Learning to See Moving Objects in the Dark

Haiyang Jiang
University of Southern California
California, USA
haiyangj@usc.edu

Yinqiang Zheng*
National Institue of Informatics
Tokyo, Japan
yqzheng@nii.ac.jp

Abstract

Video surveillance systems have wide range of utilities, yet easily suffer from great quality degeneration under dim light circumstances. Industrial solutions mainly use extra near-infrared illuminations, even though it doesn't preserve color and texture information. A variety of researches enhanced low-light videos shot by visible light cameras, while they either relied on task specific preconditions or trained with synthetic datasets. We propose a novel optical system to capture bright and dark videos of the exact same scenes, generating training and ground truth pairs for authentic low-light video dataset. A fully convolutional network with 3D and 2D miscellaneous operations is utilized to learn an enhancement mapping with proper spatial-temporal transformation from raw camera sensor data to bright RGB videos. Experiments show promising results by our method, and it outperforms state-of-the-art low-light image/video enhancement algorithms.

1. Introduction

Video surveillance systems have been vastly used throughout industry, military, and academia. Nonetheless, they commonly encounter situations with extremely low level of illumination, *e.g*. security cameras at night [41], and long time continuous wild animal monitoring for research purposes [5]. Under these circumstances, in order not to expose recording devices and not to interfere with photographed objects, extra visible-light source is usually not a viable option.

Current solutions mainly involve near-infrared (NIR) LED [42] or diodes [11] as shown in Figure 1 (a). Infrared illumination helps gain better vision in low-light environments, in the mean time, however, introducing several additional drawbacks compared to natural light camera systems. Inevitable are additional energy consumption and heat generation with the presence of extra light sources,

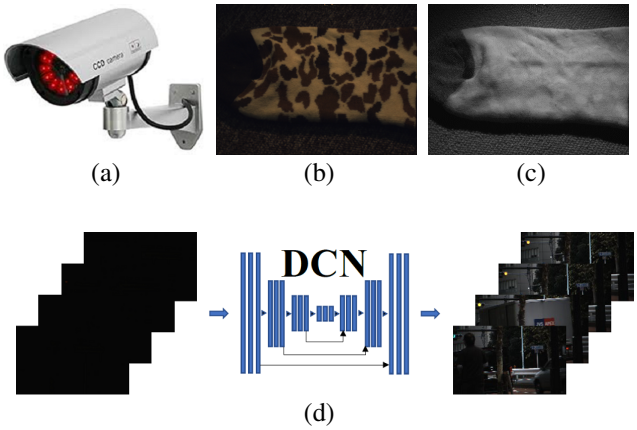


Figure 1. Comparison between widely used industrial systems for low-light situations and our proposed solution. (a) A surveillance camera equipped with NIR diodes; (b) Texture on the sleeve is visible in the RGB image; (c) Texture information disappears in the NIR image; (d) Our low-light video enhancement solution.

which can increase operation and maintenance costs of a system. More importantly, visible color and even texture information, which can be crucial in some cases *e.g*. patterns on clothes of a suspect or physical properties of an animal's coat, suffer from extensive loss [25], as demonstrated by a typical example in (b) and (c) of Figure 1. Besides, even with NIR LED sources, there is still risk of becoming invasive recording, especially when target creatures have different spectra of visible light from human being. In such situations, infrared radiation has potential to trigger uncontrollable animal reactions.

Based on these reasons, a more reasonable choice is to directly enhance videos captured by ordinary camera systems. Researchers have proposed a variety of approaches to enhance low light images and videos, including contrast enhancement [26, 1, 2, 20, 34], denoising [28, 15, 43], retinex algorithms [37, 21, 22] and so on. Among all those re-

*Corresponding Author

2380-7504/19/$31.00 ©2019 IEEE
DOI 10.1109/ICCV.2019.00742

searches, we found two closest related researches handling insufficient light problem [3, 24]. These two researches both adopt deep convolutional network (DCN) approach to learn an enhancement mapping. In order to train and test the networks, low-light and well-lighted image/video pairs are generated as basic elements in a dataset. The state-of-the-art low-light image enhancement technique, proposed by Chen *et al*. [3], trained a DCN on a dataset called See-in-the-Dark (SID) that was built by taking pictures of identical scenes with different exposure time. Their network can brighten extreme dark images with amplification ratio being as high as 300. This long-exposure-time approach of building a dataset, nevertheless, limits its scope to static settings. Lv *et al*. [24] proposed a multi-branch network (MBLLEN) and trained it on synthetic dataset. Both of these training datasets lose details of real world low light videos. They either omit dynamic temporal information in a video, or distort video information by artificial transformation.

In this paper, we propose a novel optical system to shoot bright videos, as ground truth, and dark videos, as training input, of the exact same scenes simultaneously. In this manner, we introduce a dataset of 179 pairs of videos consisting of 35800 extremely low-light raw images and their corresponding well-lighted RGB counterparts. This enabled us to train on them a modified U-Net [29] with miscellaneous 2D and 3D operations capable of manipulating temporal details embedded in a video better than usual 2D networks. Our experiments show that the newly captured dataset with this network structure outperforms state-of-the-art low-light image/video enhancement techniques by a large margin both in our test cases and real world tasks.

The major contributions of this paper can be summarized as

1. We build a novel co-axis optical system to capture temporally synchronized and spatially aligned low-light and well-lighted video pairs.

2. We collect the first low-light/well-lighted video pair dataset of street views with moving vehicles and pedestrians, which will be publicly released to facilitate further researches.

3. We propose a 3D U-Net based network for low-light video enhancement, which reveals superior performance in color fidelity and temporal consistency.

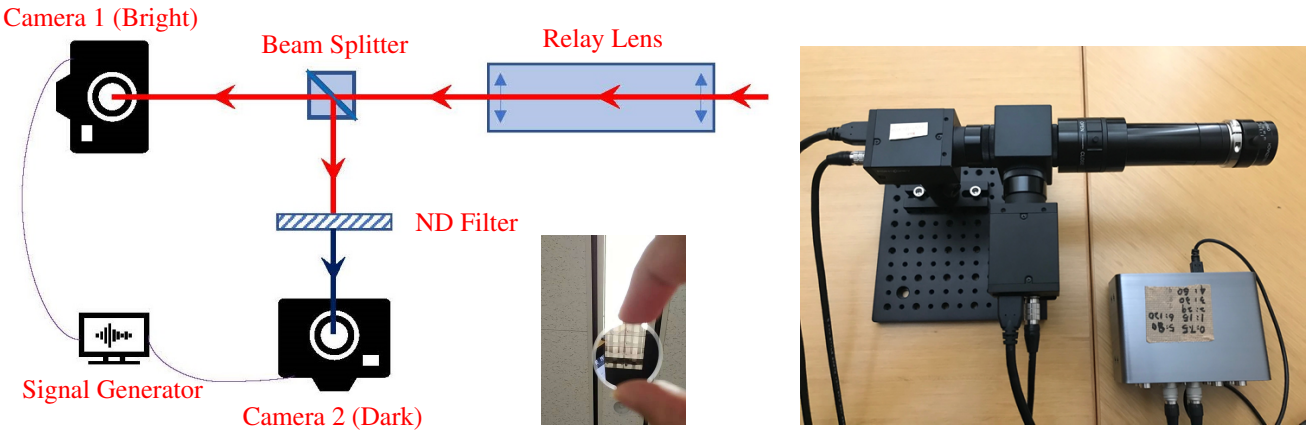
2. Related Work

Low-light video enhancement spans a variety of research fields, all of which have tremendous amount of studied literature. This section provides a quick review on related approaches.

Low-light image enhancement. Methods for dark image enhancement can be applied to dark videos in a frame-by-frame pattern. There are a vast number of conventional enhancement approaches derived from histogram equalization. Arici *et al*. [1] proposed an algorithm to exploit conditional histogram information and enhance image contrast while preserving naturalness. CVC [2] uses 2D histogram to take interpixel contextual information into account. LDR [20] solves optimization problems at different layer of a 2D histogram to generate unified transformation function. DHE [26] captures edge information of input images to enhance contrast.

Retinex theory [18] is another remarkable foundation in poorly-lighted image enhancement researches. Jobson *et al*. [14, 13] based on retinex theory discovered best placement of log function and Gaussian form, and developed a multiscale algorithm along with a color restoration method to deal with cases where gray-world assumptions are violated. LIME [10] estimates a pixel-wise illumination map, and refines it by a structure prior to achieve enhancement. Fu *et al*. [7] proposed a fusion-based enhancing method that estimates illumination and fuses derived images by a multiscale strategy. Fu *et al*. [8] pointed out disadvantages of using log transformation and proposed a weighted variational model to estimate reflectance and illumination from an image. AMSR (adaptive multiscale retinex) [21] infers weights for multiple SSR (single-scale retinex) to produce naturally enhanced images. Liu *et al*. [22] applied Gaussian filters to get illumination and reflection components, with Gamma correction to enhance saturation in HSI color space. NPE [35] tries to overcome naturalness preserving problem of retinex-based algorithms by balancing between image details and natural view. There are also researches that define novel models similar to retinex theory or other optical systems. A dual-exposure fusion algorithm, proposed by Ying *et al*. [39], imitates human visual system to enhance image contrast. Ying *et al*. [40] established a camera response model and used that model to adjust illumination of each pixel in an image. Generally, low-light image enhancement methods requires task specific adjustment to reinforce model hypotheses' validity.

Deep learning based methods. Data driven models have been demonstrating powers in the field of image and video enhancement. CNN is one of the most common models. Remez *et al*. [28] illustrated that class information obtained by a DCN helps image denoising. DnCNN [43] uses residual learning and batch normalization to speed up and improve performances. CNNs are also directly applied to low-light image enhancement tasks, *e.g*. DCNN [32] and LLCNN [33]. More recently, GANs [9] have been drawing increasing attention. They are capable of generating generic image transformation, *e.g*. conditional GAN [12] and cycle-GAN [44], and super-resolution tasks, *e.g*. SRGAN [19]. Other deep learning based methods include LLNET [23], which uses stacked-sparse denoising autoencoder to enhance and denoise low-light images, and deep Retinex-Net [37] that combines retinex theory and deep learning.



(a) (b) (c)

Figure 2. The camera system we built to capture bright and dark videos of the same scenes concurrently. (a) Optical paths of the system; (b) a demonstration of an ND filter's effect; (c) a photo of the installed system.

Low light video enhancement and training dataset. A commonly shared theoretical model has not been put forward in the field of low-light video enhancement, except for methods that borrow aforementioned image enhancement models to process videos frame-by-frame or those built upon retinex theory *e.g.* a piecewise-based framework [34]. There are methods that adopt similar approaches of dehazing inverted frames [6, 27]. Kim *et al.* [15] proposed a method that separates temporal and spatial filtering to denoise extremely low-light videos, and uses gamma correction histogram adjustment with clipping thresholds to increase dynamic range. Ko *et al.* [17] used similar patches among adjacent frames to accumulate extra information. They also used a guide map to better preserve bright regions and avoid color distortion, ghosting effect, or flickering. MBLLVEN [24] is the generalized 3D version of MBLLEN. It incorporate temporal information into their network process. This network is trained on synthetic dataset generated by uniformly randomized gamma correction and Poisson noise [23]. There are authentic low-light image training sets, *e.g.* SID [3] and Low-Light dataset (LOL) [37]. However, to the best of our knowledge, public authentic dark video training dataset does not exist yet. Therefore, we collected sufficient number of video pairs to establish a systematic low-light video enhancement pipeline.

3. See-Moving-Objects-in-the-Dark Dataset

3.1. Camera System

In order to overcome blurry effects introduced by long-exposure approach, we constructed a camera system (as shown in figure 2 (c)) to capture low light videos.

This system is designed to work under sufficiently illuminated circumstances so that generated bright videos (ground-truth) could have acceptable quality. Ambient light goes through a relay lens in order to adjust focal lengths and light beam directions. A beam splitter is then set up to divide input light and feed identical brightness information to two cameras separately. One of the cameras is equipped with a neutral density (ND) filter so that weakly brightened videos could be produced. An ND filter is able to diminish light intensity throughout all wavelengths without modifying hue of color rendition. In our settings, we chose an ND filter with 1% ratio between transmitted through optical power and incident light intensity. Its effect can be seen in Figure 2 (b), where halation of overexposed light tubes becomes visible shapes after passing through an ND filter.

Two cameras are perfectly synchronized by a signal generator to take raw pictures at video frame rate. Specifically, we set it to be 15, 30, 60, and 120 frames per second. With these rates, we are able to gather thousands of sequential images within a few minutes. Furthermore, our system's mobility allows us to take pictures in multiple places effortlessly. This gave us an advantage of creating our whole dataset within 12 hours, indicating large potentials in data augmentation for future applications.

Cameras' ADC modules generate image pixel values with 12-bit depth, and save them in 16 bits. Table 1 gives more detailed information on equipment we selected.

Components	Specifications
Camera	Model: FLIR GS3-U3-23S6C
	Shutting: Video Mode 7
	Filter: Bayer pattern
ND Filter	Model: Thorlabs Reflective ND Filter ND20B

Table 1. Camera system specifications.

Although the whole optical system is mounted on a sturdy base, with standard screw interfaces and C-mount tube systems and two cameras are aiming at matched sights,

Figure 3. A subset of training dataset. From top to bottom, each row represents a distinct gain level of 0, 5, 10, 15, 20 in ADC module. Ground truth video frames are in the front, for training input video frames are basically black.

there are still pixel-level misalignments, both translational and rotational. We addressed this issue by removing the ND filter first and taking 50 pairs of bright pictures with two installed cameras respectively. We then applied homography feature matching operations on these clear image pairs to acquire optimal geometric mappings in which differences after alignment was the minimum between two camera's visual fields. All bright videos were transformed accordingly to enforce pixel-wise rigorous correspondence between video pairs. In each pair, bright and dark videos were cropped to their overlapping rectangle area (around 1800 pixels $\times$ 1000 pixels) after geometric mapping. We compelled the height and width of the area to be even numbers for the purpose of a following demosaicing operation.

3.2. Dataset

With the constructed optical system and alignment preprocessing, we collected 179 video pairs of street views with moving vehicles and pedestrians under different conditions. We referred to this dataset as See-Moving-Objects-in-the-Dark (SMOID) dataset. Five various ADC gain levels, ranging from 0 to 20 at an interval of 5, were utilized to maintain a diverse dataset. Each video is of 200 frames. All video frames are in a 16-bit unsigned integer format organized in a "GBRG" (left to right, top to bottom, in each 2x2 cells starting from top left) Bayer pattern. Bright videos had gone through a demosaicing procedure to form ground truth videos in a normal 8-bit sRGB form. Specific converters employed for ground truth videos' demosaicing were adapted depending on parities of starting coordinates when we cropped videos, *e.g*. an even starting row number and even starting column number case would require a GR2RGB converter, while an odd starting row number and even starting column number case demanded a BG2RGB converter, etc.

Among all the pairs, 125 of them are chosen as training samples, 27 as validation set, and the rest 27 as test set.

Figure 3 displays a subset of SMOID dataset. Five rows from top to bottom correspond to videos with different ADC gain levels in an increasing order. Each row contains extracted key frames from ground truth and training input videos. Because of the fact that dark images are mainly all black, bright images are displayed in the foreground. In spite of the fact that key frames presented here are all street views, actual training videos incorporate a vast number of features that are adequate for our network to grasp a generic enhancement mapping for various dark scenes. Detailed experiment results in section 5 corroborates this conclusion.

4. Method

4.1. Pipeline

An overview of our pipeline is shown in Figure 4. Our pipeline reconstructs RGB formatted bright videos from raw sensor data. This is because of the fact that extreme black situations results in extraordinarily low signal-to-noise rate (SNR) images, in which case RGB 8-bit format fails to keep fidelity to the real world information. In our SMOID dataset, raw images were obtained by cameras with Bayer filters. Thus, before feeding videos into the network, we packed each frame into 4 channels in a "GRBG" order, which reduced videos' spatial sizes by a factor of two. Resolution restoration was taken care by inserting a sub-pixel depth-to-space layer [31] applied to each frame at the end of our network.

Given the fact that raw values can fluctuate in much wider range than common RGB values, particularly in high ADC gain levels, we adjusted entire video clip values by a

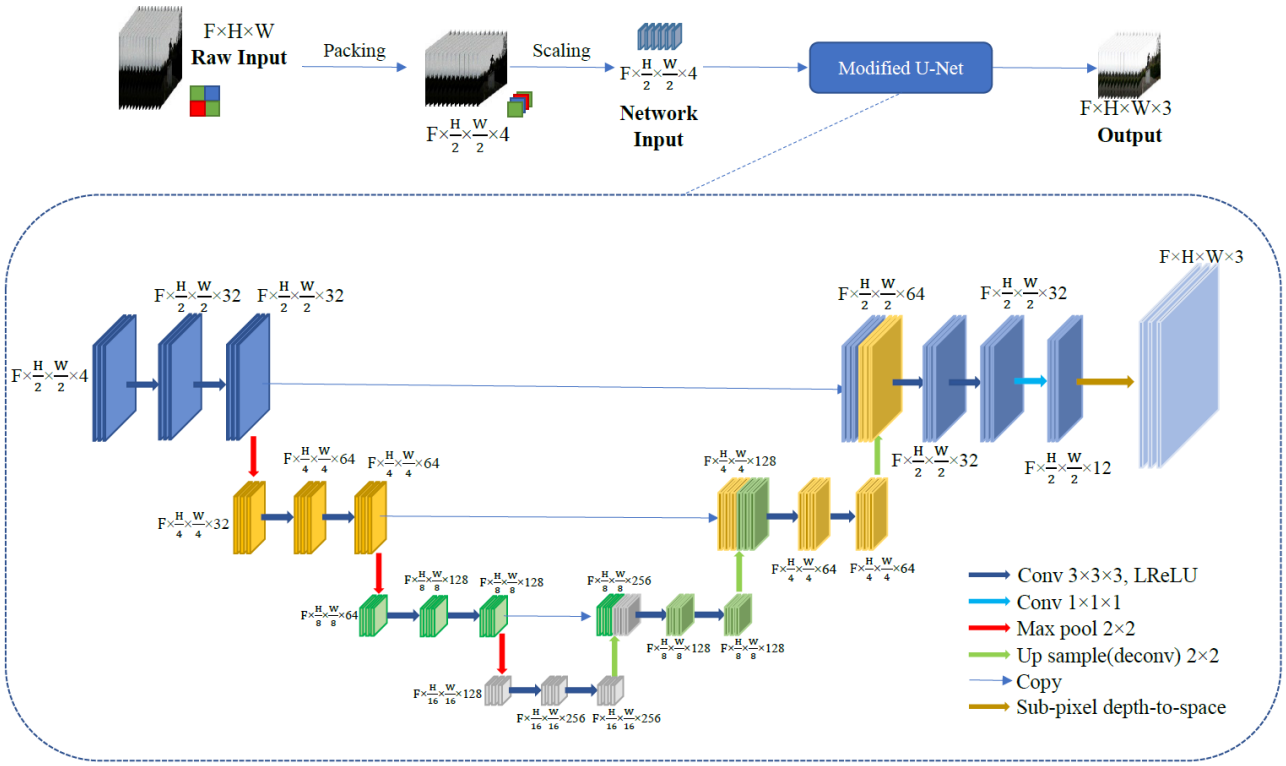


Figure 4. An overview of our modified U-Net enhancement pipeline. On the top is modularized structure illustration; on the bottom, details of the U-Net are demonstrated.

linear scaling process so that their mean values after adjustment lie in the vicinity of $\frac{1}{5}$ of the maximum values determined by storing formats. This ratio is an empirical value properly chosen after preparatory experiments and was maintained throughout subsequent processes. Still, this could bring a degree of freedom to fine-tune a model in deployment stage.

4.2. Network

Recent researches [30, 4, 38] have explored DCN's effectiveness in image-to-image tasks, especially U-Net [3]. Inspired by this fact, we based our main enhancing module on an end-to-end U-Net framework.

Low-light image enhancement algorithms directly applied to each frame of a video can easily cause flickering problems. To avoid such drawbacks and take advantage of temporal information, 3D convolution layers were adopted in our network to substitute for traditional 2D ones. In contrast to MBLLVEN derived from MBLLEN [24], we didn't simply upgrade all layers to 3D calculations, as we found out that even though 2D pooling and deconvolution layers help networks perceive abstract elements in feature maps, 3D versions of them could mishandle sequential temporal information. Hence our modified U-Net combined 3D convolutions and 2D pooling/deconvolutions together to better integrate spatial-temporal information.

During preliminary trials, we tried to reduce the depth of a U-Net and channel numbers of feature maps to determine whether redundancy exists. After a few investigations, we settled down to three downsampling/upsampling layer combinations only (compared to originally four pairs) with the same channel numbers as before.

Final step of the U-Net changes output channel number to 12 by a 1×1 convolutional layer to accommodate depth-to-space layer's dimension requirements. Detailed dimensional information of our modified U-Net can be found in Figure 4.

4.3. Training

Our network was trained from scratch with the L_1 loss function and Adam Optimizer [16]. For a packed training input video, it was cut to $16 \times 256 \times 256$, where 16 is the frame number and the rest are spatial size. The beginning frame numbers of cropped data were multiples of 4 starting from 0, meaning that each video clip could generate (200 - 16) / 4 = 46 training data arrays in one epoch. 256×256 spatial windows were randomly chosen for each data array. Random transpose along spatial-temporal dimensions were also applied for data augmentation. Learning rate was set to 10^{-4} initially and dropped to 10^{-5} after 15 epochs, 10^{-6} after 30 epochs. Training process proceeded for 60 epochs.

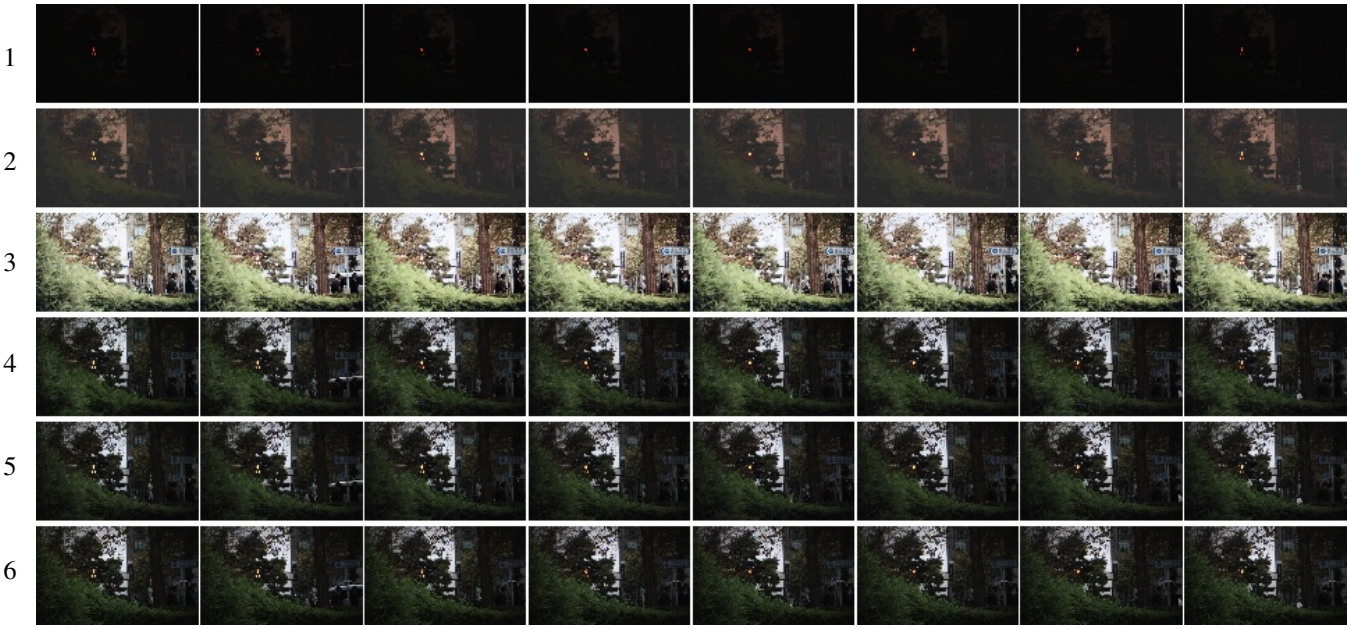


Figure 5. Key frames of qualitative results comparison. From top to bottom: 1. RGB formatted input; 2. Linearly scaled input; 3. MBLLVEN result; 4. Chen *et al*.'s work; 5. Our result; 6. Ground truth.

5. Experiments

5.1. Qualitative Results

Comparisons overview. We compared our pipeline's results to direct linearly scaled result/input, MBLLVEN[24] trained on our dataset, and Chen *et al*.'s work [3] trained on our dataset in a frame-by-frame approach. An overview of the qualitative comparison results is demonstrated in Figure 5.

From top to bottom, each row corresponds to key frames of: 1) RGB formatted input, 2) scaled frames, 3) MBLLVEN result, 4) Chen *et al*.'s work's result, 6) our result, and 7) ground truth. Linearly scaled videos are presented here as a reference of our network's direct input and provides a concept of noise level inside the video.

MBLLVEN was implemented and trained on our dataset according to Lv *et al*.'s work [24]. Original MBLLVEN enhances videos from ordinary RGB format to ordinary RGB format. We changed the shape of its first layer's kernel, so that 12-bit raw data can be processed directly. From the comparison results it can be seen that results produced by MBLLVEN suffer from noise and over adjusting problems, losing naturalness of videos.

For Chen *et al*.'s work, we shuffled all the frames we had in training dataset to train their 2D U-Net and applied it to each frame during test phase. From the overview picture, we could see that both Chen *et al*.'s results and our network's recover brightness and color accurately. With that being said, pictures from Chen *et al*.'s network result showed subtle evidence of overshooting in terms of light intensity restoration, *e.g*. the fifth key frame in their result seemed brighter compared to ours and ground truth. Our results, on the contrary, followed ground truth's luminance perfectly.

Comparison with Chen *et al*.'s method. We further compared our method with Chen *et al*.'s work on different aspects to demonstrate clearer distinctions. We first tested to what extent did our SMOID dataset serve as a helpful tool by reporting results from a SID trained network. As can be seen from 6, using SID trained network didn't accurately learn a color mapping of enhancing low light video frames.

Because Chen *et al*.'s work could only process videos frame-by-frame, we would then like to demonstrate easily occurring flickering problem. However, we present only a brief analysis here to help capture the gist. It is strongly recommended that effects of flickering are to be perceived by video clips bundled in supplementary materials of this paper.

In Figure 7, separate adjacent frames were extracted from representative videos generated by Chen *et al*.'s work and our pipeline. Picture (a) and (b) in Figure 7 are starting points from Chen *et al*.'s network and our network generated results separately. The following 10 frames are on the top right side of the figure, showing brightness and color differences.

In picture (d), two curves of relative luminances related to each frame in the clip are portrayed. We can see despite that they are on different average levels, dispersion in the red curve, corresponding to Chen *et al*.'s network result, is higher than ours. Our network, by introducing 3D opera-

Figure 6. Comparison between Chen *et al*.'s pipeline trained on different dataset. First row: trained on their own dataset, SID; Second row: trained on our dataset. SMOID.

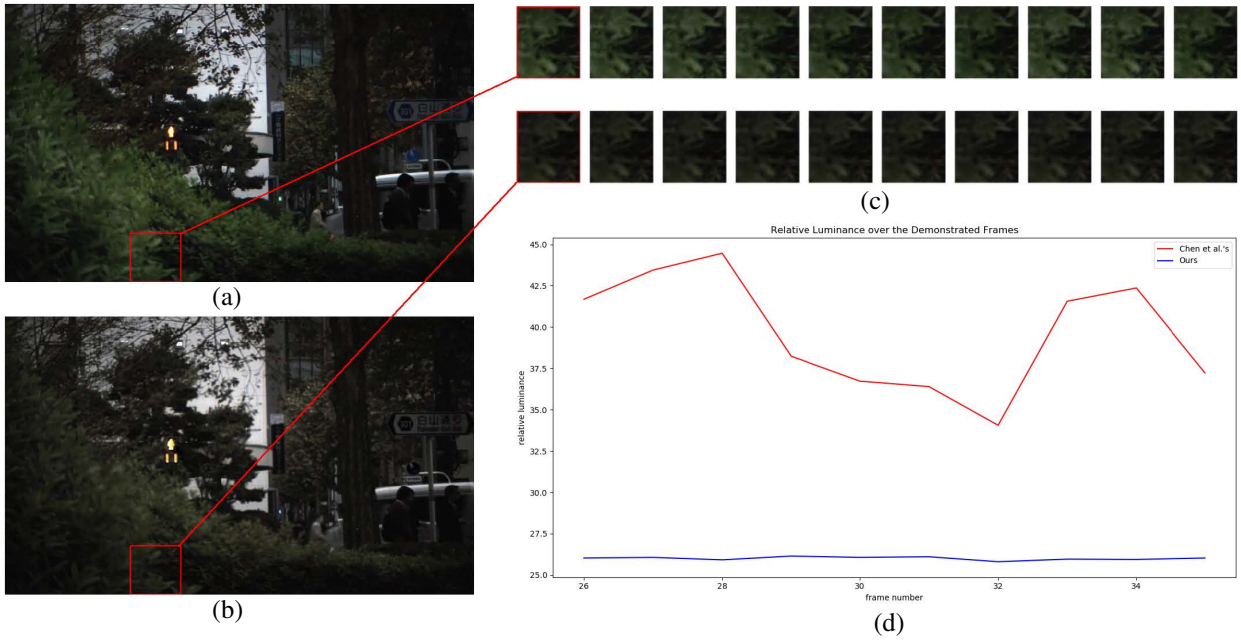


Figure 7. Example of flickering effect in results. (a) frame 26 from Chen *et al*.'s result; (b) frame 26 from our method's result; (c) continuous 10 frames from both videos; (d) average luminance in each frame of these two clips.

tions, maintains a smooth overall intensity.

Qualitative results on different cameras and of different scenes.

We also tested our trained network on raw videos obtained by other cameras, aiming at different scenes and with various gain levels. This is to examine generalization ability of our proposed pipeline.

From presented frames of result videos, we could see that our network was not overfitting to street views in SMOID, rather capable of generic dark video enhancement tasks, successfully improve visibility and color rendering. Full-length cross camera test result videos are available in supplementary materials.

Figure 8. Uncurated example frames of cross-camera tests. Two new cameras (new camera 1 and new camera 2) were used, along with different gain levels. Each image's left part is one frame from the corresponding input, and right part is one frame from the output.

5.2. Quantitative Results

For detailed quantitative measurements, we used three criteria to perform our analysis. PSNR, SSIM [36], and mean square error (MSE) of mean absolute brightness differences (MABD) vectors. PSNR results are shown in Table 2. SSIM results are shown in Table 3. These values corroborate that Chen *et al*.'s work and MBLLVEN lacks ability to accurately recover details from real world low-light videos compared to our method.

PSNR	Chen *et al.*'s	MBLLVEN	Ours
Gain 0 Test	24.73	24.33	25.19
Gain 5 Test	24.42	25.62	30.59
Gain 10 Test	24.68	26.72	30.46
Gain 15 Test	27.28	27.06	31.67
Gain 20 Test	27.96	26.97	30.06
Average	25.81	26.23	29.86

Table 2. PSNR of different methods' result on test dataset.

SSIM	Chen *et al.*'s	MBLLVEN	Ours
Gain 0 Test	0.8372	0.7840	0.8868
Gain 5 Test	0.8499	0.7991	0.9596
Gain 10 Test	0.8516	0.8112	0.9626
Gain 15 Test	0.8732	0.8188	0.9528
Gain 20 Test	0.8586	0.8192	0.9575
Average	0.8541	0.8074	0.9480

Table 3. SSIM of different methods' result on test dataset.

Mean absolute brightness differences (MABD) can be viewed as a general level of time derivatives of brightness value on each pixel location. It is calculated by:

$$MABD(k) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}|br^{k+1}(i,j) - br^{k}(i,j)| \quad (1)$$

M and N are height and width of frames. $br^k(i,j)$ is the brightness of pixel (i,j) at frame k, with k being in the range of $1 \leq k < N_{frames}$. A result of MABD vectors from different methods can be found in Figure 9.

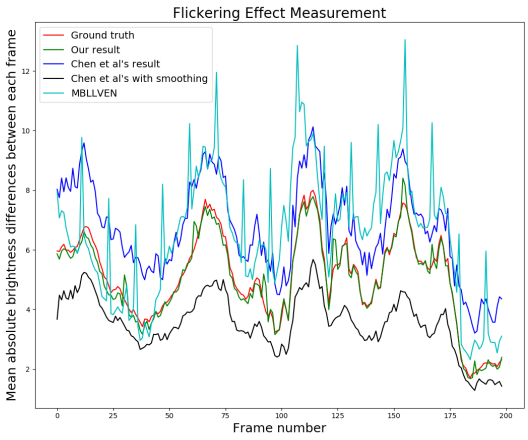


Figure 9. MABD vectors for different networks.

MABD vector of our result follows the one of ground truth better than all other curves, preserving the naturalness of the real world. Chen *et al.*'s and MBLLVEN's variation levels are all above ground truth, showing instabilities in the time axis. We also attached a black curve depicting MABD of smoothed Chen *et al.*'s result. Smoothing method was carried out by a 3-frame temporal filtering. Its over-smoothing effect can be inferred from the fact that its MABD vector is always below the level of ground truth.

Mean square error (MSE) between MABD vector of a result video and that of its ground truth could serve as an indication of its flickering effect. Those values are displayed in Table 4. In this chart, we can see that our pipeline achieved the best temporal smoothness among all approaches.

	$MSE(MABD)$
MBLLVEN	39.31
Chen *et al.*'s	29.44
Chen *et al.*'s + smooth	27.57
Ours	4.436

Table 4. MSE between MABD vectors of different methods' results and MABD vector of corresponding ground truth.

6. Conclusion

In this paper, we proposed a new synchronized dual-camera system. A large paired dataset was gathered by the system, enabling accurate enhancement network training and testing. Modified 3D U-Net was put forward, and trained on our dataset, with complicated scenes and multiple gain levels. Comparison experiments were conducted to show this pipeline's advantages over state-of-the-art low-light video enhancement algorithms. We also demonstrated impressive results of directly applying our trained network to real world dark videos recorded by other equipment different from the system we used to build up SMOID.

With all that being said, there are still various aspects to improve in the future. Dataset capacity will be continuously increased to cover more diverse scenes and objects. Network complexity can be reduced by systematically optimization for possible real-time processing. Different gain level, ISO, aperture, and ND filters with various transmission rates need to be handled by the pipeline more naturally. We hope our work could provide foundations for further exploration in the field of extreme low-light video enhancement as well as its applications.

Acknowledgements. This work was finished when Haiyang Jiang visited the Optical Sensing and Camera System Laboratory (Oscars Lab), led by Dr. Yinqiang Zheng at National Institute of Informatics (NII), Japan, through the NII International Internship Program. This work was supported by the 2018 NII Research Project Funding.

References

[1] Tarik Arici, Salih Dikbas, and Yucel Altunbasak. A histogram modification framework and its application for image contrast enhancement. *Trans. Img. Proc.*, 18(9):1921–1935, Sept. 2009.

[2] Turgay Celik and Tardi Tjahjadi. Contextual and variational contrast enhancement. *Trans. Img. Proc.*, 20(12):3431–3441, Dec. 2011.

[3] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] Qifeng Chen, Jia Xu, and Vladlen Koltun. Fast image processing with fully-convolutional networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2516–2525, 2017.

[5] Justyna Cilulko, Pawe Janiszewski, Marek Bogdaszewski, and Eliza Szczygielska. Infrared thermal imaging in studies of wild animals. *European Journal of Wildlife Research*, 59:17–23, 2012.

[6] Xuan Dong, Yi Pang, and Jiangtao Wen. Fast efficient algorithm for enhancement of low lighting video. *2011 IEEE International Conference on Multimedia and Expo*, pages 1–6, 2010.

[7] Xueyang Fu, Delu Zeng, Yue Huang, Yinghao Liao, Xinghao Ding, and John Paisley. A fusion-based enhancing method for weakly illuminated images. *Signal Processing*, 129:82–96, 2016.

[8] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2782–2790, 2016.

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[10] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26:982–993, 2017.

[11] Dirk W. Hertel, Hervé Maréchal, Daniel A. Tefera, Wensheng Fan, and Rich Hicks. A low-cost vis-nir true color night vision video system based on a wide dynamic range cmos imager. *2009 IEEE Intelligent Vehicles Symposium*, pages 273–278, 2009.

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017.

[13] Daniel J. Jobson, Zia ur Rahman, and Glenn A. Woodell. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 6 7:965–76, 1997.

[14] Daniel J. Jobson, Zia ur Rahman, and Glenn A. Woodell. Properties and performance of a center/surround retinex. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 6 3:451–62, 1997.

[15] Minjae Kim, Dubok Park, David K. Han, and Hanseok Ko. A novel approach for denoising and enhancement of extremely low-light video. *IEEE Transactions on Consumer Electronics*, 61:72–80, 2015.

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[17] Seungyong Ko, Soohwan Yu, Wonseok Kang, Chanyong Park, Sangkeun Lee, and Joonki Paik. Artifact-free low-light video enhancement using temporal similarity and guide map. *IEEE Transactions on Industrial Electronics*, 64:6392–6401, 2017.

[18] Edwin H. Land. The Retinex Theory of Color Vision. *Scientific American*, 237(6):108–128, Dec. 1977.

[19] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, 2017.

[20] Chul Lee, Chang Su Kim, and Chulwoo Lee. Contrast enhancement based on layered difference representation of 2d histograms. *IEEE Transactions on Image Processing*, 22(12):5372–5384, 1 2013.

[21] Chang-Hsing Lee, Jau-Ling Shih, Cheng-Chang Lien, and Chin-Chuan Han. Adaptive multiscale retinex for image contrast enhancement. *2013 International Conference on Signal-Image Technology & Internet-Based Systems*, pages 43–50, 2013.

[22] Huijie Liu, Xiankun Sun, Hua Han, and Wei Cao. Low-light video image enhancement based on multiscale retinex-like algorithm. *2016 Chinese Control and Decision Conference (CCDC)*, pages 3712–3715, 2016.

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

[24] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. MBLLEN: low-light image/video enhancement using cnns. In *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*, page 220. BMVA Press, 2018.

[25] Jiayi Ma, Chen Chen, Chang Li, and Jun Huang. Infrared and visible image fusion via gradient transfer and total variation minimization. *Information Fusion*, 31:100 – 109, 2016.

[26] Keita Nakai, Yoshikatsu Hoshi, and Akira Taguchi. Color image contrast enhancement method based on differential intensity/saturation gray-levels histograms. In *2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, November 2017.

[27] Jianhua Pang, Sheng Zhang, and Wencang Bai. A novel framework for enhancement of the low lighting video. *2017 IEEE Symposium on Computers and Communications (ISCC)*, pages 1366–1371, 2017.

[28] Tal Remez, Or Litany, Raja Giryes, and Alexander M. Bronstein. Deep class-aware image denoising. *2017 International Conference on Sampling Theory and Applications (SampTA)*, pages 138–142, 2017.

[29] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]).

[30] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *2015*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3431–3440, 2015.

[31] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016.

[32] Li Tao, Chuang Zhu, Jiawen Song, Tao Lu, Huizhu Jia, and Xiaodong Xie. Low-light image enhancement using cnn and bright channel prior. *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3215–3219, 2017.

[33] Li Tao, Chuang Zhu, Guoqing Xiang, Yuan Li, Huizhu Jia, and Xiaodong Xie. Llcnn: A convolutional neural network for low-light image enhancement. *2017 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4, 2017.

[34] Dongsheng Wang, Xin Niu, and Yong Dou. A piecewise-based contrast enhancement framework for low lighting video. *Proceedings 2014 IEEE International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, pages 235–240, 2014.

[35] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE Transactions on Image Processing*, 22:3538–3548, 2013.

[36] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004.

[37] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *BMVC*, 2018.

[38] Li Xu, Jimmy Ren, Qiong Yan, Renjie Liao, and Jiaya Jia. Deep edge-aware filters. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1669–1678, Lille, France, 07–09 Jul 2015. PMLR.

[39] Zhenqiang Ying, Ge Li, and Wen Gao. A bio-inspired multi-exposure fusion framework for low-light image enhancement. *CoRR*, abs/1711.00591, 2017.

[40] Zhenqiang Ying, Ge Li, Yurui Ren, Ronggang Wang, and Wenmin Wang. A new low-light image enhancement algorithm using camera response model. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 3015–3022, 2017.

[41] Iffat Zafar, Usman Zakir, Ilya V. Romanenko, Richard M. Jiang, and Eran A. Edirisinghe. Human silhouette extraction on fpgas for infrared night vision military surveillance. *2010 Second Pacific-Asia Conference on Circuits, Communications and System*, 1:63–66, 2010.

[42] Huaizhong Zhang, Chunbo Luo, Qi Wang, Matthew Kitchin, Andrew Parmley, Jesus Monge-Alvarez, and Pablo Casaseca de-la Higuera. A novel infrared video surveillance system using deep learning based techniques. *Multimedia Tools and Applications*, 77(20):26657–26676, 2018.

[43] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26:3142–3155, 2017.

[44] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017.