

A Dual Sensor Computational Camera for High Quality Dark Videography

Yuxiao Cheng, Runzhao Yang, Zhihong Zhang, Jinli Suo, and Qionghai Dai

Abstract—Videos captured under low light conditions suffer from severe noise. A variety of efforts have been devoted to image/video noise suppression and made large progress. However, in extremely dark scenarios, extensive photon starvation would hamper precise noise modeling. Instead, developing an imaging system collecting more photons is a more effective way for high-quality video capture under low illuminations. In this paper, we propose to build a dual-sensor camera to additionally collect the photons in NIR wavelength, and make use of the correlation between RGB and near-infrared (NIR) spectrum to perform high-quality reconstruction from noisy dark video pairs. In hardware, we build a compact dual-sensor camera capturing RGB and NIR videos simultaneously. Computationally, we propose a dual-channel multi-frame attention network (DCMAN) utilizing spatial-temporal-spectral priors to reconstruct the low-light RGB and NIR videos. In addition, we build a high-quality paired RGB and NIR video dataset, based on which the approach can be applied to different sensors easily by training the DCMAN model with simulated noisy input following a physical-process-based CMOS noise model. Both experiments on synthetic and real videos validate the performance of this compact dual-sensor camera design and the corresponding reconstruction algorithm in dark videography.

Index Terms—Computational photography, Video denoising, Low light video, Dark vision, RGB-NIR, Dual-channel network.



1 INTRODUCTION

UNDER low light conditions, video photographers usually have to set high ISO (sensitivity) and short exposure time to avoid motion blur, and thus suffer from noise deterioration badly. To address this issue, researchers have been attempting to develop various low-light image/video reconstruction methods in the past decades. The basic idea of noise suppression or removal is introducing prior information of nature scenes to recover the latent clean version, and some methods can achieve excellent reconstruction results, such as [1] [2] [3] [4]. The reconstruction can be conducted on either a single image or a video sequence. Generally speaking, multi-frame noise reduction methods usually get better results than those based on a single frame, thanks to the utilization of priors from temporal redundancy of dynamic scenes.

In spite of the extensive studies in spatial and temporal priors of nature image/videos and their applications in reconstruction tasks, high-quality imaging under extremely low illuminations is still quite challenging. For example, in Fig. 1, the raw measurement is quite low, and after ~ 10 -time magnification, only the general scene structure can be perceived while the details are buried in severe noise and there exists color distortion. To achieve more decent imaging under environments with poor illuminations, one can resort to collecting more photons and introducing new priors. Actually, there exist a bunch of near-infrared lights under low illuminations and a typical camera sensor can respond to photons from wavelength from 350-1200nm [5].



Fig. 1: An example of paired RGB (top) and NIR (bottom) video frame captured computationally by our approach under dark environment. The three segments in each image show the extremely dark raw measurement, linearly scaled version with severe noise and color distortion, and the clean final reconstruction, respectively.

All the authors are with the Department of Automation, Tsinghua University. Jinli Suo and Qionghai Dai are also affiliated with the Institute of Brain and Cognitive Sciences, Tsinghua University. Emails: chengyx18@mails.tsinghua.edu.cn; yangrz20@mails.tsinghua.edu.cn; zhangzh19@mails.tsinghua.edu.cn; jlsuo@tsinghua.edu.cn; qh-dai@tsinghua.edu.cn.

Corresponding authors: Jinli Suo and Qionghai Dai.

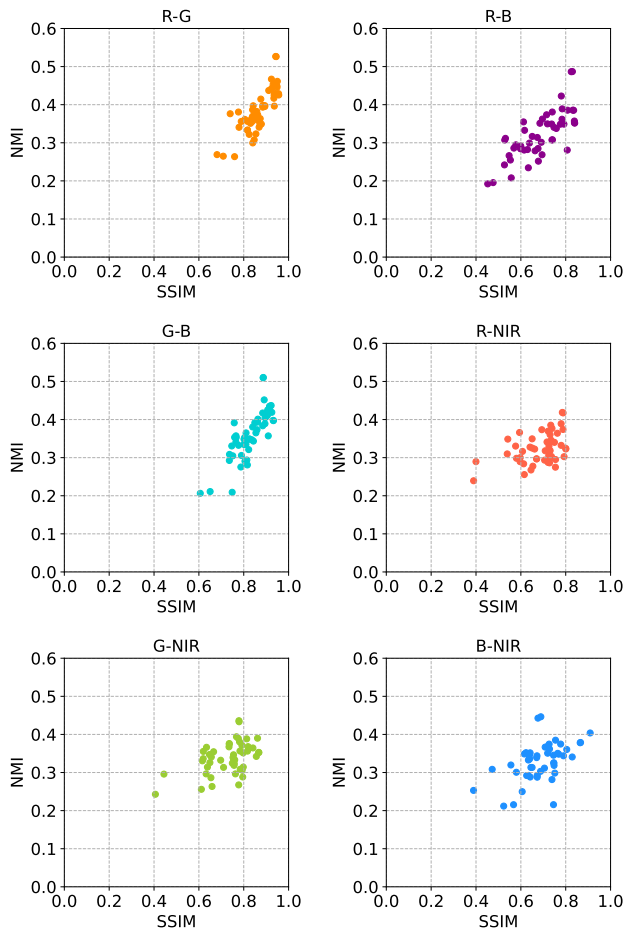


Fig. 2: Analysis of the cross channel correlation among three visible channels (i.e., red, green, blue), and between RGB and NIR channels, with SSIM and NMI as score metrics. These plots are calculated from 50 RGB/NIR image pairs.

Therefore, recording the photons in NIR band additionally can improve the quality of dark imaging. However, existing approaches are lacking in the following aspects: (i) An infrared cut-off filter is placed before the image sensor of RGB/gray camera to cut off NIR light for a natural tone matching the perception of human eyes (RGB, roughly 400-700nm) [6]. As a result, severe noise occurs due to the insufficient photons in visible wavelength. (ii) RGB and NIR images share similar features on scene structure and relative brightness, as shown in the two rows in Fig. 1, and it is possible to make use of such cross-channel correlations between RGB and NIR images to enhance the imaging quality algorithmically. However, current denoising methods rarely utilize this cross channel prior.

Inspired by the above analysis, we design a compact double-sensor camera system to acquire paired RGB and NIR images at the same time, with a beam splitter placed between the primary lens and image sensors. Using this setup, we capture a dataset of RGB-NIR nature video pairs, and analyze the image similarity among the 4 channels (R, G, B, and NIR), in terms of SSIM (Structural Similarity) and NMI (Normalized Mutual Information) scores. The results are plotted in Fig. 2 with the scattered dots concentrated in the right-top region of each subfigure, which validates

that the red, green, blue, and infrared channels are indeed strongly correlated, and the correlations between visible and NIR wavelength are only slightly lower than that among R, G, and B channels. In other words, we can introduce priors across visible and NIR channels to improve the final reconstruction.

Computationally, we design a Dual-channel Multi-frame Attention Network (DCMAN) based on bi-directional ConvLSTM and CNN-based encoder-decoder with mutual multi-level feature fusion [7]. In other words, the data from the two arms contribute mutually to recovering the latent structure of nature videos buried in the severe noise. By additionally introducing the photons in the near-infrared wavelength, and taking advantage of spatial-temporal as well as spectral prior information, we can achieve better reconstruction quality. Our method takes noisy NIR and RGB videos as inputs and gives clean results of both videos at the same time, as shown in Fig. 1. The result demonstrates that the visual quality is largely raised with fine details being reconstructed.

Overall, this paper reports a computational photography approach for dark video capturing and mainly contributes in following aspects:

- We build a dual-sensor camera system capturing RGB and NIR video pairs simultaneously, being able to increase the collected photons effectively for higher imaging quality in dark environments.
- We design a Dual-Channel Multi-frame Attention Network (DCMAN) with Guided Skip Connections (GSCs) to utilize spatial, temporal, and spectral video prior. This network can serve as a general structure applicable for similar tasks taking multiple input channels.
- The approach can be adapted to different sensors and noise levels easily by learning the model from synthetic data, produced by adding noise to our collected high quality RGB + NIR video dataset on real scenes.
- The proposed approach can achieve superior performance to existing methods in most dark scenarios.

2 RELATED WORKS

2.1 Video Denoising

Video denoising has been extensively studied in the past decades [8] [9] [10], and the large amount of literature falls into several groups based on the ways of specifying priors.

Statistically, the distribution of natural images/videos is discriminative from noises globally or over local patches. One way of noise removal is applying filters spatiotemporally or in a transformed domain after motion compensation, making use of the fact that noise components are of high frequency while the latent image/video concentrates at low frequency band [11] [12] [13]. Another way is conducting video denoising over patches, exploiting temporal and spatial patch similarity in the neighborhood. For example, Buates et al. [14] propose to conduct patch-based denoising after compensating motions among neighboring frames, and some other methods can suppress noise without motion compensation, such as non-local means algorithm

[15], and V-BM3D [16] and its variants. V-BM3D [16] is the extension of patch-based method BM3D [17], which extracts similar 2D patches from consecutive video frames and then stacks them for filtering in a transformed domain. Later, they extend V-BM3D to V-BM4D [18] searching similar 3D spatio-temporal blocks instead of 2D patches, and VNLB [19] that conducts inference with Bayesian estimation.

With the rapid development of Convolutional Neural Networks (CNN) and deep learning techniques, learning-based image reconstruction exhibits better performance and is attracting wide attention. In this stream, video denoising, however, is much less explored than image denoising, since extracting spatio-temporal features jointly is non-trivial. On the one hand, researchers try to extend patch-based image denoising by introducing temporal information with CNN feature fusion. Tassano et al. [20] propose DVDnet by explicitly splitting the denoising process into two successive stages. Specifically, consecutive frames are first individually denoised and then temporally denoised with motion compensation. Later, building on DVDnet, they further propose FastDVDnet [3] without flow estimation and achieved higher performance. Using a similar two-stage method, Claus and Gemert [2] design ViDeNN, and Mildenhall et al. [21] design a Kernel Prediction Network (KPN) to suppress noises in a burst of dark images. On the other hand, some researchers explore Recurrent Neural Networks for extracting temporal information. Chen et al. [22] develop an RNN-based method to utilize temporal video prior for noise reduction. Wang et al. [23] introduce an LSTM-based method, extracting both short and long-term dependencies from image sequences. Godard et al. [4] propose a global recurrent network and append it to existing CNN-based single image denoising networks. Beyond these works on spatio-temporal feature extraction, we explore the possibility of making use of spatial, temporal, as well as spectral corrections in one end-to-end network—CNN for spatial information, LSTM for temporal information, and finally network channel fusion for spectral RGB-NIR information.

2.2 Noise Modeling

The noise in low light photography is complex and closely correlated with camera sensors. A precise description of the sensor noise is crucial for the final denoising performance. In the early stage, most of the denoising methods assumes identically distributed (i.i.d.) additive white Gaussian noise (AWGN) and are of limited performance. Recently, several new noise models are explored, such as mixture of Gaussian (MoG) [24], Poisson-Gaussian [25] and other physical-process-based mixture model [26] [27] [28] [29]. Since high sensitivity cameras are preferred in low light imaging, Wang et al. [23] model CMOS noise with a more complex high sensitivity noise model and propose a noise calibration method assisting generating plausible synthetic noisy images/videos. In this paper, we build an integrated physical-process-based noise model covering rich noise sources, with high performance and good compatibility with different sensors.

2.3 RGB-NIR Imaging

The short-wavelength NIR images are of similar brightness and structures to RGB images. Especially, under low-light conditions NIR images serve good performance [30] and exhibit much more details invisible to eyes [31], so introducing NIR to RGB cameras can be a promising way for low light photography. Fortunately, widely used CCD and CMOS can detect photons in short-wavelength NIR range up to 1200 nm, which largely facilitate RGB-NIR hybrid imaging. Towards this direction, researchers have made some progress. Some groups use NIR images to assist with high-quality RGB imaging. For example, Krishnan and Fergus [5] and Zhuo et al. [32] enhance RGB images captured with ambient light using images captured with NIR flashlight (dark flash), whereas the former captures two images sequentially and the latter uses two synchronized hybrid cameras to take the pair simultaneously. Similarly, Sugimura et al. [33] construct an imaging system to take RGB and NIR photos with different exposure times and recover clean color images from those image pairs. Conversely, Han et al. [30] propose to super-resolve low-resolution NIR images using a high-resolution RGB image. There are also works trying to capture both high-resolution NIR and RGB images computationally, for example, Hu et al. [34] propose RGB-NIR reconstruction techniques in a single sensor with a novel RGB-NIR Color Filter Array (CFA). In spite of the above progress, there is still no working can capture and enhance both RGB and NIR for dark visual recording.

In this paper, we build a compact RGB/NIR imaging system to capture paired videos, which is largely different from [33] that uses two separate cameras. Such a design is of high stability and can be used in the cases requiring lightweight imaging setups. The reconstruction is also different from the above hybrid imaging by enhancing both RGB and NIR videos to maximize the mutual assistance, instead of focusing on only one of them. In addition, a high-quality RGB-NIR dataset is important for data-driven algorithms, but several existing datasets, such as [35] and [36], either consists of only single images (i.e., static scenes) or are constrained to a specific type of scenes. To address these problems, we capture a dataset with paired RGB and NIR nature videos.

2.4 Channel-Fusion Learning

Channel fusion is an effective way of taking multiple input data and learning the latent structure from both intra- and inter- multiple data sources and can incorporate CNN-based deep networks. Specifically, the informative feature extracted by CNN can be fused at the input level, early level, or late level [6] [37]. However, to leverage features at different levels, researchers propose networks with feature fusion across multiple levels [7]. The advantages of such multi-level fusion have been successfully applied in high-level vision tasks such as object or scene change detection [7] and [38], but the studies are still at an early stage in the field of image reconstruction.

Meng et al. [39] use consecutive frames as dominant data channel while guided maps serve as an assistant arm, which guides the network to focus more on certain regions. Like in [7], these two channels are fused at multiple levels to

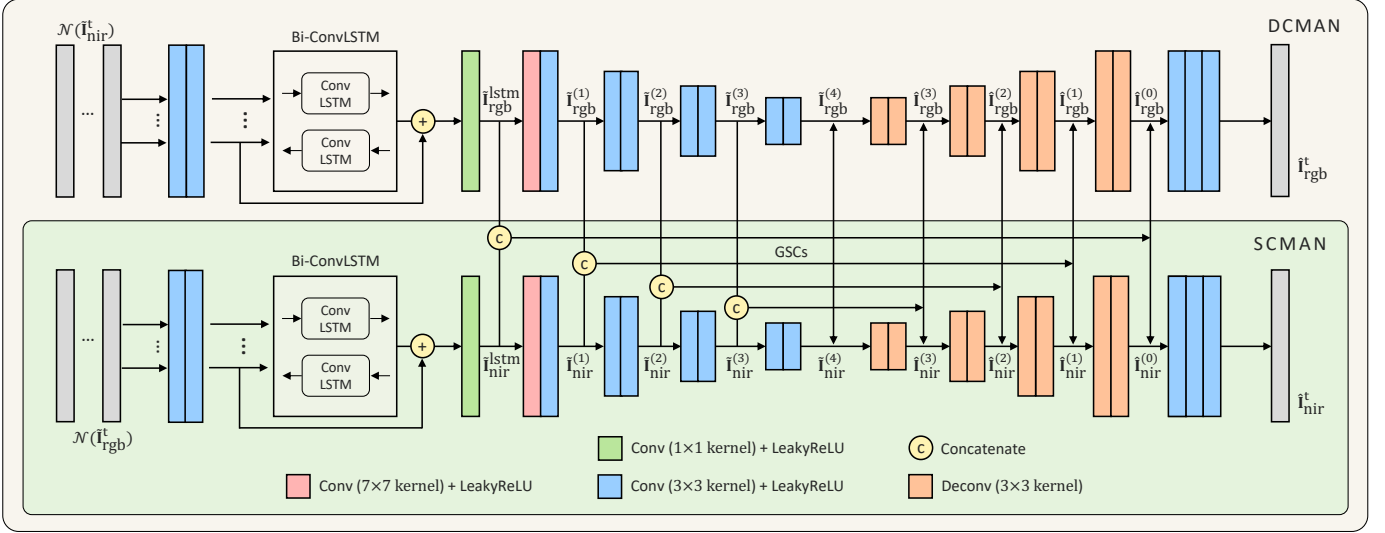


Fig. 3: The structure of our proposed Dual-Channel Multi-Frame Attention Network (DCMAN). Our network includes two channels: each channel consists of a ConvLSTM module and a U-Net-shaped CNN module; two channels are fused with Guided Skip Connections (GSCs). This network explores the spatial prior in nature videos with the U-Net CNN module, temporal prior with the Bi-ConvLSTM module, and spectral prior with inter-channel GSCs.

achieve high performance. Similarly, Zha et al. [40] design a dual-channel multi-scale deep network to fuse the features from multi-spectral bands. Inspired by that work, we can utilize RGB and NIR correlations explicitly. In this paper, the proposed dual-channel multi-frame attention network (DCMAN) is a symmetrical structure, which treats two channels—RGB and NIR—in a parallel manner to mutually enhance each other’s quality, considering the similar photon counts and noise levels at two corresponding input light paths.

3 NETWORK STRUCTURE

We propose a Dual-Channel Multi-frame Attention Network (DCMAN) with two channels taking noisy RGB and NIR inputs: each channel is equipped with a ConvLSTM module and a U-Net-shaped CNN module to explore the spatio-temporal redundancies for noise suppression; two channels are fused with Guided Skip Connections at multiple layers to explore cross-spectrum prior in nature videos.

The training dataset is composed of paired noisy and clean video patches cropped at the same location in consecutive $2T + 1$ frames, each entry can be described as

$$\left\{ \langle \mathbf{I}_{\text{rgb}}^{1 \dots (2T+1)}, \mathbf{I}_{\text{nir}}^{1 \dots (2T+1)} \rangle, \langle \tilde{\mathbf{I}}_{\text{rgb}}^{1 \dots (2T+1)}, \tilde{\mathbf{I}}_{\text{nir}}^{1 \dots (2T+1)} \rangle \right\} \quad (1)$$

Here $\mathbf{I}_{\text{rgb}}^t$ and $\mathbf{I}_{\text{nir}}^t$ denote a clean RGB and corresponding NIR patch captured under good illumination, $\tilde{\mathbf{I}}_{\text{rgb}}^t$ and $\tilde{\mathbf{I}}_{\text{nir}}^t$ denote noisy counterparts generated following the physical based noise model, and t denotes the frame index. We define the neighboring frames of \mathbf{I}^t as

$$\mathcal{N}(\mathbf{I}^t) = \{\mathbf{I}^{t-T}, \mathbf{I}^{t-T+1}, \dots, \mathbf{I}^{t-1}, \mathbf{I}^{t+1}, \dots, \mathbf{I}^{t+T-1}, \mathbf{I}^{t+T}\}. \quad (2)$$

In the next subsections, we describe the processing of t th frame, and omitted the subscript t for simplicity.

3.1 Bi-ConvLSTM Subnet

Utilizing the redundancies among neighboring video frames, the noise deterioration in a video frame can be corrected by neighboring frame patches, including past and future ones. To explore the latent temporal correlation, we first equip the network with two identical bi-directional ConvLSTM [39] [41] modules, which process the input RGB patch sequence $\mathcal{N}(\tilde{\mathbf{I}}_{\text{rgb}})$ and NIR patches $\mathcal{N}(\tilde{\mathbf{I}}_{\text{nir}})$ separately after an affiliated pre-convolutional layers. The residual connections in bi-directional ConvLSTM (bi-ConvLSTM) will guide the network to penalize large differences between adjacent frame patches. Specifically, the bi-ConvLSTM $F_{\text{lstm}}(\cdot)$ can be formulated as follows:

$$\tilde{\mathbf{I}}_{\text{rgb}}^{\text{lstm}} = F_{\text{lstm}} \left(\mathcal{N}(\tilde{\mathbf{I}}_{\text{rgb}}); \theta_{\text{rgb}}^{\text{lstm}} \right) \quad (3)$$

$$\tilde{\mathbf{I}}_{\text{nir}}^{\text{lstm}} = F_{\text{lstm}} \left(\mathcal{N}(\tilde{\mathbf{I}}_{\text{nir}}); \theta_{\text{nir}}^{\text{lstm}} \right), \quad (4)$$

with $\theta_{\text{rgb}}^{\text{lstm}}$ and $\theta_{\text{nir}}^{\text{lstm}}$ being the network parameters of $F_{\text{lstm}}(\cdot)$ to be learned. The receptive field of this module depends on the number of pre-Conv and bi-ConvLSTM layers. Usually the motion between adjacent frames is relatively mild, a shallow bi-ConvLSTM module would be sufficient.

3.2 Dual-Channel Encoder

To further extract the spatio-temporal-spectral information, we design a Dual-Channel subnet based on encoder-decoder networks. The encoder module of each channel has 4 encode layers, each layer takes output of the former layer as input. The input of the first layer $\tilde{\mathbf{I}}_{\text{rgb}}^{(0)}$ and $\tilde{\mathbf{I}}_{\text{nir}}^{(0)}$ are the output of bi-ConvLSTM module, i.e., $\tilde{\mathbf{I}}_{\text{rgb}}^{(0)} = \tilde{\mathbf{I}}_{\text{rgb}}^{\text{lstm}}$ and $\tilde{\mathbf{I}}_{\text{nir}}^{(0)} = \tilde{\mathbf{I}}_{\text{nir}}^{\text{lstm}}$. The encoder can be formulated as

$$\tilde{\mathbf{I}}_{\text{rgb}}^{(k)} = F_{\downarrow}^{(k)} \left(\tilde{\mathbf{I}}_{\text{rgb}}^{(k-1)}; \theta_{r\downarrow}^{(k)} \right) \quad (5)$$

$$\tilde{\mathbf{I}}_{\text{nir}}^{(k)} = F_{\downarrow}^{(k)} \left(\tilde{\mathbf{I}}_{\text{nir}}^{(k-1)}; \theta_{n\downarrow}^{(k)} \right), \quad k \in [1, 4]. \quad (6)$$

In these two equations, $F_{\downarrow}^{(k)}$ is the encoder layer at the k th level with $k \in [1, 4]$, $\theta_{\text{rgb}\downarrow}^{(k)}$ and $\theta_{\text{nir}\downarrow}^{(k)}$ are parameters of subnets in two channels, respectively.

3.3 Guided Skip Connections in Decoder

The RGB and NIR channels guide each other mutually with the help of the Guided Skip Connections (GSCs), an extension to Skip Connections [42]. GSCs in decoder module can not only synthesize high-resolution low-level features, like the skip connections does, but can also combine features from another channel (RGB or NIR). In our GSCs, the features from encoder are concatenated and passed to the 4-layer decoder $F_{\uparrow}^{(k)}$ with $k \in [1, 4]$, specifically defined as

$$\hat{\mathbf{I}}_{\text{rgb}}^{(k-1)} = F_{\uparrow}^{(k)} \left(\left\{ \hat{\mathbf{I}}_{\text{rgb}}^{(k)}, \tilde{\mathbf{I}}_{\text{rgb}}^{(k)}, \tilde{\mathbf{I}}_{\text{nir}}^{(k)} \right\}; \theta_{\text{rgb}\uparrow}^{(k)} \right) \quad (7)$$

$$\hat{\mathbf{I}}_{\text{nir}}^{(k-1)} = F_{\uparrow}^{(k)} \left(\left\{ \hat{\mathbf{I}}_{\text{nir}}^{(k)}, \tilde{\mathbf{I}}_{\text{nir}}^{(k)}, \tilde{\mathbf{I}}_{\text{rgb}}^{(k)} \right\}; \theta_{\text{nir}\uparrow}^{(k)} \right). \quad (8)$$

Here $\hat{\mathbf{I}}_{\text{rgb}}^{(4)} = \tilde{\mathbf{I}}_{\text{rgb}}^{(4)}$, $\hat{\mathbf{I}}_{\text{nir}}^{(4)} = \tilde{\mathbf{I}}_{\text{nir}}^{(4)}$, i.e., output of the encoder module. The final output of our network are

$$\hat{\mathbf{I}}_{\text{rgb}} = F^{(0)} \left(\left\{ \hat{\mathbf{I}}_{\text{rgb}}^{(0)}, \tilde{\mathbf{I}}_{\text{rgb}}^{(0)}, \tilde{\mathbf{I}}_{\text{nir}}^{(0)} \right\}; \theta_{\text{rgb}}^{(0)} \right) \quad (9)$$

$$\hat{\mathbf{I}}_{\text{nir}} = F^{(0)} \left(\left\{ \hat{\mathbf{I}}_{\text{nir}}^{(0)}, \tilde{\mathbf{I}}_{\text{nir}}^{(0)}, \tilde{\mathbf{I}}_{\text{rgb}}^{(0)} \right\}; \theta_{\text{nir}}^{(0)} \right) \quad (10)$$

3.4 Loss Function

The loss function for our network is defined to favor better image sharpness and higher color fidelity jointly. Specifically, the loss function can be written as

$$L = \sum_i \lambda_1 \left(\left\| \hat{\mathbf{I}}_{i,\text{rgb}} - \mathbf{I}_{i,\text{rgb}} \right\|_1 + \left\| \hat{\mathbf{I}}_{i,\text{nir}} - \mathbf{I}_{i,\text{nir}} \right\|_1 \right) + \sum_i \lambda_2 L_{\text{cos}}(\hat{\mathbf{I}}_{i,\text{rgb}}, \mathbf{I}_{i,\text{rgb}}), \quad (11)$$

in which i indexes patches in the training batch, and L_{cos} denotes cosine embedding loss [43]

$$L_{\text{cos}}(\mathbf{x}, \mathbf{y}) = 1 - \cos(\mathbf{x}, \mathbf{y}). \quad (12)$$

4 INTEGRATED CMOS NOISE MODEL

Capturing clean-noisy paired data for model training is time-consuming and inflexible for different cameras, so here we add noise to the high-quality NIR-RGB database to synthesize training data instead. A precise noise modeling is crucial for the final results. Many denoising algorithms assume the noise to be Gaussian or Poisson [3] [44]. However, test results in Fig. 4 show that these models can not describe the true low light noise well. Besides, low-light photos taken with different cameras vary a lot. These facts inspire us to use an integrated physical-process-based noise model being able to incorporate the various noise sources and compatible with different sensors.

A typical CMOS camera sensor takes three steps to convert photon hit on the photosensor to the final digital images: the photosensor detects photons and converts them to electrons, integrated circuits convert and amplify electrons to voltage, and an ADC quantizes analog voltage signals to digital signals. During these stages, various sources of

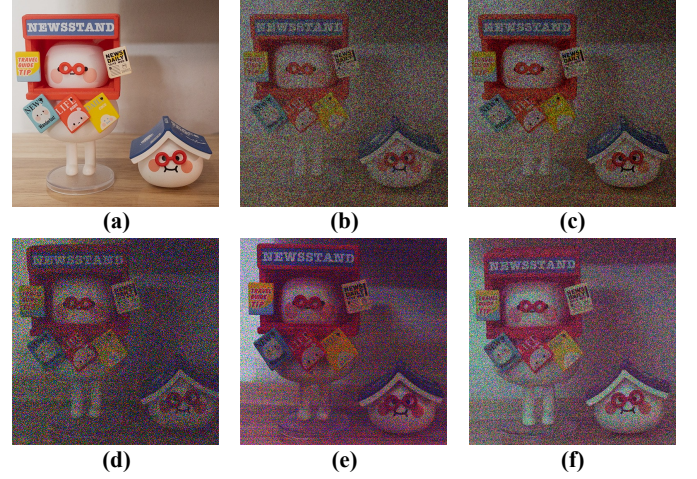


Fig. 4: Comparison among the synthetic noisy images of the scene in (a) following different noise models (b-c) and real noisy images captured by different cameras (d-f). (a) Bright image captured under long exposure. (b)(c) Images superimposed with Gaussian noise and Poisson noise, respectively. (d)(e)(f) Noisy images captured with iPhone 11, Fujifilm X100T and Sony A7R2, respectively.

noise exist. We build an integrated noise model taking all the fundamental ones into account and adopt proper probabilistic distribution for each term to approximate the physical process best.

(i) *Shot noise.* The shot noise is caused by the randomness of photon arrival and can be modeled with Poisson distribution with the mean value being the expected photon number N [45]. The shot noise is determined by the signal itself and the photons at different color channels vary. The shot noise at channel c is

$$n_c^{\text{sh}} \sim P(N_c), c \in \{R, G, B, \text{NIR}\}, \quad (13)$$

with $P(\cdot)$ denoting the Poisson distribution.

(ii) *Dark current.* Dark current is caused by the random generation of electrons and holes, leading to a signal-independent noise [45]. The modeling of this noise is, however, more complex. In some prior works, dark current is assumed to follow a Gaussian or Poisson distribution [23] [45]. However, researchers in [28] observe the long-tail shape of noise data and propose to model this noise by Tukey lambda distribution, which matches the true dark current well. Inspired by this work, we model this noise using a clipped Poisson distribution with better approximation and compatibility with different sensors:

$$n^{\text{dk}} = \max\{0, n_d - N_d\}, n_d \sim P(N_d). \quad (14)$$

Here N_d denotes the expected number of dark current electrons of each pixel.

(iii) *Read noise and Dynamic streak noise.* During the conversion from electrons to voltage signals, read noise appears. This kind of noise does not depend on the signal and can be modeled with a Gaussian distribution and added to the signal directly as

$$n^{\text{rd}} \sim G(0, \sigma_r^2). \quad (15)$$

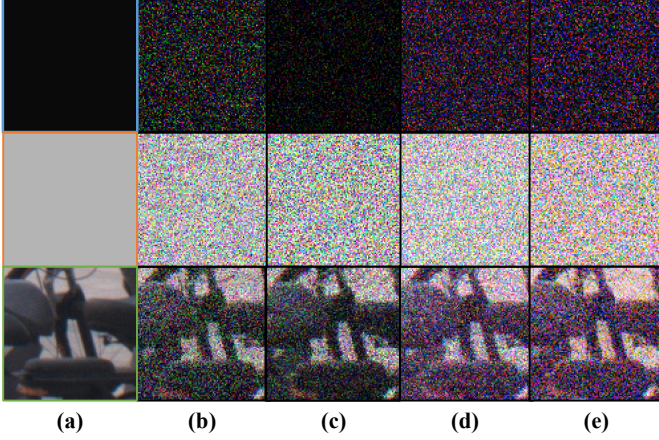


Fig. 5: Comparison of noisy images produced by different noise models with the real images captured under low illumination. (a) Clean image patch for simulation. (b) Results by additive white Gaussian noise. (c) By Poisson noise. (d) By our physical-process-based noise model. (e) Real noisy image patch of the same scene.

In low light imaging, image quality would be further worsened by dynamic streak noise (DSN) [23], which appears to be horizontal streaks in low light noisy photos. This noise heavily depends on the pixel location. Combining shot noise, dark current, read noise and DSN, the equation would be

$$y_{r,c} = \beta_{r,c}(n_c^{\text{sht}} + n^{\text{dk}} + n^{\text{rd}}) \quad (16)$$

where r denotes the row index of pixel and $\beta_{r,c} \sim G(1, \sigma_\beta)$.

(iv) *Quantization.* Under low illumination, one usually set a high information gain to amplify the analog signal before quantization, and magnify further digitally for better visualization and subsequent processing. Mathematically, the voltage signals from camera sensor (with shot noise and dark current) are amplified $K_{a,c}$ times before digitalized, and the quantized signals can be expressed as

$$y_{q,c} = \lfloor K_{a,c} \beta_{r,c} (n_c^{\text{sht}} + n^{\text{dk}} + n^{\text{rd}}) \rfloor, \quad (17)$$

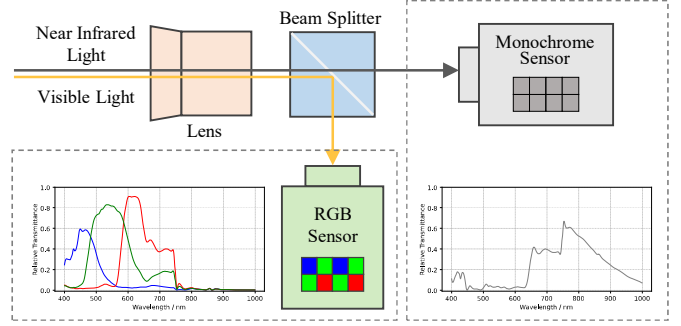
where $\lfloor \cdot \rfloor$ denotes the floor function. Afterwards, the images can be amplified to fit a proper scale, e.g. [0, 255], with digital gain K_d .

In this paper, we integrated all above terms to model the noise based on the physical process of image recording. The model can be formulated as

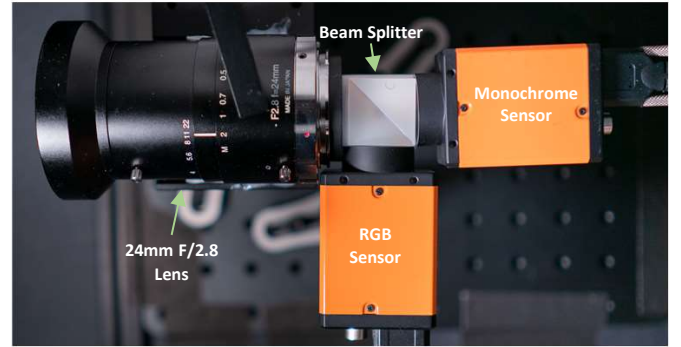
$$y_c = K_d (\lfloor K_{a,c} \beta_{r,c} (n_c^{\text{sht}} + n^{\text{dk}} + n^{\text{rd}}) \rfloor), \quad (18)$$

in which the parameters of noise terms n_c^{sht} , n^{dk} , and n^{rd} are all defined and calibrated in a pixel-wise manner.

To synthesize noisy low light images, we set parameter N_c based on the target flux and clean image pixel and simulate system gain $K = K_a K_d$. Suppose we capture videos in extremely dark environments, and set the highest analog amplification factor $K_{a,c}$ and a digital gain adjust $K_d = \frac{K}{K_a}$ is set. Then the number of photons N_c can be inferred as $N_c = \frac{I_c}{K}$ to maintain the overall brightness. Parameters are calibrated following the procedure described



(a)



(b)

Fig. 6: Our RGB-NIR imaging system. (a) Diagram of our two sensor acquisition system, with the sensor response of the visible band and near-infrared plotted. (b) A photo of our prototype.

in [23], and then randomly fluctuates within a range to improve robustness. The ranges of the key parameters are shown in Tab. 2. Note that image noise intensity is mainly determined by the system gain K . Fig. 5 demonstrates the simulated noisy data and comparison with results by other models. We can see that our noise model provides simulation results much closer to real captured low light video frames, especially in the dark regions.

5 IMAGE ACQUISITION

5.1 The RGB-NIR Imaging System

To capture paired RGB and NIR videos, we build an RGB+NIR camera with two 1" CMOS image sensors (HIKROBOT MV-CH089-UC and MV-CH089-10UM) recording the visible and near-infrared bands respectively. As illustrated in Fig. 6(a), the target scene is firstly captured by the primary lens (Nikon BlueVision BV-L1024, $f=24\text{mm}$, F-Mount), and then the outgoing light is split by a cubic beam splitter which reflects RGB light (stops at roughly 700nm) and transmits NIR light. The spectral transmission curve of the beam splitter and simulated spectral response of red, green, blue, and near-infrared channels are also plotted in Fig. 6(a). Later, two arms arrive at two orthogonally placed sensors: the RGB sensor UC with Bayer CFA captures RGB images, and monochrome sensor UM without a near-infrared cut-off filter captures monochromatic images). These two arms are complementary in collecting photons falling within the broad spectral region (350-1100 nm) of

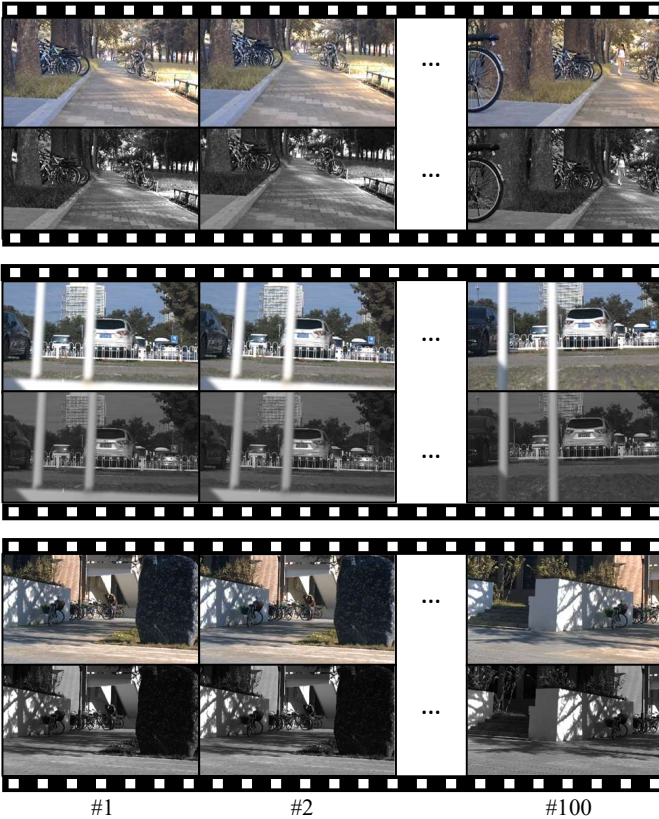


Fig. 7: Several examples from our RGB-NIR video dataset, with pixel-wise registration between measurements in two wavelength bands.

two CMOS sensors. Two cameras are wired to a PC and are software synchronized. After capturing the raw video pairs, we need to align the images from two sensors. Since the RGB and NIR images share the same optical axis, there is no parallax and a precise alignment can be achieved with a simple SIFT detector [46] and affine transformation. The pixel resolution of the aligned video pair is around 1280×720 .

5.2 Data for Model Training and Testing

For simulating the clean-noisy data for supervised model training and quantitative performance evaluation, we collect a dataset including 118 high-quality RGB-NIR video clips with 11,444 frames of real nature scenes under sufficient illuminance. Using a high-end commercial camera as a reference, we adjust the color tone and image contrast by building a Look-Up-Table (LUT) to improve the quality of our dataset. Several examples are shown in Fig. 7. The training data for the network is generated by adding noise to the bright dataset following the aforementioned CMOS noise model. For quantitative evaluation, 104 videos are used for model training and the rest for testing.

To test the performance on real low light videos, we captured noisy input in dark environments with around 0.125%~0.5% illuminance of daily bright photography. These testing data are preprocessed in the same way as the dataset and linearly scaled to normal brightness before being fed into the trained deep network.

6 EXPERIMENTS AND ANALYSIS

6.1 Experiment Settings

The training set for the DCMAN is synthesized by adding noise to our RGB-NIR clean video pairs following the CMOS noise model, with the parameters listed in Table. 2. To raise the robustness towards different noise levels, parameters K_a and K_d are randomly selected for each simulated low light video.

We conduct experiments on a PC with Intel Core CPUs and NVIDIA GeForce GTX 1080Ti GPUs. The DCMAN network is trained with 20,000 paired video patches randomly cropped from original synthesized videos with a patch size of 120×120 pixels. We use the Adam optimizer to train the network. The batch size is 10, the learning rate is 10^{-4} and decreases by a factor of 0.1 every epoch. The color loss parameter λ_2 is 0.1 initially and increases to 0.25 and 0.6 after the 20th and 30th epoch, respectively, and $\lambda_1 = 1 - \lambda_2$. We terminate the training after 40 epochs. The whole training process takes around 40 hours.

As the first to discuss low light video reconstruction with RGB and NIR video pairs, in order to prove our superior performance towards RGB video methods, our method is compared with three state-of-the-art video denoising methods, one filtering based and two data-driven methods: V-BM4D [18], FastDVDnet [3] and TOFlow [44]. For a fair comparison, we conduct the following preprocessing to optimize the performance of three benchmark methods. For V-BM4D, information on noise level is required and we use its in-built noise estimator to estimate the sigma value (standard variation of the Gaussian noise). For FastDVDnet, we remove the noise map and re-train the model on RGB videos in our RGB-NIR dataset to maintain fairness. TOFlow requires the optical flow between adjacent frames, so we use the pre-trained flow estimation network SpyNet [47] and fine-tune it on our dataset with superimposed Gaussian noise, as mentioned in its original paper.

6.2 Experiments on Synthetic Noisy Videos

To quantify our performance, we first apply our methods on the simulated test dataset, using PSNR (Peak signal-to-noise ratio) and SSIM (The Structural Similarity Index) as evaluation metrics. We also compare with three state-of-the-art methods, as shown in Tab. 1, including a patch-based video denoising method V-BM4D [18], a deep-learning-based method FastDVDnet [3], and an optic-flow-based deep learning method TOFlow [44].

The experiment is performed under the setting $K_{RGB} = K_{NIR}$. Other noise parameters are set to match Tab. 2. The results show that our method outperforms other methods by a big gap. At $K = 40$, the PSNR increment goes up to 8.76 dB, 6.20 dB, and 6.06 dB higher than V-BM4D, TOFlow, and FastDVDnet, respectively. The visual comparison in Fig. 8 displays similar advantages. One can see that V-BM4D and TOFlow leave some noise and FastDVDnet tends to produce oversmooth results, while the proposed approach preserves much more details and better color fidelity.

The superior performance benefits from multiple aspects: firstly, more information from the infrared channel is utilized in our methods, while V-BM4D and FastDVDnet only use RGB input. Secondly, 7 adjacent frames are used to

TABLE 1: Comparison of performance in terms of PSNR (dB) / SSIM on our RGB-NIR video dataset. $K = K_a K_d$ denotes the total gain factor during the simulation process. S: Spatial, T: Temporal, C: Spectral. "Ours S" denotes the results from SCMAN_Fn1, with only spatial prior. "Ours S+" denotes the results from SCMAN_Fn7, with explicit exploration of temporal prior. "Ours S+T+C" denotes the results from DCMAN_Fn7 (or DCMAN), with spatio-temporal and spectral prior jointly. "Ours-LSTM" denotes the results by removing bi-ConvLSTM module from DCMAN_Fn7.

Simu.	V-BM4D [18]	TOFlow [44]	FastDVDnet [3]	Ours S	Ours S+T	Ours-LSTM	Ours S+T+C
$K = 10$	27.974 / 0.867	28.976 / 0.914	29.020 / 0.919	29.530 / 0.906	31.062 / 0.935	31.121 / 0.936	31.150 / 0.937
$K = 20$	24.017 / 0.790	25.943 / 0.873	26.012 / 0.883	28.127 / 0.879	29.838 / 0.916	29.863 / 0.918	29.965 / 0.920
$K = 40$	19.667 / 0.659	22.227 / 0.802	22.371 / 0.819	26.442 / 0.841	27.947 / 0.882	28.256 / 0.889	28.427 / 0.892

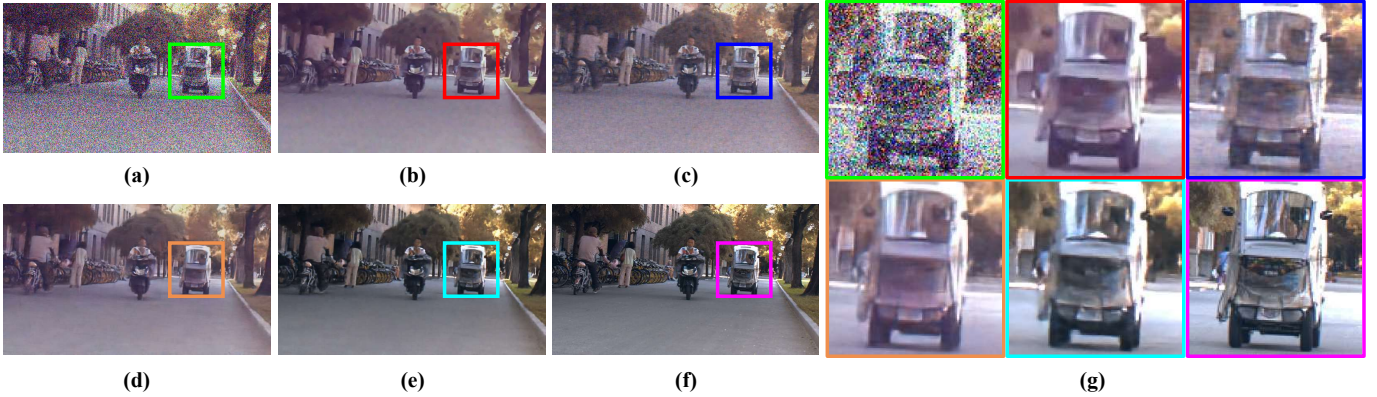


Fig. 8: Simulation test results of our model compared with state of the art methods V-BM4D [18], TOFlow [44], and FastDVDnet [3], and the degenerated single-channel version of our deep network (SCMAN), i.e., model without NIR channel. (a) Simulated noisy frame. (b) V-BM4D. (c) TOFlow. (d) FastDVDnet. (e) Ours. (f) Groundtruth. In (g) we show the comparison in the highlighted region for better visualization, with the same layout as (a)-(f).

TABLE 2: Parameter settings of our CMOS noise model for generating paired training data. Each parameter is randomly selected from a uniform distribution over a range instead of a fixed value, which can help cover the intensity variation among different scenes.

Parameters	Distribution
N_d : Dark current electron number	$U(2, 10)$
σ_β : Standard deviation of DSN	$U(0.02, 0.08)$
σ_r : Standard deviation of read noise	$U(0.5, 2)$
K_{RGB} : Gain in RGB channel	$U(10, 40)$
K_{NIR} : Gain in NIR channel	$U(K_{RGB}, 3K_{RGB})$

denoise the center frame instead of 5 in FastDVDnet. Finally, two previous deep-learning-based methods (TOFlow and FastDVDnet) are retrained on our data but both assume Gaussian noise, while our models are trained with data synthesized following an integrated CMOS noise model and calibrated parameters.

6.3 Ablation Studies

Temporal prior information. In our network, the LSTM module combines information from $F_n = 2T + 1$

frames to denoise the center frame. To evaluate the influence from temporal information utilization, we set F_n to be 1, 3, 5, and 7 respectively to train 4 models DCMAN_Fn1, DCMAN_Fn3, DCMAN_Fn5, and DCMAN_Fn7. Their performances are shown in Tab. 3. Generally, the denoise performance grows as F_n increases. Empirically, in our final network, we set $F_n = 7$, which provides the network with a massive amount of temporal prior assistance. Comparing DCMAN_Fn1 with DCMAN_Fn7 ($K = 40$), i.e., the "Ours S" and "Ours S+T" columns in Tab. 1, introducing temporal information increases PSNR by 1.55 dB.

Spectral prior information. One of the key contributions of our work is the utilization of additional NIR information. To quantitatively evaluate the performance improvement brought by the NIR information, we disable the NIR channel in our network and remove the corresponding input to train a network recovering only RGB video frames. The network is called Single-Channel Multi-frame Attention Network (SCMAN), as shown in the two parallel insets of Fig. 3.

We firstly vary the frame number ($F_n = 1, 3, 5, 7$) to train SCMAN_Fn1, SCMAN_Fn3, SCMAN_Fn5, and SCMAN_Fn7, with the results shown in Tab. 3. We can see that introducing NIR information indeed raises the performance at almost all of the settings. Comparing SCMAN_Fn7 with DCMAN_Fn7 ($K = 40$), we can calculate that spectral information increases PSNR by 0.48 dB on our test dataset, as shown in the "Ours S+T+C" column in Tab. 1. We also

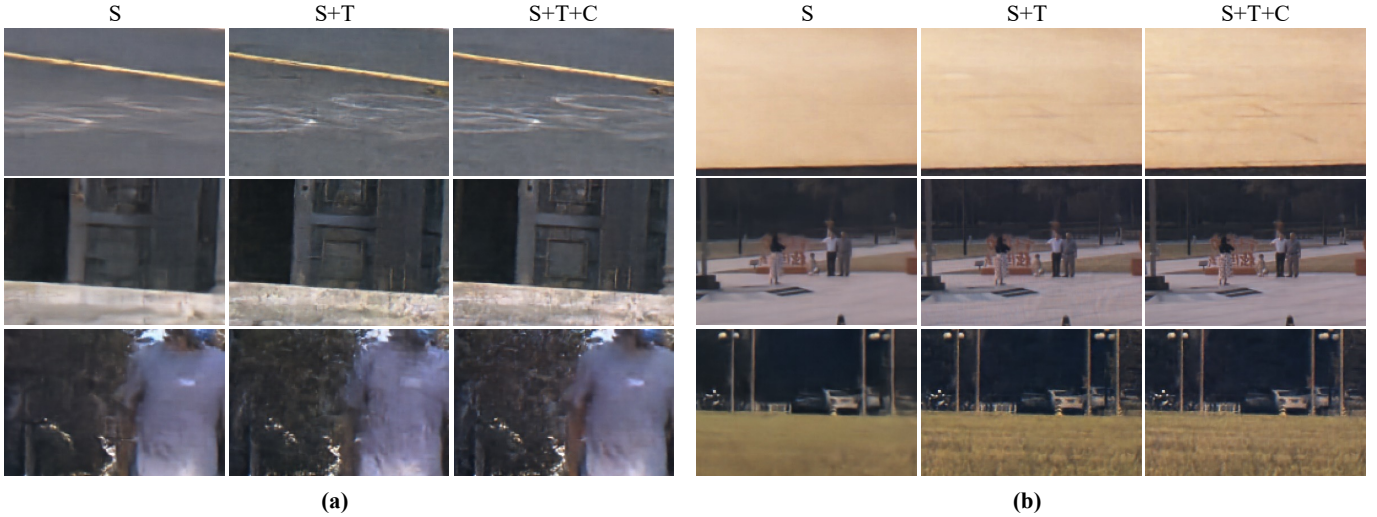


Fig. 9: Comparison of SCMAN_Fn1 (Spatial information), SCMAN_Fn7 (Spatial+Temporal information) and DCMAN_Fn7 (Spatial+Temporal+Spectral information) results to verify the effectiveness of our multi-frame dual-channel network. (a) and (b) show the comparison on synthetic and real data, respectively.

TABLE 3: Performance on simulation test for several F_n and system gain K . Analytical data are shown as $PSNR_{DCMAN}/PSNR_{SCMAN}(dB)$. $PSNR_{DCMAN}$ denotes PSNR of DCMAN (with NIR channel), and $PSNR_{SCMAN}$ denotes that of SCMAN (without NIR channel).

Simulation Test	$F_n = 1$	$F_n = 3$	$F_n = 5$	$F_n = 7$
$K = 10$	29.851 / 29.530	30.826 / 30.603	31.166 / 30.889	31.150 / 31.062
$K = 20$	28.515 / 28.127	29.521 / 29.192	29.898 / 29.533	29.965 / 29.838
$K = 40$	26.875 / 26.442	27.875 / 27.475	28.275 / 27.847	28.427 / 27.947

compare the reconstruction results from DCMAN_Fn7 and SCMAN_Fn7 visually in Fig. 9. The conclusion consists well with the quantitative results, more details and fewer artifacts occur in DCMAN results than in SCMAN.

Conv-LSTM. Bidirectional Conv-LSTM modules are used to encode temporal correlation among multiple frames. However, features from different frames may also be combined directly with a multi-channel CNN, named CNN early fusion. To quantify the contribution from the Conv-LSTM module, we design a new network with CNN early fusion [37]. The result with Conv-LSTM and trivial CNN early fusion are shown in the ‘Ours S+T+C’ and ‘Ours-LSTM’ columns in Tab. 1. It is observed that on our data the Bi-ConvLSTM module increases PSNR by 0.17.

The design of the deep network. The superior performance of our final model can be attributed to both better network design, additional information from the NIR channel, and adopting a physical-process-based noise model. Considering that most state-of-the-art denoising methods are trained with Gaussian or other simple noise models, we conduct an experiment here to verify the advantages of the proposed network itself. Specifically, we train a DCMAN and SCMAN model on additive white Gaussian noise (AWGN) and name it DCMAN_G and SCMAN_G. Compared to state-of-the-art deep learning-based video denoising methods, TOFlow [44] and FastDVDnet [3], retrained

with synthetic data by adding AWGN on our dataset, both our single-channel and bi-channel networks perform much better, as shown in Tab. 4.

TABLE 4: Test results on image synthesized with Additive White Gaussian Noise (AWGN) compared with two deep learning based method TOFlow and FastDVDnet. DCMAN_G and SCMAN_G are our methods with and without NIR channel trained on simulated AWGN noise. Our model outperform those methods on both noise levels.

PSNR(dB) / SSIM	$\sigma = 20$	$\sigma = 40$
TOFlow	31.792 / 0.941	29.328 / 0.902
FastDVDnet	32.509 / 0.947	29.605 / 0.911
SCMAN_G	32.833 / 0.952	30.259 / 0.923
DCMAN_G	32.988 / 0.954	30.418 / 0.926

Integrated noise model. Our training data are synthesized by a physical-process-based integrated noise model. To validate the necessity of using such a noise model incorporating various noise sources, we compare the performance of DCMAN_G (Gaussian noise) and DCMAN (calibrated noise following the integrated noise) on real low light images. The results are shown in Fig. 10, from which one can see noticeable performance improvements brought by learning

from data generated by an integrated noise model with calibrated model parameters.

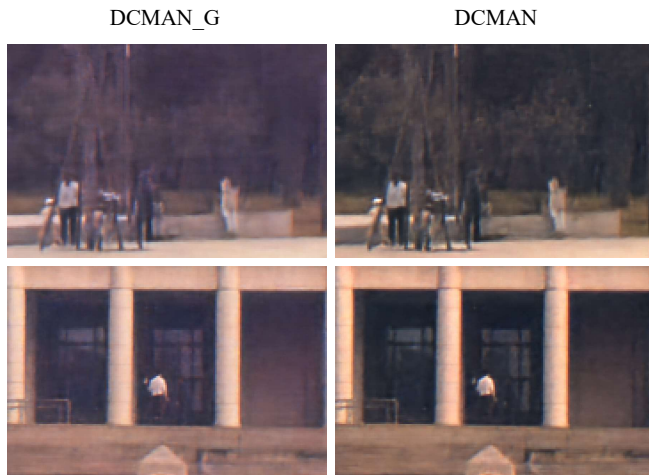


Fig. 10: Comparison of results by DCMAN_G (trained from data with Gaussian noise) and DCMAN (trained with synthetic data following physical CMOS noise model) on real low light videos.

6.4 Experiments on Real Captured Low Light Videos

We test our approach on real low-light videos captured with our dual sensor camera, the results are shown in Fig. 11-12, and please refer to supplementary data for video demonstration. The images are captured at dark hours with illumination of less than 1 lux. From the image in Fig. 11, one can see that under such low illuminations the input raw images are severely deteriorated by noise, with structure details washed out and color deviating largely. After reconstruction, we can achieve striking quality improvement. The flat regions (e.g., road, sky), structures (e.g., road lamp, driving line, car, traffic sign), and textured regions (e.g., trees, shrubbery) are all recovered with high quality. The high performance indicates that our approach is quite promising in assisting night traffic surveillance and auto driving.

Comparatively, three previously published state-of-the-arts show limited performance in such challenging scenarios, as shown in Fig. 12. The results from V-BM4D and TOfFlow tend to suffer from residual noise, while FastDVDnet causes over-smoothness and cannot recover thin structures. Besides, all these three methods have apparent color distortion, while our DCMAN removes purplish color bias and preserves image details well. For example, in the 1st row, we can notice that the details in the dark background are recovered, such as the tree trunks and the shrubs. In the 2nd and 4th scenes, the details are reconstructed at a much higher quality, e.g., the textures on the marble, and the bike on the side of the road. Besides, the proposed approach produces high performance on the dense striped patterns (the railing in the 5th row) that are prone to oversmoothness and the highlight region (car wheel in the 6th row). Our results are also advantageous in the texture regions, with more details and less color distortion, such as the lawn regions in the 3rd scene. Overall, the proposed method

is of stronger noise suppression, better structure preservation, and higher color fidelity. The superior performance is mainly attributed to the additional information from the NIR channel and successful exploration of the cross-channel and multi-frame prior, including both the network design and effective training strategy.

7 SUMMARY AND DISCUSSIONS

This paper reports a dual-channel computational dark videography approach with superior performance than state-of-the-arts, by collecting more photons optically and introducing cross channel priors computationally. The high performance benefits from technical contributions in three folds: (i) We design a compact RGB+NIR dual-sensor camera to largely increase the collected photons of a conventional RGB camera by additionally capturing a NIR wavelength band, and two paths are with pixel-wise registration and high precision synchronization. (ii) We then propose a Dual-Channel Multi-frame Attention Network (DCMAN) to enhance the RGB and NIR video pair in an end-to-end manner by exploring the spatial, temporal, and spectral information jointly. (iii) The model can be learned effectively from synthetic data produced by superimposing noise on our high quality NIR+RGB dataset following an integrated physical-process-based noise model and adapted to different sensors.

So far, the application of our approach is somewhat limited due to its high computing complexity. For a 360×640 video frame, we need around 700ms to recover the high-quality video with an NVIDIA GeForce GTX 1080Ti GPU. A low weight model is under development currently. In the hardware, one can replace the Bayer pattern filter with a customized one with NIR transmittance to collect more photons in dark environments, and developing corresponding reconstruction algorithms is a worth studying topic. The approach can also be extended further to other wavelength bands for different imaging scenarios or tasks, such as UV, mid-NIR, etc.

ACKNOWLEDGMENTS

Our code will be soon available at <https://github.com/jarrycyx/dual-channel-low-light-video.git>

REFERENCES

- [1] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9446–9454.
- [2] M. Claus and J. van Gemert, "Videnn: Deep blind video denoising," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.
- [3] M. Tassano, J. Delon, and T. Veit, "Fastdvdnet: Towards real-time deep video denoising without flow estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1354–1363.
- [4] C. Godard, K. Matzen, and M. Uyttendaele, "Deep burst denoising," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 538–554.
- [5] D. Krishnan and R. Fergus, "Dark flash photography," *ACM Trans. Graph.*, vol. 28, no. 3, p. 96, 2009.
- [6] J. Jiang, X. Feng, F. Liu, Y. Xu, and H. Huang, "Multi-spectral RGB-NIR image classification using double-channel CNN," *IEEE Access*, vol. 7, pp. 20 607–20 613, 2019.



Fig. 11: Our RGB and NIR reconstruction results on a dark video frame captured by our setup at night.

- [7] P. Zhang, W. Liu, Y. Lei, and H. Lu, "Hyperfusion-Net: Hyperdensely reflective feature fusion for salient object detection," *Pattern Recognition*, vol. 93, pp. 521–533, 2019.
- [8] Y. Zhang, X. Guo, J. Ma, W. Liu, and J. Zhang, "Beyond brightening low-light images," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 1013–1037, Apr. 2021.
- [9] C. Tian, Y. Xu, L. Fei, and K. Yan, "Deep learning for image denoising: A survey," *arXiv:1810.05052 [cs]*, Oct. 2018.
- [10] R. S.R and N. P. Kavya, "Noise reduction in video sequences the state of art and the technique for motion detection," *International Journal of Computer Applications*, vol. 58, no. 8, pp. 31–36, Nov. 2012.
- [11] F. Jin, P. Fieguth, and L. Winger, "Wavelet video denoising with regularized multiresolution motion estimation," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, pp. 1–11, 2006.
- [12] V. Zlokolica, A. Pizurica, and W. Philips, "Wavelet-domain video denoising based on reliability measures," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 8, pp. 993–1007, 2006.
- [13] J. C. Brailean, R. P. Kleihorst, S. Efstratiadis, A. K. Katsaggelos, and R. L. Legendijk, "Noise reduction filters for dynamic image sequences: A review," *Proceedings of the IEEE*, vol. 83, no. 9, pp. 1272–1292, 1995.
- [14] A. Buades, J.-L. Lisani, and M. Miladinović, "Patch-based video denoising with optical flow estimation," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2573–2586, 2016.
- [15] A. Buades, B. Coll, and J.-M. Morel, "Denoising image sequences does not require motion estimation," in *IEEE Conference on Advanced Video and Signal Based Surveillance*. IEEE, 2005, pp. 70–74.
- [16] K. Dabov, A. Foi, and K. Egiazarian, "Video denoising by sparse 3D transform-domain collaborative filtering," in *European Signal Processing Conference*, 3, pp. 145–149.
- [17] M. Lebrun, "An analysis and implementation of the BM3D image denoising method," *Image Processing On Line*, vol. 2012, pp. 175–213, 2012.
- [18] M. Maggioni, G. Boracchi, A. Foi, and K. Egiazarian, "Video denoising, deblocking, and enhancement through separable 4-D nonlocal spatiotemporal transforms," *IEEE Trans Image Process*, vol. 21, no. 9, pp. 3952–66, Sep. 2012.
- [19] P. Arias and J.-M. Morel, "Video denoising via empirical Bayesian estimation of space-time patches," *Journal of Mathematical Imaging and Vision*, vol. 60, no. 1, pp. 70–93, 2018.
- [20] M. Tassano, J. Delon, and T. Veit, "Dvdnet: A fast network for deep video denoising," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1805–1809.
- [21] B. Mildenhall, J. T. Barron, J. Chen, D. Sharlet, R. Ng, and R. Carroll, "Burst denoising with kernel prediction networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2502–2510.
- [22] X. Chen, L. Song, and X. Yang, "Deep rnns for video denoising," in *Applications of Digital Image Processing XXXIX*, vol. 9971. Inter-

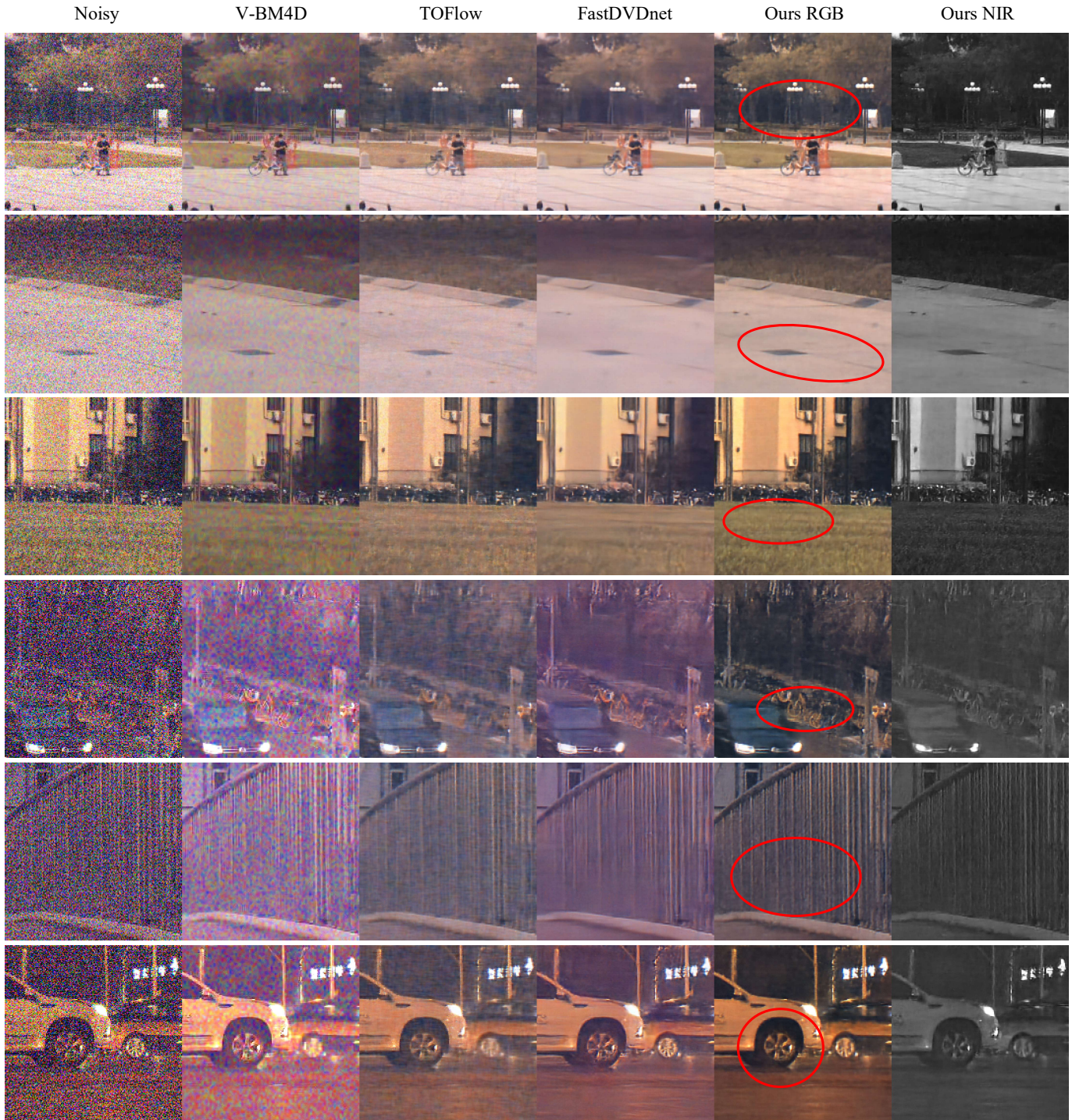


Fig. 12: Our RGB reconstruction results on real low light video frames and comparison with state-of-the-arts. Here some regions with large performance diversity are highlighted with red ellipses.

national Society for Optics and Photonics, 2016, p. 99711T.

- [23] W. Wang, X. Chen, C. Yang, X. Li, X. Hu, and T. Yue, "Enhancing low light videos by exploring high sensitivity camera noise," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4111–4119.
- [24] F. Zhu, G. Chen, and P.-A. Heng, "From noise modeling to blind image denoising," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 420–429.
- [25] M. Mäkitalo and A. Foi, "Noise parameter mismatch in variance stabilization, with an application to poisson–gaussian noise estimation," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5348–5359, 2014.
- [26] G. E. Healey and R. Kondepudy, "Radiometric CCD camera calibration and noise estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 3, pp. 267–276, 1994.
- [27] Y. Tsin, V. Ramesh, and T. Kanade, "Statistical calibration of CCD imaging process," in *Proceedings Eighth IEEE International Conference on Computer Vision*, vol. 1. IEEE, 2001, pp. 480–487.
- [28] K. Wei, Y. Fu, J. Yang, and H. 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*, 2020, pp. 2758–2767.

- [29] H. Yue, C. Cao, L. Liao, R. Chu, and J. 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*, 2020, pp. 2301–2310.
- [30] T. Y. Han, D. H. Kim, S. H. Lee, and B. C. Song, "Infrared image super-resolution using auxiliary convolutional neural network and visible image under low-light conditions," *Journal of Visual Communication and Image Representation*, vol. 51, pp. 191–200, 2018.
- [31] Z. Chen, X. Wang, and R. Liang, "RGB-NIR multispectral camera," *Optics Express*, vol. 22, no. 5, pp. 4985–4994, 2014.
- [32] S. Zhuo, X. Zhang, X. Miao, and T. Sim, "Enhancing low light images using near infrared flash images," in *IEEE International Conference on Image Processing*. IEEE, 2010, pp. 2537–2540.
- [33] D. Sugimura, T. Mikami, H. Yamashita, and T. Hamamoto, "Enhancing color images of extremely low light scenes based on RGB/NIR images acquisition with different exposure times," *IEEE Trans Image Process*, vol. 24, no. 11, pp. 3586–97, Nov. 2015.
- [34] X. Hu, F. Heide, Q. Dai, and G. Wetzstein, "Convolutional sparse coding for RGB+NIR imaging," *IEEE Trans Image Process*, vol. 27, no. 4, pp. 1611–1625, Apr. 2018.
- [35] X. Niu, H. Han, S. Shan, and X. Chen, "VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 562–576.
- [36] M. Brown and S. Süssstrunk, "Multi-spectral SIFT for scene category recognition," in *CVPR*. IEEE, 2011, pp. 177–184.
- [37] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi, "Real-time video super-resolution with spatio-temporal networks and motion compensation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4778–4787.
- [38] Y. Lei, D. Peng, P. Zhang, Q. Ke, and H. Li, "Hierarchical paired channel fusion network for street scene change detection," *IEEE Trans Image Process*, vol. 30, pp. 55–67, 2021.
- [39] X. Meng, X. Deng, S. Zhu, X. Zhang, and B. Zeng, "A robust quality enhancement method based on joint spatial-temporal priors for video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 6, pp. 2401–2414, Jun. 2021.
- [40] S. Zha, W. Jin, C. He, Z. Chen, G. Si, and Z. Jin, "Detecting of Overshooting Cloud Tops via Himawari-8 Imagery Using Dual Channel Multiscale Deep Network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1654–1664, 2020.
- [41] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Advances in Neural Information Processing Systems*, 2015, pp. 802–810.
- [42] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.
- [43] T. Wilkinson and A. Brun, "Semantic and verbatim word spotting using deep neural networks," in *International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2016, pp. 307–312.
- [44] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, vol. 127, no. 8, pp. 1106–1125, 2019.
- [45] M. Konnik and J. Welsh, "High-level numerical simulations of noise in CCD and CMOS photosensors: Review and tutorial," *arXiv preprint arXiv:1412.4031*, 2014.
- [46] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2. IEEE, 1999, pp. 1150–1157.
- [47] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4161–4170.



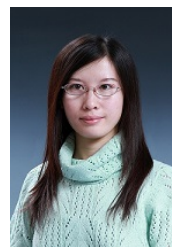
Yuxiao Cheng is an undergraduate student in the Department of Automation, Tsinghua University, Beijing, China. His research interests include computational imaging and computer vision.



Runzhao Yang received his B.E. degree in the School of Electrical Engineering and Automation from Wuhan University, Wuhan, China, in 2020. He is pursuing his Ph.D. degree in the Department of Automation at Tsinghua University, Beijing, China. His research interests include computer vision, data compression, and machine learning.



Zhihong Zhang received the B.Eng. degree from the School of Electronics Engineering, Xi'an University, Xi'an, China, in 2019. He is currently pursuing the Ph.D. degree in the Department of Automation, Tsinghua University. His research interests include computational imaging, computer vision, and machine learning.



Jinli Suo received the BS degree in computer science from Shandong University, Shandong, China, in 2004 and the Ph.D. degree from the Graduate University of Chinese Academy of Sciences, Beijing, China, in 2010. She is currently an associate professor with the Department of Automation, Tsinghua University, Beijing, China. Her research interests include computer vision, computational photography, and statistical learning. She is an Associate Editor for the *IEEE Transactions on Computational Imaging*.



Qionghai Dai received the M.S. and Ph.D. degrees in computer science and automation from Northeastern University, Shenyang, China, in 1994 and 1996, respectively. He is currently a Professor with the Department of Automation, an Adjunct Professor with the School of Life Sciences, Tsinghua University, and an Academician with the Chinese Academy of Engineering. He has authored or coauthored more than 200 conference and journal papers and two books. His research interests include computational photography and microscopy, computer vision and graphics, and intelligent signal processing. He is an Associate Editor for the *Journal of Visual Communication and Image Representation*, the *IEEE Transactions on Neural Networks and Learning Systems*, and the *IEEE Transactions on Image Processing*.