Seeing Dynamic Scene in the Dark: A High-Quality Video Dataset with Mechatronic Alignment

Ruixing Wang[1,2*] Xiaogang Xu[1*] Chi-Wing Fu[1] Jiangbo Lu[2] Bei Yu[1,2] Jiaya Jia[1,2]
[1] The Chinese University of Hong Kong [2] SmartMore
{rxwang,xgxu,cwfu,byu,leojia}@cse.cuhk.edu.hk, jiangbo@smartmore.com

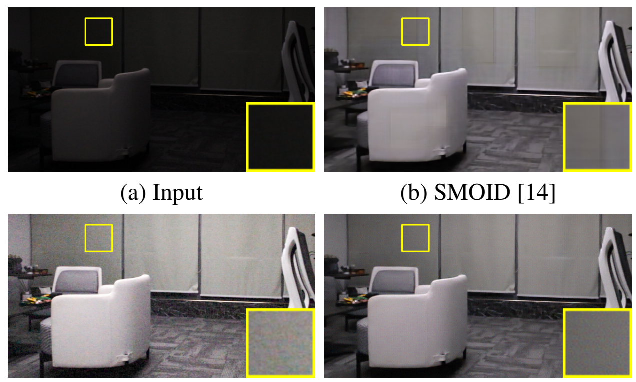
Abstract

Low-light video enhancement is an important task. Previous work is mostly trained on paired static images or videos. We compile a new dataset formed by our new strategy that contains high-quality spatially-aligned video pairs from dynamic scenes in low- and normal-light conditions. We built it using a mechatronic system to precisely control the dynamics during the video capture process, and further align the video pairs, both spatially and temporally, by identifying the system's uniform motion stage. Besides the dataset, we propose an end-to-end framework, in which we design a self-supervised strategy to reduce noise, while enhancing the illumination based on the Retinex theory. Extensive experiments based on various metrics and large-scale user study demonstrate the value of our dataset and effectiveness of our method. The dataset and code are available at https://github.com/dvlab-research/SDSD.

1. Introduction

To enhance underexposed images and videos captured in low light is a longstanding task in computer vision. It is challenging since underexposed input does not has much scene structural information. Also, dark areas are typically dominated by noise with low signal-to-noise ratios (see Figure 1(a)). When enhancing such input, one may end up with amplified noise and undesirable visual artifacts in results, as shown in Figure 1(b)&(c). These issues could be exaggerated for videos taken from dynamic scenes, in which the cameras move largely. In this paper, we focus on *enhancing underexposed videos taken from low-light dynamic scenes*.

Many methods [34, 18, 9, 6, 25, 20, 4] have been proposed to enhance underexposed images/videos based on deep neural networks via supervised learning. Often these methods learn a mapping from images/videos taken in low-light condition to those with normal lighting. They generally do not deal with videos of dynamic scenes or severely-



(a) Input (b) SMOID [14]
(c) Auto-Tone in Lightroom (d) Ours

Figure 1: An example frame (a) from a challenging underexposed frame enhanced by a SOTA method (b), a commercial software (c), and our method (d). Our result exhibits clearer details with distinct contrast and less noise.

underexposed videos corrupted by heavy noise. A major reason comes from the lack of suitable datasets – there is no real-world spatially-aligned video pair in high quality for dynamic scenes.

The inherent difficulty of constructing such a dataset is the following. First, to prepare this type of video pair means that one needs to capture two videos – one in low-light and the other in normal light of the same dynamic scene with identical camera motion. Second, it has to precisely align every pair of corresponding frames in the two videos, both spatially and temporally. Lastly, while beam splitters could be used to alleviate some of the constraints for building a dynamic-scene high-quality dataset, quality of captured videos would be limited [14].

As a result, existing datasets, such as those of [1, 5, 27], provide mainly paired images. Chen *et al*. [4] built a paired video dataset of static scenes, and Jiang *et al*. [14] released a paired-video dataset of dynamic scenes in limited quality. Our first goal in this work is to construct a new dataset with high-quality spatially-aligned video pairs that feature dynamic scenes.

Besides, for videos in low-light conditions, noise often dominates. When we light up video frames, noise can be

*Equal Contribution.

undesirably amplified, leading to various visual artifacts in the enhancement results. In this work, our second goal is to develop a new solution to enhance underexposed videos, taking noise into account.

Our contribution is the following. First, we release a new dataset of 150 high-quality spatially-aligned videos that feature the same dynamic scenes in low- and normal-light conditions. To ensure the alignment and quality of the videos, we built a mechatronic alignment system, in which we assembled an electric slide rail and mounted a professional camera on it; see Figure 2. Using this system, we captured videos of nearly-identical camera motion, thereby reducing the effort needed to align the low- and normal-light videos for temporal and spatial consistency. The constructed dataset is named as SDSD dataset, standing for "Seeing Dynamic Scenes in the Dark."

Second, we formulate an end-to-end framework for enhancing underexposed videos. We emphasize noise reduction and illumination enhancement simultaneously in our method. For noise reduction, we formulate a self-supervised strategy for learning, while for the illumination enhancement, we predict an illumination map from each input frame based on the Retinex theory [16].

Our dataset is the first high-quality paired video dataset for dynamic scenes, featuring high-resolution video pairs of the same scene and motion in low- and normal-light conditions. Trained on our new dataset, our framework works decently for enhancing underexposed videos, even in extremely low-light conditions. To evaluate and demonstrate the applicability and robustness of our new approach, we conducted comprehensive experiments to compare it with a rich set of state-of-the-art methods on our constructed dataset and SMID dataset [4]. Further, we conducted a large-scale user study with 100 participants, showing that our results are visually more pleasing and accurate than previous methods.

Figure 2: The devices in our mechatronic system. In the top row, from left to right is Canon EOS 6D Mark II, the electric machine (to drive the motion of the camera), the controller (to set the starting and ending points for motion), and an ND filter. We mount the camera and the electric machine on the electric slide rail, as shown in the bottom row.

2. Related Work

2.1. Low-light Image Enhancement and Datasets

To enhance a low-light video, one may apply an image enhancement method in a frame-by-frame manner. Histogram equalization and gamma correction are fundamental tools to increase image contrast and expand the dynamic range. Recently, Retinex-based methods [24, 8, 33, 10, 2, 35] produce impressive results enhancing low-light images.

Learning-based low-light image enhancement methods receive increasing attention in recent years [30, 31, 17, 3]. Wang *et al*. [23] proposed to enhance underexposed photos by learning the illumination map. Sean *et al*. [20] learned spatially local filters of three different types to enhance low-light images. Xu *et al*. [28] proposed a frequency-based decomposition and enhancement model to enhance low-images with a low-light dataset based on SID [5]. Yang *et al*. [32] presented a semi-supervised learning method to recover a linear band representation of an enhanced image.

Also, unsupervised learning has been explored for photo enhancement [6, 13, 9]. Guo *et al*. [9] trained a lightweight neural network to estimate pixel-wise and high-order curves for dynamic range adjustment of a given image. However, applying image enhancement algorithms to individual frames likely causes flickering problems.

To improve the enhancement performance, various datasets were built. Bychkovsky *et al*. [1] compiled the large MIT-Adobe FiveK dataset, in which the photos are paired with expert-retouched results for tone adjustment. Chen *et al*. [5] collected raw images of short/long exposure pairs with a U-Net to learn a raw image enhancement system. Recently, Wei *et al*. [27] presented a dataset containing low- and normal-light image pairs and proposed a deep Retinex-Net learned on this dataset.

2.2. Low-light Video Enhancement and Datasets

Zhang *et al*. [34] presented an approach for underexposed video enhancement using a perception-driven progressive fusion. Lv *et al*. [18] proposed a multi-branch network to extract features up to different levels, applicable to both image and video domains. Jiang *et al*. [14] employed a standard CNN to learn enhancement mapping for the transformation from low-light raw camera sensor data to bright videos. However, these methods are not applicable to severe noise conditions.

Xue *et al*. [29] designed a flow representation tailored for specific video processing tasks. Wang *et al*. [25] mathematically defined the practical high sensitivity noise in digital cameras and proposed to enhance low-light videos based on the noise model using a recurrent neural network. Chen *et al*. [4] collected a static dataset of raw low-light videos and learned the low-light to normal-light transformation for videos. Danai *et al*. [22] provided a data synthesis mecha-

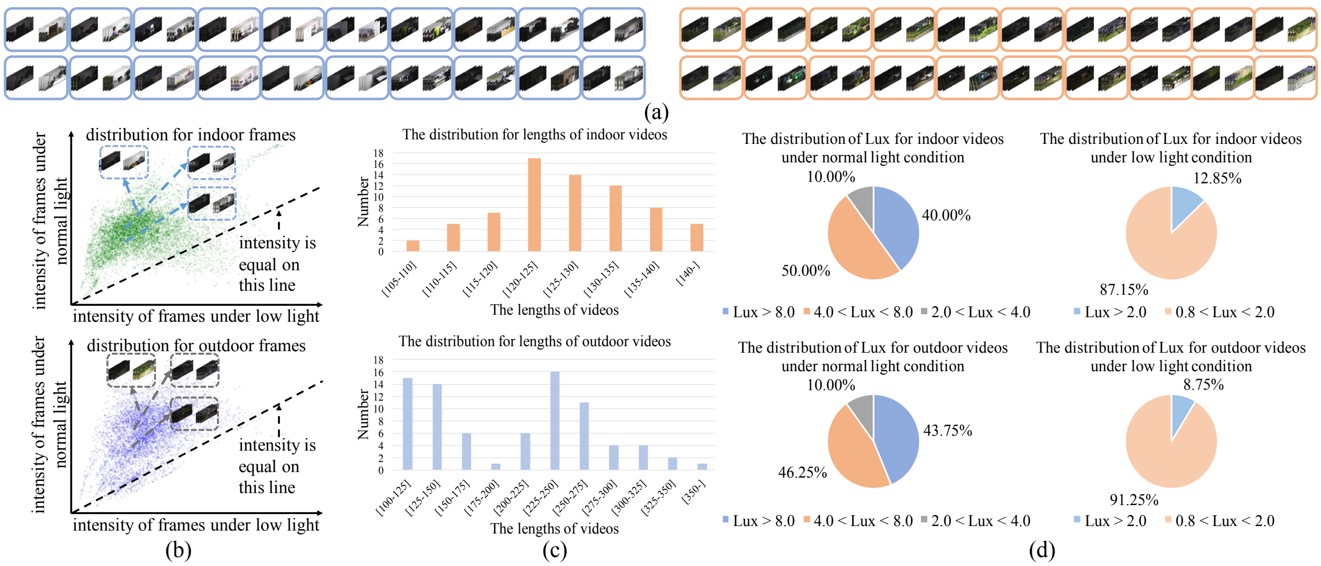


Figure 3: (a) Overview of 25% data in our dataset (zoom in to see more details). (b) Distribution of the intensity for a video pair. We randomly crop two patches at the same location for a video pair under different light conditions and compute the intensity of each patch that corresponds to a point in the 2D plot. Obviously, most cropped patches from normal-light videos have a higher intensity than the corresponding patches in the low-light counterpart. (c) Distribution of video length (the number of frames) in our dataset. (d) Distribution of video Lux in our dataset.

Table 1: Comparison between our dataset and those from the previous work.

Dataset	Release	Status	Capture Device	Numbers
EHSC [25]	×	Dynamic	Canon 5D Mark III	900
SMID [4]	√	Static	Sony RX100 VI	22,220
SMOID [14]	×	Dynamic	FLIR GS3-U3-23S6C	35,800
Ours	√	Dynamic	Canon 6D Mark II	37,500

nism to generate dynamic video pairs from SID [4].

Although several datasets for video enhancement have been proposed recently, they are with different limitations. For example, the dataset is not released yet for [14, 25, 22]. The dataset consists of only static videos [4], while the video quality is limited in the one proposed in [14]. Current representative dataset issues are summarized in Table 1.

In contrast to the previous work, we provide a high-quality dataset via mechatronic alignment. It is made publicly available. Besides, our method is complementary to current learning-based methods. Particularly, we design a new network to handle the underexposed dark areas in videos, and enable the correction of illumination and noise suppression in these areas simultaneously.

3. SDSD Dataset with Mechatronic Alignment

Supervised low-light video enhancement for dynamic scenes is challenging. Spatially paired video data in high quality from real dynamic scenes needs much effort to collect. If we use two cameras to gather paired data, the first way is to use camera pose estimation, like DPED [12]. But this solution causes misalignment. Another way is to utilize a beam splitter to build an optical system. Nevertheless, it is hard to capture high-quality videos since professional cameras cannot be mounted on such an optical system. The dataset of SMOID [14] is not released yet up to our submission time.

In contrast to these strategies, we collected paired videos by employing an electric slide rail as shown in Figure 4, which can repeatedly move along a path precisely within 1mm error. This allows us to manage dynamics in scenes by accurately controlling the camera motion with the electric slide rail. Therefore, we can capture a pair of videos under different light conditions from a scene by running the electric slide rail with the camera for two rounds, as shown in Figure 4. Such a pair can be later spatially aligned. In a nutshell, collecting data consists of capturing and aligning, which will be described as follows.

3.1. Capture Video Data

To control the trace of the camera, we set the starting point $\boldsymbol{A}$ and endpoint $\boldsymbol{B}$ on the electric slide rail. The camera starts capturing videos at point $\boldsymbol{A}$, then moves towards point $\boldsymbol{B}$. To capture a pair of videos, we run the slide rail by two rounds. In the first round, we capture a noise-free bright video with good contrast and vivid color. In the second round, we put the ND filter on the camera lens and increase the camera ISO to capture a low-light video with severe noise.

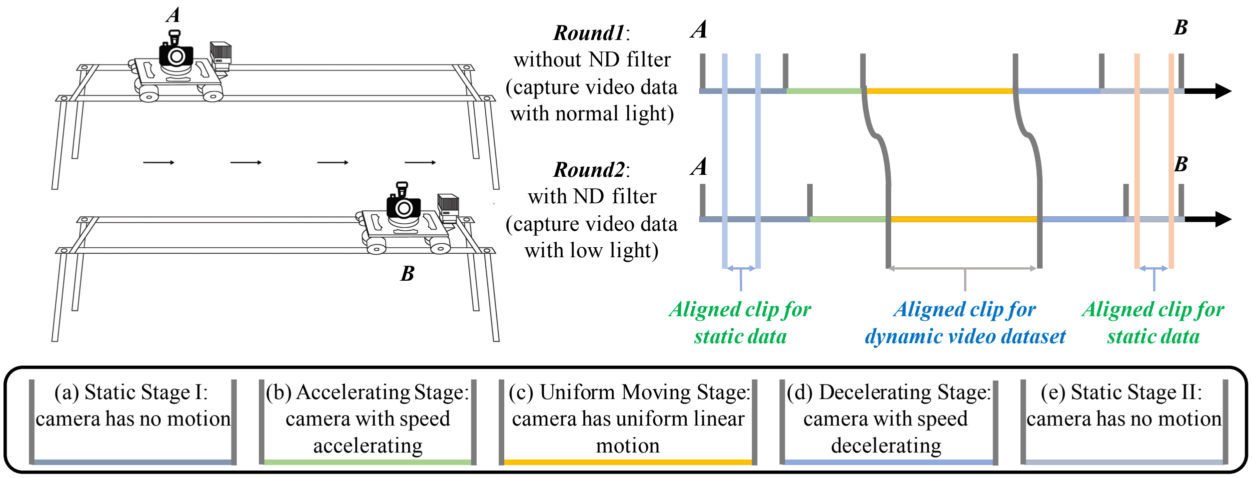


Figure 4: To capture videos with our electric slide rail system, we mount the camera and use an electric machine to drive it. The camera motion path is governed by the controller and this motion process consists of five stages. The start point of the path is $\boldsymbol{A}$ and the endpoint is $\boldsymbol{B}$. To collect a video pair, we run this system for two rounds. In the first round, we capture a video in normal light, and then we use an ND filter to collect a low-light video in the second round. The captured videos are aligned by finding the uniform moving stage. Videos in the static stage can be utilized as static data pairs.

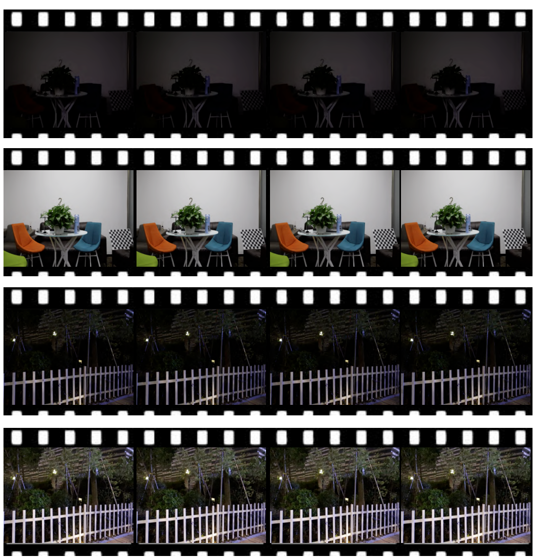


Figure 5: Two video clips in our dataset. For each clip, the first row is captured by Canon EOS 6D Mark II with an ND filter, and the second row is captured under normal light.

3.2. Align Video Data

Alignment of videos was conducted according to the camera trace, which consists of five stages (Figure 4) — that is, static stage I, accelerating stage, uniform moving stage, decelerating stage, and static stage II.

The camera in static stages I and II locates at points $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively, and has no motion. The accelerating stage leads to the camera with speed accelerating at the beginning of the moving trace and the decelerating stage is at the ending of the moving trace. Aligning frames of the two sequences in the same position is easy in the uniform moving stage, where the camera motion is stable. Thus, we choose the frames in the uniform moving stage to construct our video dataset.

First, we find the first frame in the uniform moving stage from the normal/low-light video. Then we manually pick the aligned frame from the uniform moving stage in a frame-wise style, until finding the dis-alignment frame in the decelerating stage. Specifically, we adopt the reference objects in the top, bottom, left, right of the frames to measure the alignment where the reference objects should locate at the same position for the two aligned frames.

We collected 150 paired video sequences in total, including 80 outdoor videos and 70 indoor videos. Each video consists of 100-300 frames, and the resolution is $1,920 \times 1,080$. Our dataset is called SDSD, and Figure 3 shows 25% of the data in our dataset and the statistical indicators of the overall dataset. In our dataset, there are various scenes, such as cityscapes, grassland, and indoors. In Figure 5, we provide two examples for indoor/outdoor sequence under low- and normal-light conditions.

4. Method

Besides our constructed dataset with mechatronic alignment, we design an effective video enhancement framework. For simplicity, $I_t \in \mathbb{R}^{H \times W \times 3}$ denotes a low-light video frame and $\bar{I}_t \in \mathbb{R}^{H \times W \times 3}$ represents the paired frame under normal-light condition. Given a sequence of frames $(I_{t+i}, i \in [-2, 2])$, we aim to enhance the illumination of the middle frame I_t.

It is observed that severe noise inevitably occurs in videos

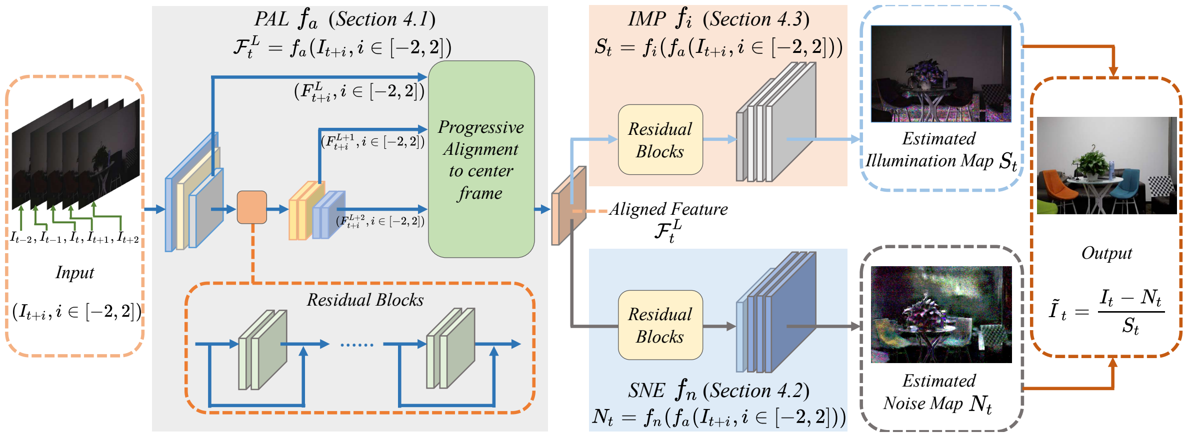


Figure 6: Overview of our framework, which achieves noise reduction and illumination enhancement. It consists of the modules of progressive alignment (PAL), noise estimation (SNE), and illumination prediction (IMP).

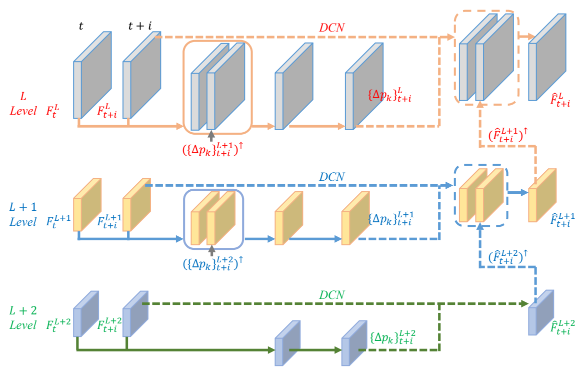


Figure 7: Detailed illustration for "progressive alignment to center frame" in Figure 6.

taken in a dark environment [25] and enhancing illumination would further amplify it. We first formulate physical noise in dark as $I_t = \mathcal{I}_t + \epsilon$, where $\mathcal{I}_t$ is the clean frame under a low-light condition without noise and ϵ is the noise.

According to the Retinex-based enhancement theory [16], an illumination map S_t can be computed to recover this frame to a normal illumination condition as

$$\frac{I_t}{S_t} = \frac{\mathcal{I}_t + \epsilon}{S_t} = \frac{\mathcal{I}_t}{S_t} + \frac{\epsilon}{S_t},$$

where $\frac{\mathcal{I}_t}{S_t}$ is the enhanced frame without noise, and $\frac{I_t}{S_t}$ is the frame with amplified noise $\frac{\epsilon}{S_t}$. Thus, directly enhancing frame illumination will cause noise amplification.

Instead of solely enhancing the illumination like traditional methods, we propose an end-to-end network as shown in Figure 6, simultaneously achieving noise reduction and illumination enhancement. This network consists of the modules of progressive alignment (PAL), noise estimation (SNE), and illumination prediction (IMP).

4.1. Progressive Alignment (PAL)

Directly conducting low-light image enhancement [23, 9, 20] to each frame causes flickering. To avoid it and take advantage of temporal information, the input of existing video enhancement methods is a sequence. Meanwhile, to produce frames without blur, existing video enhancement methods consider aligning neighboring frames into the middle one [36, 21, 15]. Such alignment can be executed at the feature level. In this section, we illustrate the process of feature extraction and our progressive strategy for alignment.

Given the input sequence under the low-light condition as $(I_{t+i}, i \in [-2, 2])$ with shape $\mathbb{R}^{5\times H\times W\times 3}$, we extract these frame features with three convolution layers and two down-sampling layers, to propagate the information spatially and temporally, as shown in Figures 6 and 7. The obtained features are denoted as $(F_{t+i}^L, i \in [-2, 2])$ with shape $\mathbb{R}^{5\times \frac{H}{4}\times \frac{W}{4}\times C}$, where C is the number of feature channels, and L is one level in the progressive alignment.

The alignment module spatially aligns features of neighboring frames to the central one, which is realized by the deformable convolution (DCN) [7] progressively, as illustrated in Figure 7. To align F_{t+i}^L, $i \in [-2, 2]$, we extract features with different levels as $F_{t+i}^{L+1}, i \in [-2, 2]$ and F_{t+i}^{L+2}, $i \in [-2, 2]$ that have shape $\mathbb{R}^{5\times \frac{H}{8}\times \frac{W}{8}\times C}$ and $\mathbb{R}^{5\times \frac{H}{16}\times \frac{W}{16}\times C}$.

We first compute the offset $\{\Delta p_k\}_{t+i}^{L+2}$ for the DCN in level $L+2$. The offset is learned from F_{t+i}^{L+2} and F_t^{L+2}, and the aligned feature is obtained with the learned offset as

$$\begin{aligned} \{\Delta p_k\}_{t+i}^{L+2} &= f^{L+2}(F_{t+i}^{L+2} \odot F_t^{L+2}), \\ \widehat{F}_{t+i}^{L+2} &= g^{L+2}(DCN(F_{t+i}^{L+2}, \{\Delta p_k\}_{t+i}^{L+2})), \end{aligned} \tag{1}$$

where $\{\Delta p_k\}_{t+i}^{L+2}$ is the learned offset for DCN at $(L + 2)$-th level, $\odot$ denotes channel concatenation, f^{L+2} and g^{L+2} are the mapping function completed by several convolution layers, and DCN is the operation of DCN. To implement the progressive learning, we employ the computed offset at $(L+2)$-th and $(L+1)$-th levels for offset computation at $(L+1)$-th and L-th levels. Further, we set the progressive learning for updating features at each level by incorporating the features from other levels. The process can be written as

$$\begin{aligned}\{\Delta p_k\}_{t+i}^{L+j} &= f^{L+j}(F_{t+i}^{L+j} \odot F_t^{L+j} \odot (\{\Delta p_k\}_{t+i}^{L+j+1})^{\uparrow}),\\ \widehat{F}_{t+i}^{L+j} &= g^{L+j}(DCN(F_{t+i}^{L+j}, \{\Delta p_k\}_{t+i}^{L+j}) \odot (\widehat{F}_{t+i}^{L+j+1})^{\uparrow}),\end{aligned} \tag{2}$$

where $(\{\Delta p_k\}_{t+i}^{L+j+1})^{\uparrow}$ is the upsampled offset, $(\widehat{F}_{t+i}^{L+j+1})^{\uparrow}$ is the upsampled feature and $j \in \{0,1\}$. With the aligned features $\widehat{F}_{t+i}^{L}$, we fuse them with the similarity between $\widehat{F}_{t+i}^{L}$ and $\widehat{F}_t^{L}$. The process to obtain the aligned feature can be denoted as $\mathcal{F}_t^L = f_a(I_{t+i}), i \in [-2,2]$, where $\mathcal{F}_t^L$ has shape of $\mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$. Such alignment works well for videos with the smooth local motion, since the motion in the corresponding input can be simulated as the translation.

4.2. Self-Supervised Noise Estimation (SNE)

After obtaining the aligned feature $\mathcal{F}_t^L$, we utilize it for two purposes: noise estimation and illumination map prediction. Implementation of noise estimation is described in this section. With the input I_{t+i}, we aim to predict a noise map N_t with shape as $\mathbb{R}^{H\times W\times 3}$, and the recovered frame can be obtained as $I_t - N_t$. The module of noise estimation is trained with the principle of "Noisier2Noise" [19] where we add the crafted noise to an input noisy/clean frame and train the network to regress the added noise. Thus, the noise estimation module can be learned in a self-supervised way.

As shown in Figure 6, to estimate noise, the aligned feature $\mathcal{F}_t^L$ forwards through a network f_n that consists of residual blacks and two layers for up-sampling. This SNE network produces the computed noise map N_t as $N_t = f_n(\mathcal{F}_t^L) = f_n(f_a(I_{t+i})), i \in [-2,2]$. For training, we compute the average RGB value of I_{t+i} to create noise to be added to I_{t+i}, so that the noise magnitude can be more relevant to image contents of I_{t+i}. We compute the loss as

$$\begin{aligned}\widehat{N}_{t+i} &= I_{t+i} - \mathcal{M}(I_{t+i}), i \in [-2,2],\\ \mathcal{L}_n &= \mathbb{E}(\|f_n(f_a(I_{t+i} + \widehat{N}_{t+i}, i \in [-2,2])) - \widehat{N}_t\|),\end{aligned} \tag{3}$$

where $\widehat{N}_{t+i}$ is the created noise, $\mathcal{M}(I_{t+i})$ is the average RGB value of I_{t+i}, $\mathbb{E}$ is the operation to compute the average value, and $\mathcal{L}_n$ is the loss term for training f_n.

4.3. Illumination Map Prediction (IMP)

According to the Retinex-based methods [23], we enhance the illumination of I_t by predicting an illumination map $\frac{I_t}{\bar{I}_t}$. Unlike existing Retinex-based methods, we propose to train a noise-aware network for estimating an illumination map. The illumination map should be consistent with the content of frames and not be influenced by the noise.

As shown in Figure 6, we adopt another network f_i with the input of $\mathcal{F}_t^L$ to predict the illumination map. This IMP module also consists of residual blocks and two layers for up-sampling. We formulate this process to acquire the illumination map S_t as $S_t = f_i(f_a(I_{t+i}, i \in [-2,2]))$ and the output size of the illumination map is $\mathbb{R}^{H\times W\times 3}$. The loss term to train f_i is written as

$$\begin{aligned}\mathcal{L}_{ic} &= \mathbb{E}(\|f_i(f_a(I_{t+i}, i \in [-2,2])) - \frac{I_t}{\bar{I}_t}\|),\\ \mathcal{L}_{in} &= \mathbb{E}(\|f_i(f_a(I_{t+i} + \widehat{N}_{t+i}, i \in [-2,2])) - \frac{I_t}{\bar{I}_t}\|),\end{aligned} \tag{4}$$

where $\widehat{N}_t$ is the crafted noise defined in Eq. (3). $\mathcal{L}_{ic}$ and $\mathcal{L}_{in}$ are the loss terms for training f_i.

4.4. Overall Loss Function

We denote the output of $(I_{t+i}, i \in [-2,2])$ from our network as $(S_t, N_t) = f(I_{t+i}, i \in [-2,2])$, where f denotes the function implemented by our network. The final enhanced frame can be obtained as $\tilde{I}_t = \frac{I_t - N_t}{S_t}$. To this end, we add a loss function for $\tilde{I}_t$ as the constraint for f_a, f_n and f_i simultaneously, which can be written as

$$\mathcal{L}_b = \mathbb{E}(\|\tilde{I}_t - \bar{I}_t\|). \tag{5}$$

Moreover, to ensure the effect of enhancement with noisy input, we set another constraint for f_a, f_n and f_i as

$$\begin{aligned}(S_t', N_t') &= f(I_{t+i} + \widehat{N}_{t+i}, i \in [-2,2]),\\ \mathcal{L}_{bn} &= \mathbb{E}(\|\frac{I_t + \widehat{N}_t - N_t'}{S_t'} - \bar{I}_t\|),\end{aligned} \tag{6}$$

where $\widehat{N}_t$ is the crafted noise defined in Eq. (3).

The overall loss function to train this framework is summarized as

$$\mathcal{L}_a = \lambda_1\mathcal{L}_n + \lambda_2(\mathcal{L}_{ic} + \mathcal{L}_{in}) + \lambda_3\mathcal{L}_b + \lambda_4\mathcal{L}_{bn}, \tag{7}$$

where λ_1, λ_2, λ_3, and λ_4 are weights of the loss terms. We empirically set $\lambda_1 = 2$, $\lambda_2 = 0.25$, $\lambda_3 = 0.5$, and $\lambda_4 = 0.5$.

5. Experiments

5.1. Experiment Setup

We demonstrate the superiority of our method and the impact of SDSD through experiments in this section. To illustrate the effect of our method, we retrain seven previous representative methods on the SDSD and SMID [4] datasets for comparison and provide an ablation study for our method. Besides, we conduct user study to evaluate the results of our method and the chosen baselines.

Further, we compare the performance of two models with our designed network structure that are trained on SDSD and SMID [4], respectively, and conduct the evaluation on real-world videos captured from mobile devices. For the SMID dataset, we use SMID pre-processing to process the RAW data to produce the sRGB data. The comparison between

Table 2: Quantitative comparison among our method, state-of-the-art baselines, and ablation settings on our SDSD and the SMID [4] dataset. PSNR is in dB.

Methods	SDSD		SMID	
	PSNR	SSIM	PSNR	SSIM
DeepUPE [23]	21.82	0.68	23.91	0.69
ZeroDCE [9]	20.06	0.61	22.62	0.67
DeepLPF [20]	22.48	0.66	24.36	0.69
DRBN [32]	22.31	0.65	24.42	0.69
MBLLEN [18]	21.79	0.65	22.67	0.68
SMID [4]	24.09	0.69	24.78	0.72
SMOID [14]	23.45	0.69	23.64	0.71
Ours w/o PAL, w/o IMP, w/o SNE	22.61	0.64	25.04	0.71
Ours with PAL, w/o IMP, w/o SNE	24.47	0.65	25.32	0.71
Ours with PAL, with IMP, w/o SNE	24.53	0.67	25.71	0.74
Ours	**24.92**	**0.73**	**26.03**	**0.75**

these two models via subjective evaluations show that our SDSD is perceptually better than the static SMID dataset to enhance videos captured from dynamic scenes. One visual example is shown in Figure 10.

Similar to previous work [11, 23, 20], we employ two commonly-used metrics of peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [26]. High PSNR and SSIM values suggest good results.

5.2. Comparison on the SDSD Dataset

We compare our approach with seven state-of-the-art methods on the SDSD dataset and provide the qualitative results in Figure 8. It is clear that our result mainly has two advantages over other methods. First, the result from our proposed method has high contrast, clear details, and natural color constancy. Thus, frames processed by our method are more realistic. Second, see the wall and floor areas, our output has fewer visual artifacts and less noise, and looks cleaner than all baselines. These facts show that our model can achieve noise reduction and illumination enhancement. The effect is better than the already strong baselines.

Moreover, we provide quantitative results in Table 2 for comparison. As exhibited in Table 2, our method achieves the highest PSNR and SSIM on the SDSD dataset. Especially, our PSNR is higher than all baselines with a large margin (more than 0.8dB). This superiority validates that our method yields greater performance for low-light video enhancement compared with all baselines.

Besides, we perform ablation study to evaluate the effectiveness of the components in our method. Comparing the values presented in the last row ("Ours") with the values listed in the three rows above "Ours" (Ours with/without PAL, IMP, and SNE) in Table 2, we observe clear progress brought by adding PAL, IMP, and SNE into our framework.

Table 3: User preference comparison in the user study. "Ours" is the percentage that our result is preferred, "Other" is the percentage that some other method is preferred, "Same" is the percentage that the users have no preference.

Methods	Other	Same	Ours
DeepUPE [23]	27.0%	8.7%	64.3%
ZeroDCE [9]	11.9%	4.4%	83.7%
DeepLPF [20]	19.8%	6.8%	73.4%
DRBN [32]	12.5%	19.2%	68.3%
MBLLEN [18]	6.8%	13.5%	79.7%
SMID [4]	13.9%	13.5%	72.6%
SMOID [4]	19.8%	25.4%	54.8%

5.3. Comparison on the SMID Dataset

To demonstrate the generalization of our method, we evaluate the effect of our method/baselines that are trained on SMID [4]. The testing sequences are in 8bit sRGB format.

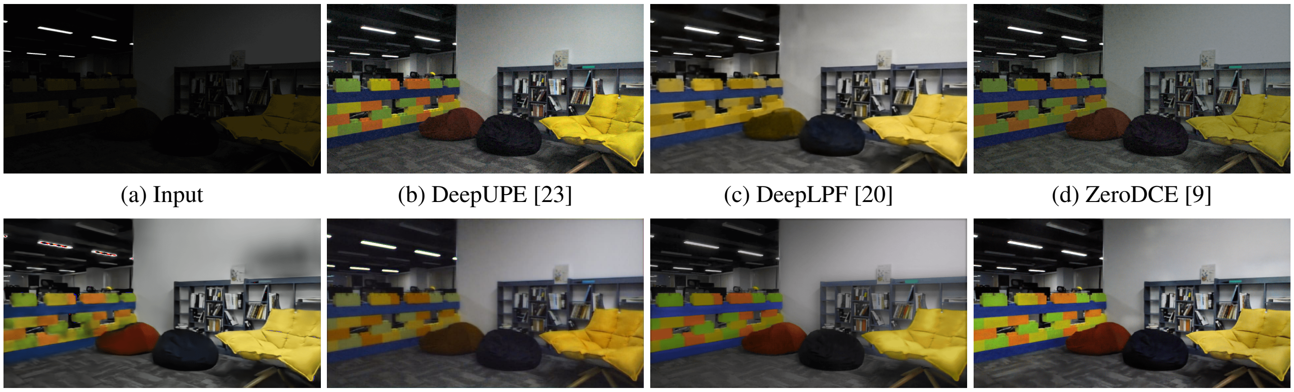
In Figure 9, we provide the results of our method and baselines, which are trained on the training set of SMID [4] while evaluated on the testing set of SMID [4]. Our method restores the underexposed video frames into those with normal brightness and natural color. Also, we provide quantitative results in Table 2 for comparison. Our network achieves the best PSNR and SSIM, and performs better than the baselines with a large margin (more than 1.3dB).

5.4. User Study on the Real Testing Videos

To compare our method with the seven baselines based on human perception, we conduct user study with 100 persons using totally 12 videos, which are captured by iPhone7plus and iPhoneX with real camera motion and local subjects motion. We compute the results of different methods on these videos to conduct an AB-test. All network models are trained on the SDSD dataset.

Each participant saw two videos (called videos A and B) simultaneously, which were synthesized by different methods, and has to choose among three options: "Video A is better", "Video B is better", and "I cannot make a decision on which one is better". For evaluation accuracy, we invited 100 persons to participate in our user study, and each participant was asked to complete 14 pairs of AB-test. Each AB-test was conducted between our result and one of the seven baselines — they were presented in a random left-right order. Participants made decisions according to the following five properties: suitable brightness, clear details, distinct contrast, vivid color, and well-preserved photo realism.

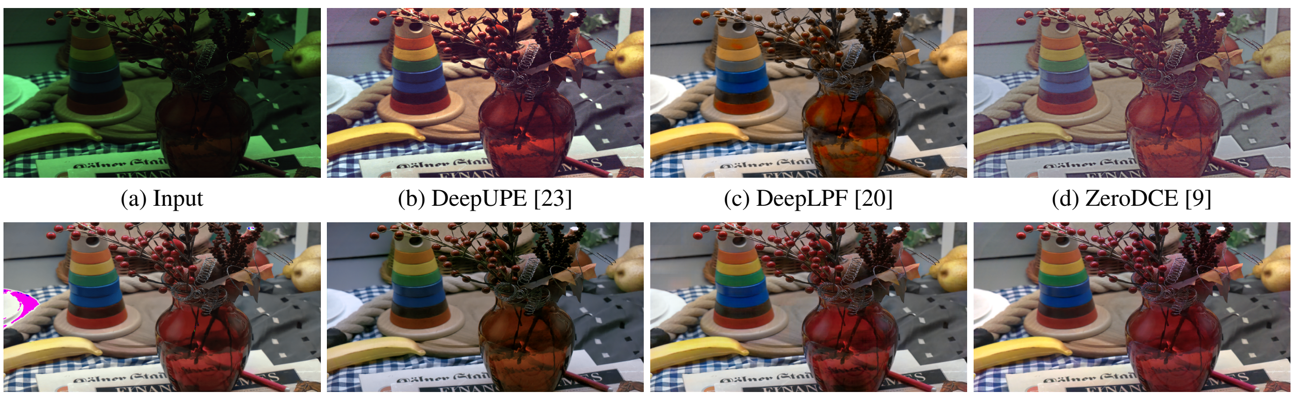
The results of this user study are given in Table 3, where we report the proportion that our results are preferred by participants. It proves that our method yields more appealing and natural results, as the participants often preferred our predicted videos rather than those from the baselines.



(a) Input (b) DeepUPE [23] (c) DeepLPF [20] (d) ZeroDCE [9]

(e) DRBN [32] (f) SMID [4] (g) SMOID [14] (h) Our result

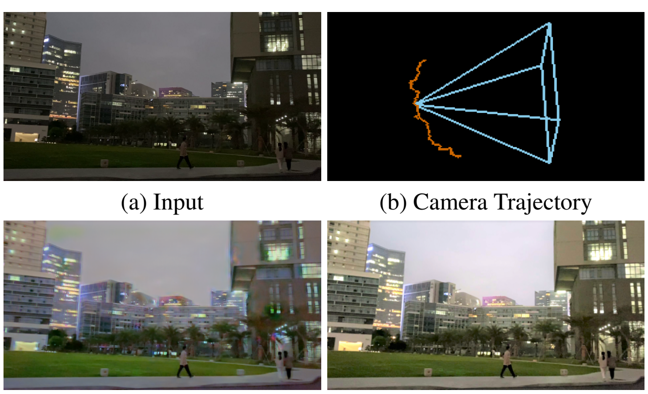
Figure 8: An underexposed video frame (a) enhanced by various methods (b)-(h). Results from baselines exhibit blurry details, noise, distorted color, weak contrast, abnormal brightness, and unnatural white balance (zoom in to see details).



(a) Input (b) DeepUPE [23] (c) DeepLPF [20] (d) ZeroDCE [9]

(e) DRBN [32] (f) SMID [4] (g) SMOID [14] (h) Our result

Figure 9: Another underexposed video frame (a) enhanced by various methods (b)-(h) that are trained on the SMID dataset [4] (zoom in to see details).



(a) Input (b) Camera Trajectory

(c) Ours with SMID [4] (d) Ours with SDSD

Figure 10: A real video frame (a) captured by iPhone X enhanced by our method trained on SMID [4] (c) and our SDSD (d). (b) shows the phone's camera setting (light blue) and its trajectory (orange) in the 3D-reconstructed scene (please zoom in to see details).

6. Conclusion

We have presented a paired high-quality video dataset built using a mechatronic system. Each video pair in our dataset is captured from an indoor/outdoor dynamic scene, containing two spatially-aligned videos taken from low- and normal-light conditions, respectively. Besides, we propose an end-to-end framework for video enhancement. Our framework achieves noise reduction and illumination enhancement simultaneously. Extensive experiments with user study are conducted, demonstrating the value of our dataset and the effectiveness of our method.

The methods trained with the SDSD dataset effectively enhance videos that are captured with a real camera trajectory, e.g., the panning and rotation motion, as shown in the user study. But they may not perfectly handle videos captured with serious camera shaking. Hence, we envision to build another dataset using a robot arm that can precisely repeat the most challenging trajectories.

References

[1] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2011.

[2] Bolun Cai, Xianming Xu, Kailing Guo, Kui Jia, Bin Hu, and Dacheng Tao. A joint intrinsic-extrinsic prior model for retinex. In *Int. Conf. Comput. Vis.*, 2017.

[3] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Trans. Image Process.*, 2018.

[4] Chen Chen, Qifeng Chen, Minh N Do, and Vladlen Koltun. Seeing motion in the dark. In *Int. Conf. Comput. Vis.*, 2019.

[5] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[6] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Int. Conf. Comput. Vis.*, 2017.

[8] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[9] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[10] Xiaojie Guo, Yu Li, and Haibin Ling. LIME: Low-light image enhancement via illumination map estimation. *IEEE Trans. Image Process.*, 2017.

[11] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. Exposure: A white-box photo post-processing framework. *ACM Trans. Graph.*, 2018.

[12] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *Int. Conf. Comput. Vis.*, 2017.

[13] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. WESPE: Weakly supervised photo enhancer for digital cameras. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[14] Haiyang Jiang and Yinqiang Zheng. Learning to see moving objects in the dark. In *Int. Conf. Comput. Vis.*, 2019.

[15] Tae Hyun Kim, Mehdi SM Sajjadi, Michael Hirsch, and Bernhard Scholkopf. Spatio-temporal transformer network for video restoration. In *Eur. Conf. Comput. Vis.*, 2018.

[16] Edwin H Land. The retinex theory of color vision. *Scientific American*, 1977.

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

[18] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. Mbllen: Low-light image/video enhancement using cnns. In *Brit. Mach. Vis. Conf.*, 2018.

[19] Nick Moran, Dan Schmidt, Yu Zhong, and Patrick Coady. Noisier2noise: Learning to denoise from unpaired noisy data. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[20] Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. Deeplpf: Deep local parametric filters for image enhancement. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[21] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[22] Danai Triantafyllidou, Sean Moran, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. Low light video enhancement using synthetic data produced with an intermediate domain mapping. In *Eur. Conf. Comput. Vis.*, 2020.

[23] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[24] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE Trans. Image Process.*, 2013.

[25] Wei Wang, Xin Chen, Cheng Yang, Xiang Li, Xuemei Hu, and Tao Yue. Enhancing low light videos by exploring high sensitivity camera noise. In *Int. Conf. Comput. Vis.*, 2019.

[26] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 2004.

[27] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *Brit. Mach. Vis. Conf.*, 2018.

[28] Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to restore low-light images via decomposition-and-enhancement. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[29] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *Int. J. Comput. Vis.*, 2019.

[30] Jianzhou Yan, Stephen Lin, Sing Bing Kang, and Xiaoou Tang. A learning-to-rank approach for image color enhancement. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014.

[31] Zhicheng Yan, Hao Zhang, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. Automatic photo adjustment using deep neural networks. *ACM Trans. Graph.*, 2016.

[32] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[33] Zhenqiang Ying, Ge Li, Yurui Ren, Ronggang Wang, and Wenmin Wang. A new low-light image enhancement algorithm using camera response model. In *Int. Conf. Comput. Vis.*, 2017.

[34] Qing Zhang, Yongwei Nie, Ling Zhang, and Chunxia Xiao. Underexposed video enhancement via perception-driven progressive fusion. *IEEE Trans. Vis. Comput. Graph.*, 2016.

[35] Qing Zhang, Ganzhao Yuan, Chunxia Xiao, Lei Zhu, and Wei-Shi Zheng. High-quality exposure correction of underexposed photos. In *ACM Int. Conf. Multimedia*, 2018.

[36] Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. Spatio-temporal filter adaptive network for video deblurring. In *Int. Conf. Comput. Vis.*, 2019.