (HSE) which directly reflects human's perceptual judgment on *video* results to compare the performance of our method and other methods. Specifically, we randomly select 6 videos from the test set. For each video, it is first enhanced by these methods respectively. We then ask 10 volunteers to independently watch the output video of these methods and assign an integer score ranging from 1 (bad quality) to 5 (excellent quality) respectively. During this process, the volunteers are instructed to consider: 1) whether there exists visible noise; 2) whether the video contains over- or under-exposure effects; 3) whether the video shows color/structural distortions; and 4) the temporal consistency of the video. The quantitative results on FMSVD and third-party dataset are reported in Tab. 2. We can see that our method achieves the best results on two datasets in terms of both quantitative and human subjective evaluation.

Qualitative results. Visual results on FMSVD and the third-party dataset are shown in Fig. 8. More results in video form can be found in supplementary materials. We can observe that: 1) methods that only take VIS as input (EnGAN, Z-DCE, SCI, Jiang *et al*., URetinex, UTVNet) fail to recover the structure loss caused by lack of photons and heavy noise. For multi-spectral fusion methods, DRF heavily counts on NIR to recover structure information, which promises good overall structure and noise suppression, but generates results with obvious color distortion and content loss. 2) EnGAN, Z-DCE, SCI, and URetinex fail to eliminate (sometimes even amplify) the visible noise. Although Jiang *et al*. and UTVNet consider noise interference during the algorithm design, the denoising effects of these methods are still far from being satisfactory. DVN successfully suppresses the visible noise, but suffers from over-fitting issues and gives perceptually inferior results. 3) Color bias exists in previous methods such as EnGAN (orange), Jiang *et al*. (green), and DRF (color distortion). In contrast, our method successfully recovers the structure loss in RGB frames via retrieving the complementary information from NIR. Our method also suppresses visible noise and produces results with no obvious color distortion.

Visual results for pseudo nighttime NIR. Fig. 9 shows visual results for pseudo NIR images generated by our proposed relighting algorithm. Although our method is heuristic, it reflects the core observations in physics, and the results look visually pleasing. More results and a comparison with quadratic decay and Lambert based shading can be found in the supplementary materials.

5.3. Ablation Study

We perform several ablation studies to demonstrate the effectiveness of each component of our model. Different variants are tested on both FMSVD and the third-party dataset, including model without relighting algorithm (w/o RL), without the noise GAN (w/o NG), without resampling trick (w/o RST), without saturation loss (w/o f_{sat}), and replacing relighting algorithm with CycleGAN [75] to produce pseudo nighttime NIR (*Cycle). The results are reported in Tab. 3 and Fig. 10. From the results, we can observe that without the relighting algorithm, though our

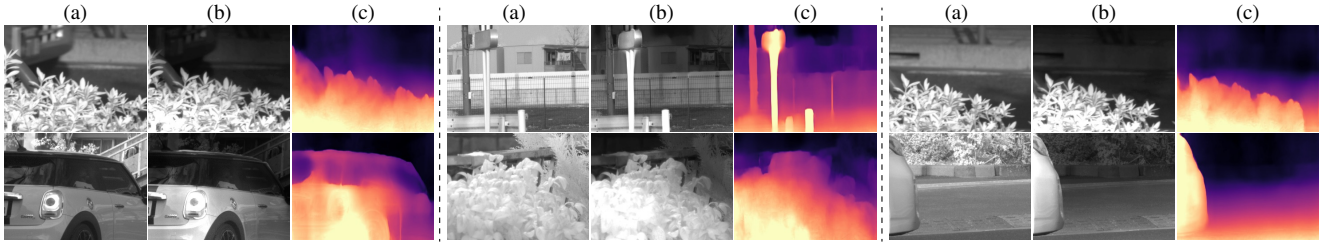


Figure 9: **Visual results for pseudo NIR.** (a) original daytime NIR. (b) pseudo nighttime NIR. (c) depth map from G_{depth}.

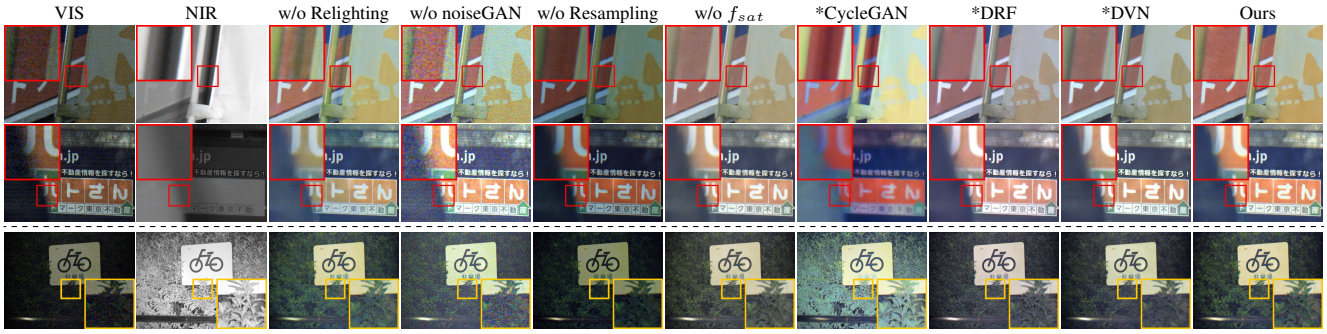


Figure 10: **Visual results for ablation studies** on FMSVD (above dash) and third-party dataset (below dash).

Table 3: **Quantitative ablation results** on FMSVD and the third-party dataset. The best results are in **bold**.

Methods	FMSVD			Third-Party		
	PI↓	NIQE↓	HSE↑	PI↓	NIQE↓	HSE↑
w/o RL	5.052	5.783	4.16	3.039	3.796	3.83
w/o NG	5.367	8.837	3.16	4.511	6.809	3.33
w/o RST	5.679	6.250	2.83	3.226	4.088	3.00
w/o f_{sat}	5.111	5.864	4.33	2.823	**3.571**	4.16
*Cycle	5.146	6.147	2.66	3.197	3.772	2.83
*DRF	5.195	6.536	3.83	2.944	4.234	3.67
*DVN	5.214	6.186	3.67	3.128	3.990	3.16
Ours	**4.786**	**5.469**	**4.50**	**2.819**	3.666	**4.33**

model can give relatively acceptable results when NIR has uniform distributions like daytime (third row in Fig. 10), it starts to produce obvious artifacts and color distortion when nighttime NIRs have large gaps from daytime NIR in terms of light distribution (first and second row in Fig. 10). Without noise GAN, our model fails to suppress the obvious noise from input, leading to unsatisfying results. Without resampling trick, our model cannot enhance the low-light frames to adequate lighting conditions. Our model also outputs plain results without the saturation loss. If we replace relighting algorithm with CycleGAN [75], the model produce results with color distortion and structure discrepancy.

Retraining multi-spectral methods with our data. We conduct experiments in which we use paired data generated by our stage one pipeline to retrain multi-spectral fusion methods including DRF [67] (*DRF) and DVN [26] (*DVN). The results are shown in Tab. 3 and Fig. 10. On the one hand, DRF [67] and DVN [26] produce obviously better visual results on FMSVD and have better generality on the third-party dataset after retraining on paired data generated by our algorithm. On the other hand, the results still suffer from color bias and artifacts, demonstrating the superiority of our enhancement network.

6. Conclusion

We proposed the first NIR-assisted low-light video enhancement paradigm which makes use of in-the-wild unpaired 24-hour VIS-NIR videos from our proposed Full-time Multi-Spectral Video Dataset (FMSVD). To address the light distribution gaps between daytime and nighttime, we performed NIR day-to-night synthesis through a heuristic yet effective relighting algorithm, and VIS day-to-night conversion via the resampling trick and a noise GAN. A video enhancement model was then optimized using pseudo data and real data. We evaluated our model on both FMSVD and the third-party dataset, demonstrating superior video quality as well as generalization ability through both evaluation metrics and human subjective judgment.

For future works, we will explore: 1) Advanced algorithm (hardware) for more accurate depth estimation to synthesize NIR (e.g., multi-view stereo, Time-of-Flight camera). 2) Adopting multiple NIR images with different wavelengths for more robust assistance.

Acknowledgement

This research was supported in part by JSPS KAKENHI Grant Numbers 22H00529, 20H05951, JST-Mirai Program JPMJMI23G1, and ROIS NII Open Collaborative Research 2023-23S1201.

References

[1] Ronen Basri and David W Jacobs. Lambertian reflectance and linear subspaces. *IEEE transactions on pattern analysis and machine intelligence*, 25(2):218–233, 2003. 4

[2] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018. 7

[3] Matthew Brown and Sabine Süsstrunk. Multi-spectral sift for scene category recognition. In *CVPR 2011*, pages 177–184. IEEE, 2011. 3

[4] Chen Chen, Qifeng Chen, Minh N Do, and Vladlen Koltun. Seeing motion in the dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3185–3194, 2019. 1, 3

[5] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3291–3300, 2018. 1, 3

[6] Jingwen Chen, Jiawei Chen, Hongyang Chao, and Ming Yang. Image blind denoising with generative adversarial network based noise modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3155–3164, 2018. 2, 5

[7] Ziang Cheng, Yinqiang Zheng, Shaodi You, and Imari Sato. Non-local intrinsic decomposition with near-infrared priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2521–2530, 2019. 4

[8] Dinu Coltuc, Philippe Bolon, and J-M Chassery. Exact histogram specification. *IEEE Transactions on Image processing*, 15(5):1143–1152, 2006. 2

[9] Arthur L Da Cunha, Jianping Zhou, and Minh N Do. The nonsubsampled contourlet transform: theory, design, and applications. *IEEE transactions on image processing*, 15(10):3089–3101, 2006. 3

[10] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 6

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[12] Xin Deng and Pier Luigi Dragotti. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3333–3348, 2020. 2

[13] Zhenyu Duan, Jinpeng Lan, Yi Xu, Bingbing Ni, Lixue Zhuang, and Xiaokang Yang. Pedestrian detection via bi-directional multi-scale analysis. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1023–1031, 2017. 3

[14] Clément Fredembach and Sabine Susstrunk. Illuminant estimation and detection using near-infrared. In *Digital Photography V*, volume 7250, pages 112–122. SPIE, 2009. 3

[15] 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*, pages 2782–2790, 2016. 2

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 3

[17] Prabath Gunawardane, Tom Malzbender, Ramin Samadani, Alan McReynolds, Dan Gelb, and James Davis. Invisible light: Using infrared for video conference relighting. In *2010 IEEE International Conference on Image Processing*, pages 4005–4008. IEEE, 2010. 3

[18] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1780–1789, 2020. 1, 3, 6, 7

[19] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on image processing*, 26(2):982–993, 2016. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[21] Zhiwei Hong, Xiaocheng Fan, Tao Jiang, and Jianxing Feng. End-to-end unpaired image denoising with conditional adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4140–4149, 2020. 5

[22] Haidi Ibrahim and Nicholas Sia Pik Kong. Brightness preserving dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics*, 53(4):1752–1758, 2007. 2

[23] Geonwoon Jang, Wooseok Lee, Sanghyun Son, and Kyoung Mu Lee. C2n: Practical generative noise modeling for real-world denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2350–2359, 2021. 2, 5

[24] Haiyang Jiang and Yinqiang Zheng. Learning to see moving objects in the dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7324–7333, 2019. 1, 3, 6, 7

[25] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30:2340–2349, 2021. 1, 3, 6, 7

[26] Shuangping Jin, Bingbing Yu, Minhao Jing, Yi Zhou, Jiajun Liang, and Renhe Ji. Darkvisionnet: Low-light imaging via rgb-nir fusion with deep inconsistency prior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1104–1112, 2022. 2, 3, 6, 7, 8

[27] Daniel J Jobson, Zia-ur Rahman, and Glenn A Woodell. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image processing*, 6(7):965–976, 1997. 2

[28] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 6

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

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

[31] Joon-Young Kwak, Byoung Chul Ko, and Jae Yeal Nam. Pedestrian tracking using online boosted random ferns learning in far-infrared imagery for safe driving at night. *IEEE Transactions on Intelligent Transportation Systems*, 18(1):69–81, 2016. 3

[32] Edwin H Land. The retinex theory of color vision. *Scientific american*, 237(6):108–129, 1977. 2

[33] Chulwoo Lee, Chul Lee, and Chang-Su Kim. Contrast enhancement based on layered difference representation of 2d histograms. *IEEE transactions on image processing*, 22(12):5372–5384, 2013. 2

[34] John J Lewis, Robert J O'Callaghan, Stavri G Nikolov, David R Bull, and Nishan Canagarajah. Pixel-and region-based image fusion with complex wavelets. *Information fusion*, 8(2):119–130, 2007. 3

[35] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2018. 3

[36] Mading Li, Jiaying Liu, Wenhan Yang, Xiaoyan Sun, and Zongming Guo. Structure-revealing low-light image enhancement via robust retinex model. *IEEE Transactions on Image Processing*, 27(6):2828–2841, 2018. 2

[37] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9039–9048, 2018. 4

[38] Matthias Limmer and Hendrik PA Lensch. Infrared colorization using deep convolutional neural networks. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 61–68. IEEE, 2016. 3

[39] Yunfei Liu, Yu Li, Shaodi You, and Feng Lu. Unsupervised learning for intrinsic image decomposition from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4

[40] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 5

[41] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61:650–662, 2017. 1, 3

[42] 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*, volume 34, pages 11725–11732, 2020. 3

[43] Jinlei Ma, Zhiqiang Zhou, Bo Wang, and Hua Zong. Infrared and visible image fusion based on visual saliency map and weighted least square optimization. *Infrared Physics & Technology*, 82:8–17, 2017. 3

[44] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5637–5646, 2022. 1, 3, 6, 7

[45] Armin Mehri and Angel D Sappa. Colorizing near infrared images through a cyclic adversarial approach of unpaired samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[46] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 7

[47] Yusuke Monno, Hayato Teranaka, Kazunori Yoshizaki, Masayuki Tanaka, and Masatoshi Okutomi. Single-sensor rgb-nir imaging: High-quality system design and prototype implementation. *IEEE Sensors Journal*, 19(2):497–507, 2018. 3

[48] Adam Nyberg, Abdelrahman Eldesokey, David Bergstrom, and David Gustafsson. Unpaired thermal to visible spectrum transfer using adversarial training. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 3

[49] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[50] Zahra Sadeghipoor, Yue M Lu, and Sabine Süsstrunk. Correlation-based joint acquisition and demosaicing of visible and near-infrared images. In *2011 18th IEEE International Conference on Image Processing*, pages 3165–3168. IEEE, 2011. 3

[51] Xiaoyong Shen, Qiong Yan, Li Xu, Lizhuang Ma, and Jiaya Jia. Multispectral joint image restoration via optimizing a scale map. *IEEE transactions on pattern analysis and machine intelligence*, 37(12):2518–2530, 2015. 2

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[53] J Alex Stark. Adaptive image contrast enhancement using generalizations of histogram equalization. *IEEE Transactions on image processing*, 9(5):889–896, 2000. 2

[54] Patricia L Suárez, Angel D Sappa, and Boris X Vintimilla. Infrared image colorization based on a triplet dcgan architecture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–23, 2017. 3

[55] Sabine Süsstrunk, Clément Fredembach, and Daniel Tamburrino. Automatic skin enhancement with visible and near-infrared image fusion. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1693–1696, 2010. 3

[56] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via

cross-modality image generation and registration. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 3508–3515. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track. 3

[57] Fengqiao Wang, Lu Liu, and Cheolkon Jung. Deep near infrared colorization with semantic segmentation and transfer learning. In *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 455–458. IEEE, 2020. 3

[58] Ruixing Wang, Xiaogang Xu, Chi-Wing Fu, Jiangbo Lu, Bei Yu, and Jiaya Jia. Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9700–9709, 2021. 1, 3

[59] 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*, pages 6849–6857, 2019. 1, 3

[60] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE transactions on image processing*, 22(9):3538–3548, 2013. 2

[61] Yuzhi Wang, Haibin Huang, Qin Xu, Jiaming Liu, Yiqun Liu, and Jue Wang. Practical deep raw image denoising on mobile devices. In *European Conference on Computer Vision*, pages 1–16. Springer, 2020. 1, 3

[62] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 1, 3

[63] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2758–2767, 2020. 3

[64] Guangming Wu, Yinqiang Zheng, Zhiling Guo, Zekun Cai, Xiaodan Shi, Xin Ding, Yifei Huang, Yimin Guo, and Ryosuke Shibasaki. Learn to recover visible color for video surveillance in a day. In *European Conference on Computer Vision*, pages 495–511. Springer, 2020. 3

[65] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5901–5910, 2022. 6, 7

[66] Xiaohe Wu, Ming Liu, Yue Cao, Dongwei Ren, and Wangmeng Zuo. Unpaired learning of deep image denoising. In *European conference on computer vision*, pages 352–368. Springer, 2020. 2, 5

[67] Jinhui Xiong, Jian Wang, Wolfgang Heidrich, and Shree Nayar. Seeing in extra darkness using a deep-red flash. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10000–10009, 2021. 2, 3, 6, 7, 8

[68] Han Xu, Pengwei Liang, Wei Yu, Junjun Jiang, and Jiayi Ma. Learning a generative model for fusing infrared and visible images via conditional generative adversarial network with dual discriminators. In *IJCAI*, pages 3954–3960, 2019. 3

[69] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2020. 3

[70] Huanjing Yue, Cong Cao, Lei Liao, Ronghe Chu, and Jingyu Yang. Supervised raw video denoising with a benchmark dataset on dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2301–2310, 2020. 1, 3

[71] Fan Zhang, Shaodi You, Yu Li, and Ying Fu. Hsi-guided intrinsic image decomposition for outdoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 313–322, 2022. 4

[72] Yi Zhang, Hongwei Qin, Xiaogang Wang, and Hongsheng Li. Rethinking noise synthesis and modeling in raw denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4593–4601, 2021. 3

[73] Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *International Conference on 3D Vision*, 2022. 4

[74] Chuanjun Zheng, Daming Shi, and Wentian Shi. Adaptive unfolding total variation network for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4439–4448, 2021. 6, 7

[75] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 3, 7, 8

[76] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019. 6