Dancing in the Dark: A Benchmark towards General Low-light Video Enhancement

Huiyuan Fu[1] , Wenkai Zheng[1], Xicong Wang[1], Jiaxuan Wang[1], Heng Zhang[2], Huadong Ma[1]
[1]Beijing University of Posts and Telecommunications, [2]Xiaomi
{fhy, ciki, wangxc, 2018213169, mhd}@bupt.edu.cn, zhangheng8@xiaomi.com

Abstract

Low-light video enhancement is a challenging task with broad applications. However, current research in this area is limited by the lack of high-quality benchmark datasets. To address this issue, we design a camera system and collect a high-quality low-light video dataset with multiple exposures and cameras. Our dataset provides dynamic video pairs with pronounced camera motion and strict spatial alignment. To achieve general low-light video enhancement, we also propose a novel Retinex-based method named Light Adjustable Network (LAN). LAN iteratively refines the illumination and adaptively adjusts it under varying lighting conditions, leading to visually appealing results even in diverse real-world scenarios. The extensive experiments demonstrate the superiority of our low-light video dataset and enhancement method. Our dataset is available at `https://github.com/ciki000/DID`.

1. Introduction

Low-light video enhancement is a crucial task in computer vision with a wide range of applications, such as surveillance, self-driving cars, and consumer electronics. It aims to improve the visibility and visual quality of videos captured in low-light conditions, which typically suffer from low brightness, low contrast, severe noise, and blur. Despite significant advances in this area, low-light video enhancement remains a challenging problem due to the complex and variable nature of low-light environments.

Recent advances in deep learning methods [14, 2, 9, 20, 4] have shown promising results in low-light video enhancement. However, the performance of deep learning-based methods heavily relies on the quality of the training dataset. Capturing high-quality spatially-aligned video pairs of dynamic scenes is particularly challenging, as it requires the camera to capture a low-light video and a normal-light video of the same scene with identical motion. As a result, existing low-light video datasets have certain limitations. For instance, SIDGAN [19] is a synthetic dataset that lacks real-world variability, and SMID [2] is a static video dataset. While SDSD [20] collected paired low-light videos of real-world scenes, the dataset has some limitations, including a restricted range of camera motion, suboptimal spatial alignment, and inconsistent frames across some scenes. Therefore, building a high-quality low-light video dataset is still a challenging task.

SDSDNet [20] with SDSD | LAN with SDSD
SDSDNet [20] with DID | LAN with DID

Figure 1: Visual comparison on a low-light video captured by an iPhone 14 Pro.

To collect a high-quality low-light video dataset, we design a camera system to capture spatially aligned dynamic video pairs. We build a large-scale paired low-light video dataset named DID, which means "Dancing in the Dark". Compared to previous low-light video datasets, our dataset contains a more diverse set of real-world scenes, larger camera motion, and more accurate spatial alignment. Moreover, our dataset is captured by multiple cameras under varying lighting conditions, which enhances its generalizability.

We propose a Light Adjustable Network (LAN) based on the Retinex theory [10] for general low-light video enhancement. LAN iteratively refines the illumination components to generate enhanced results of different luminance. Our method adapts to different scenes by adaptively selecting the magnitude of the light enhancement and can also be manually adjusted to change the illumination of the results. Thus, our method exhibits good generalization and

avoids over or underexposure in extreme cases, unlike previous one-to-one methods that could only enhance lighting degradation similar to the training low-light samples.

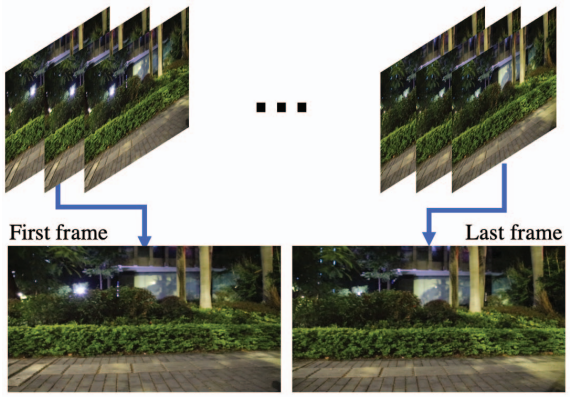
We conducted extensive experiments to demonstrate the generalization and superiority of our dataset and method. Fig. 1 shows the results of different methods trained on both SDSD dataset and our dataset for enhancing low-light videos captured by mobile phones in real scenes.

In summary, the contributions of our work are as follows:

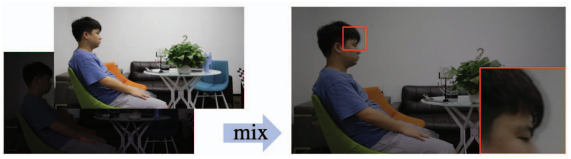
- We build a high-quality paired low-light video dataset with pronounced camera motion, strict spatial alignment, and diverse scene content.
- We propose a Retinex-based low-light video enhancement method, named Light Adjustable Network, which iteratively refines the illumination components and adaptively adjusts the illumination to generate more natural and robust enhanced results.
- We conduct extensive experiments to validate the effectiveness of our dataset and the Light Adjustable Network compared with state-of-the-art methods.

2. DID Dataset

The performance of the low-light enhancement method is affected by the quality of the training dataset, however



(a) Low camera motion.



(b) Misalignment.

(c) Inconsistent scenarios.

Figure 2: Limited quality of SDSD dataset.

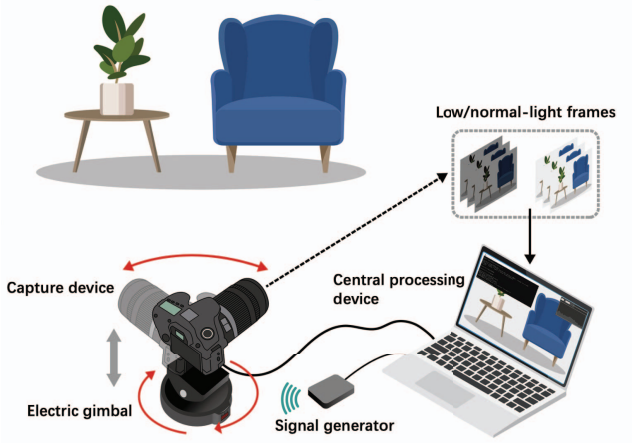


Figure 3: The camera system we built to capture the low/normal-light videos.

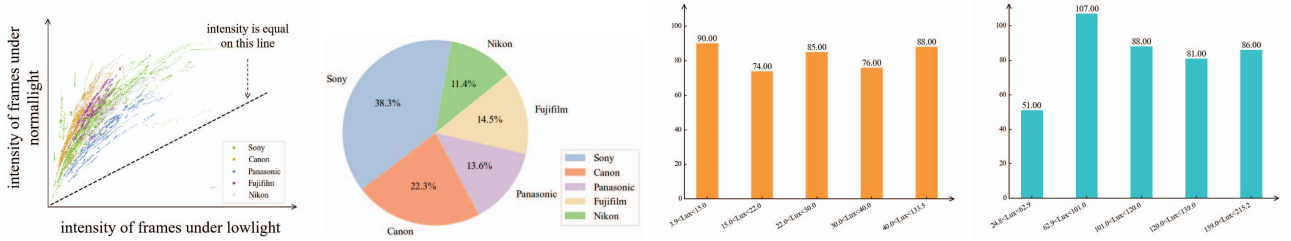
Table 1: Comparison of our dataset with previous low-light video datasets.

Dataset	Real	Status	Capture Device	Num	Release
SIDGAN [19]	×	Dynamic	-	-	✓
EHSC [22]	✓	Dynamic	Canon 5D Mark III	900	×
SMOID [9]	✓	Dynamic	FLIR GS3-U3-23S6C	35800	×
DRV [2]	✓	Static	Sony RX100 VI	22220	✓
SDSD [20]	✓	Dynamic	Canon 6D Mark II	37500	✓
Ours (DID)	✓	Dynamic	Sony RX100 M4 Canon EOSR10 Panasonic G9 Fujifilm XT4 Nikon Z5	41038	✓

collecting high-quality paired low-light video datasets is challenging due to the difficulty of ensuring that two videos with different lighting captured in the same scene have the same motion trajectory. Although paired video data can be generated [19], deep models trained on synthetic data may introduce artifacts and color bias when processing real-captured low-light videos due to the gap between synthetic and real-world data [11]. Some works [2, 20] have released real-captured paired low-light video datasets recently, but they all have some limitations in various ways. As shown in Table 1, EHSC [22] and SMOID [9] have not been released until now, while DRV [2] consists of static videos. And the video quality of SDSD [20] is limited, as shown in Fig. 2. Furthermore, if all videos of the train set are obtained by a camera with the same non-linear camera response, this can lead to severe performance degradation of the model in videos with unknown camera response functions [1]. Models trained with these datasets are often unsatisfactory in real applications scenarios. Therefore, there is a great need for a general high-quality low-light video dataset.

2.1. Camera System

To collect a general high-quality low-light video dataset, we design a camera system consisting of 5 capture devices, an electric gimbal, a signal generator and a central pro-



(a) (b) (c) (d)

Figure 4: Statistical indicators for our DID dataset. (a) Intensity distribution for low/normal-light videos. (b) Distribution of videos captured by five different brand cameras. (c) Luminance distribution for low-light videos. (d) Luminance distribution for normal-light videos.

Figure 5: Two example videos of our DID dataset.

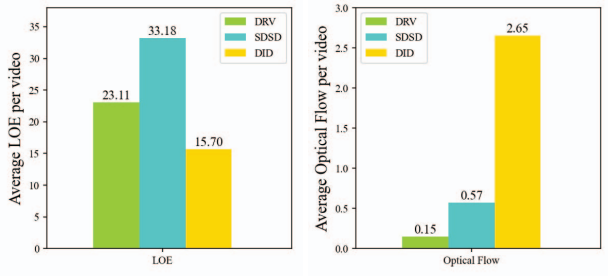


Figure 6: LOE and optical flow of different datasets.

cessing device. Our camera system collects paired video datasets by shooting frame by frame. As shown in Fig. 3, the central processing device adjusts the camera ISO to capture a series of low-light frames and normal-light frames at the same location, and then these frames are sent to the central processing device. The central processing device will check these frames and if their quality is up to standard, then it will synthesize these frames into the final low/normal-light frames, otherwise it will send a signal to recapture them until they meet the requirements. After capturing a pair of frames, the signal generator sends a signal to make the electric gimbal move slightly. To ensure continuity between adjacent frames, we limit the sum of the horizontal and vertical rotation angles of the electric gimbal to less than 1° each time. The specific inspection and synthesis methods for a series of low/normal-light frames and the specific parameters of each shooting device are described in detail in the supplementary material.

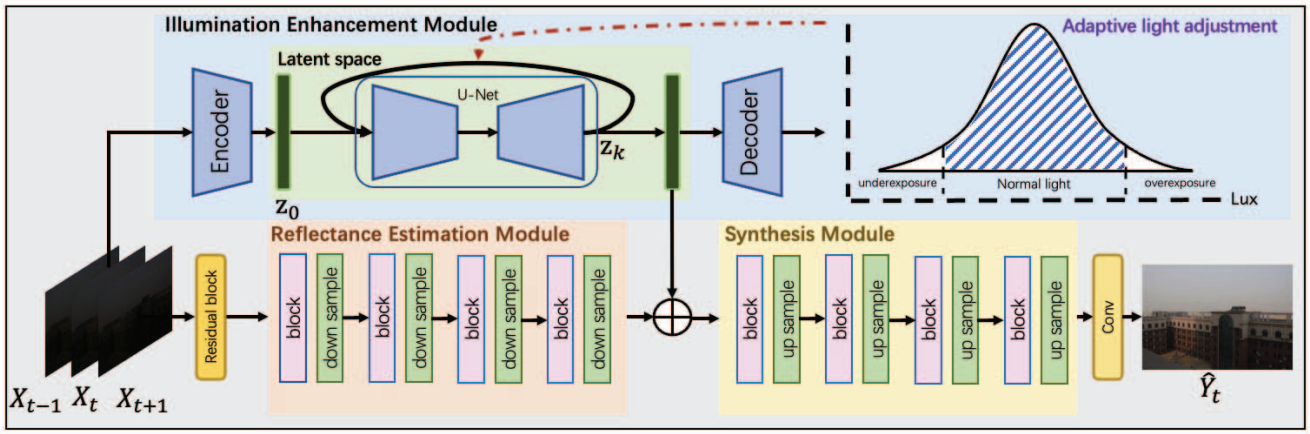
2.2. Video Data

We collected 413 paired videos with a total of 41038 frames and named them as DID dataset (standing for "Dancing in the Dark"). The resolution of our videos is 2560×1440, and more statistical indicators of the overall dataset are shown in Fig. 4. In Fig. 5, we give two samples of different scenarios in our dataset.

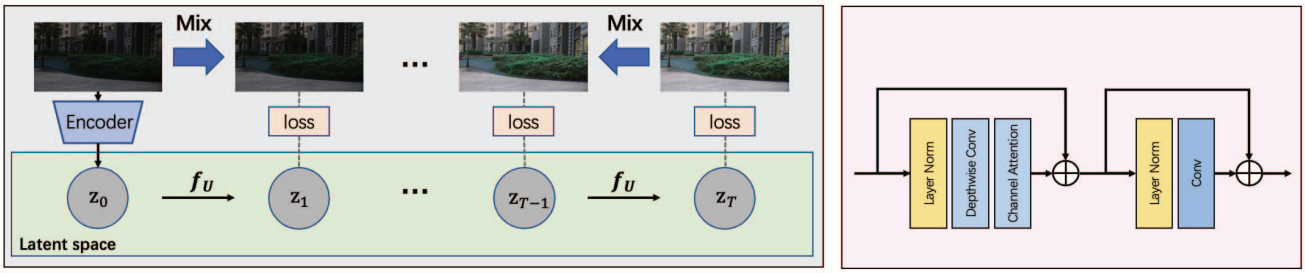
To quantitatively compare our dataset with the previous low-light video datasets, we introduce two metrics, lightness-order-error (LOE) [21] and optical flow [6]. LOE calculates the relative order of lightness in different local areas, which can be used to measure the alignment of paired low/normal-light frames. Optical flow is used for calculating pixel motion between two consecutive images, which can be used to measure the dynamics of a video. Specific implementations of LOE and optical flow are given in the supplementary material. As shown in the Fig. 6, our dataset has the lowest LOE, indicating that our paired videos are well aligned. In addition, the optical flow of our dataset is much larger than that of DRV [2] and SDSD [20], indicating stronger or faster motion in our videos while the videos in DRV and SDSD have low motion activity or are static. Therefore, the model trained on our dataset has better performance than other datasets on real scenarios with high motion activity or strong movement.

In summary, our DID dataset has the following advantages over the previous low-light video datasets:

- DID is a multi-illuminance, multi-camera low-light video dataset.
- DID is a dynamic video dataset with obvious camera motion, rather than static or with only small motion.
- DID is a high quality paired dataset with very precise spatial alignment.
- Experiments demonstrate that models have better performance when trained with our dataset.



(a)



(b) (c)

Figure 7: Overview of our method. (a) The framework of our LAN. (b) The process of iterative illumination refinement. (c) The structure of the blocks in reflectance estimation module and synthesis module.

3. Method

Low-light enhancement is a highly ill-posed problem. Like many such inverse problems, a low-light image or video may correspond to multiple suitable normal-light images or videos. Although previous low-light enhancement methods have been able to generate results close to ground truths, most of these methods are one-to-one models with fixed outputs for a single input resulting in their limited generalization performance. In real application scenarios, the distribution of low-light samples may differ somewhat from existing low-light datasets, and may even contain extremely dark scenes or slightly dark scenes. Therefore simply fitting the light degradation of the training data may lead to suboptimal solutions and often leaves the enhanced results overexposed or underexposed. For general low-light video enhancement, we propose Light Adjustable Network(LAN), which can adaptively adjust the illumination to generate appropriately exposed results, and can also be used by the user to adjust the light intensity to generate different outputs.

3.1. Light Adjustable Network

Fig. 7 (a) illustrates the framework of our proposed LAN. According to Retinex theory [10], we decompose the input video frames $X_{t+i,i\in[-k,k]}$, into reflectance components R_t and illumination components I_t, and then enhance the illumination components by iterative refinement, and finally synthesize them into normal light frames $\hat{Y}_t$.

Specifically, given a sequence of low-light frames $X_{t+i} \in \mathbb{R}^{k\times H\times W\times 3}$, we first concatenate them and project them as embedding $F_t^0 \in \mathbb{R}^{H\times W\times C}$ through a Residual Block [7]. Then, the reflectance estimation module estimates the reflectance R_t from it. The reflectacne estimation module is a hierarchical structure with 4 stages, each stage consisting of a feature extraction block and a down-sampling layer. The components of the feature extraction block are shown in Fig. 7 (c), which is inspired by [28, 3]. For the i-th stage, the feature map $F_t^{i-1} \in \mathbb{R}^{\frac{H}{2^{i-1}}\times\frac{W}{2^{i-1}}\times 2^{i-1}C}$ will be processed as feature map $F_t^i \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times 2^i C}$. F_t^4 will be considered as the reflectance component R_t and then sent to the synthesis module.

The illumination enhancement module first encodes the input frame X_t into a latent representation z_0 by a pretrained encoder, and then enhances the illumination component by iterative refinement, which will be described in detail in Section 3.2.

The estimated reflectance R_t and the enhanced illumination $\hat{I}_t$ are first aligned by a convolutional layer and then fed into the synthesis model. The synthesis module is also a hierarchical architecture with 4 stages, each stage consisting of a feature fusion block and a up-sampling layer.

The components of the feature fusion block are the same as the feature extraction block of the reflectance estimation module. In addition, we use skip connections in the corresponding stages of the synthesis module and reflectance estimation module to better recover image details. Finally, the feature map output from the synthesis module is projected by a convolutional layer as the enhanced result $\hat{Y}_t$.

3.2. Iterative illumination refinement

Since the enhancement of the illumination component requires more global information than local texture information, we think it is a good way to process it in a lower-dimensional representational space. This also prevents iterative refinement from taking up too much computational resources. Therefore we first encode the input X_t into the latent space via an autoencoder, which is perceptually equivalent to the data space.

To be able to adjust the light intensity of the enhanced result, we iteratively refine the light so that the illumination component can be adjusted by changing the number of iterations. As shown in Fig. 7 (b), the illumination enhancement module generates target illumination features z_t in T refinement steps. Starting with the latent representation z_0, the module iteratively refines the illumination component through successive iterations (z_1, z_2, ... , z_{T-1}, z_T). The ground truths for illumination of different intensities are defined as latent representations of a mixture of low light frames and corresponding normal light frames. The ground truth of $z_{k,k\in[1,T]}$ is shown below:

$$\begin{aligned} \alpha_k &= \frac{k}{T} \\ \tilde{z}_k &= \mathcal{E}\left(\alpha_k \cdot Y_t + (1-\alpha_k)\cdot X_t\right) \end{aligned} \tag{1}$$

where $\mathcal{E}$ denotes the representation of the pre-trained encoder.

Inspired by diffusion models [8, 16, 18], we use a U-Net [17] to learn the mapping from $\tilde{z}_k$ to $\tilde{z}_{k+1}$. Then the illumination component can be adjusted by changing the number of iterative refinements.

3.3. Adaptive light adjustment

Although our model has better generalization than one-to-one networks, we believe that manual adjustment to select the appropriate illumination is a suboptimal solution. Therefore, we further improve our model so that it can adjust the illumination adaptively in different scenarios.

Since the model tends to perform poorly on low-light samples that differ significantly from the distribution of the training data, we vary the number of iterations so that the recovered illumination features approximate the illumination distribution of the normal-light samples of the training data. We assume that the intensity data of the normal-light samples l follow a Gaussian distribution,

$$l \sim \mathcal{N}\left(\mu, \sigma^2\right) \tag{2}$$

where μ denotes the mean and σ denotes the standard deviation. We perform statistical analysis on normal light samples of the training data and calculate the sample mean and sample standard deviation. With a specific distribution of normal light intensities, we calculate the intensity of the generated illumination component z_k for each iteration and perform a hypothesis test on the intensities of (z_{k-2n}, z_{k-2n+1} , ... , z_k), where n is a pre-defined interval parameter. We use an one-sided Student's test, which compares the mean intensity $\bar{l}_k$ of (z_{k-2n}, z_{k-2n+1} , ... , z_k) with the known mean μ of the normal-light intensity distribution. The null hypothesis H_0 and the alternative hypothesis H_1 are shown below:

$$\begin{aligned} H_0 &: \bar{l}_k \geq \mu \\ H_1 &: \bar{l}_k < \mu \end{aligned} \tag{3}$$

The test statistic is calculated as:

$$t = \frac{\bar{l}_k - \mu}{s_k}\sqrt{2n+1} \tag{4}$$

where s_k is the standard deviation of the intensities of (z_{k-2n}, z_{k-2n+1} , ... , z_k).

The rejection region is:

$$t < t_{\alpha,2n} \tag{5}$$

where α is the significance level (we take 0.05) and $2n$ is the degree of freedom. $t_{\alpha,2n}$ is the α quantile of a t-distribution with $2n$ degrees of freedom. It represents that under the t-distribution, there is an α probability that the value is less than $t_{\alpha,2n}$.

Finally, we compare our calculated t with critical value $t_{\alpha,2n}$. If $t < t_{\alpha,2n}$ then we reject the null hypothesis (i.e. the mean intensity $\bar{l}_k$ is less than the known mean μ) and the illumination enhancement module continues to iterate to generate a higher intensity illumination component; otherwise we cannot reject the null hypothesis and the illumination enhancement module stops iterating and sends z_{k-n} to the synthesis module.

In addition, to prevent brightness jitter in the enhanced video, we limit the difference in the number of iterations of adjacent frames to no more than $\frac{T}{p}$ (where p is a hyperparameter).

3.4. Training and Loss Function

First, we follow [16] to train a VAGAN [5, 27] for encoding the input frames into latent space. Then we train the illumination enhancement module. For each latent representation of the input samples, we let the U-Net used for iterative refinement learn the mapping from $\tilde{z}_k$ to $\tilde{z}_{k+1}$, as follows:

$$\mathcal{L}_U = \sqrt{\|f_U(\tilde{z}_k) - \tilde{z}_{k+1}\|_F^2 + \epsilon^2} \tag{6}$$

where f_U denotes the mapping function learned by U-Net and $\mathcal{L}_U$ is the loss term used to train U-Net, $\|\cdot\|_F$ represents Frobenius norm and the constant ϵ is set to 0.001. $k \in [1, T]$ is a random variable. By employing a lightweight U-Net architecture and a limited number of iterations, we ensure that iterative refinement does not impose a performance bottleneck on the model.

Finally we train the entire network. According to Retinex theory, the reflectance components of low-light frames and paired normal-light frames should be consistent, so we add reflectance consistency loss as follows:

$$\mathcal{L}_R = \|f_R(X_t) - f_R(Y_t)\|_F^2 \tag{7}$$

where f_R denotes the mapping function of the reflectance estimation module.

The overall loss function to train our LAN is summarized as:

$$\mathcal{L} = (1-\lambda)\sqrt{\|\hat{Y}_t - Y_t\|_F^2 + \epsilon^2} + \lambda\mathcal{L}_{\text{SSIM}}\left(\hat{Y}_t, Y_t\right) + \frac{\mathcal{L}_R}{\tau} \tag{8}$$

where λ is a trade-off parameter, $\mathcal{L}_{\text{SSIM}}$ represents the structural similarity loss [24], and τ denotes a temperature parameter.

4. Experiments

4.1. Implementation Details

We show the superiority of our proposed approach and the effect of our constructed DID through experiments in this section. To evaluate the effect of our method, we retrain 9 previous representative methods on the DID and SDSD datasets for comparison and give an ablation study for our method. In addition, a user study is conducted to demonstrate the results of our approach and the chosen baselines.

We divide our DID dataset into a training set, a test set, and a validation set in the ratio of 3:1:1. Then we use the training set to train our LAN. We augment the data using rotation and horizontal flipping and optimize the network by AdamW optimizer [13] with the momentum terms of (0.9, 0.999). We set the learning rate to 0.001 and use the cosine decay strategy to decrease it. Our default number of iterations $T = 10$ and we train LAN for 200 epochs.

4.2. Quantitative Evaluation

To comprehensively evaluate the effectiveness of our proposed method, we conduct quantitative experiments on paired video datasets captured under various scenes, including both the DID and SDSD datasets. Specifically, we evaluate the performance of our method on the test dataset of DID, which comprises videos with diverse scenes and illumination conditions, including some challenging data with extremely low illumination levels that are difficult to recover. We compare the quality of the enhanced videos produced by our Light Adjustable Network with state-of-the-art methods. Moreover, we further evaluate the performance of our approach on the test dataset of SDSD, which includes 12 indoor video pairs and 13 outdoor video pairs.

We adopt two well-known objective evaluation metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM)[24]. PSNR is the ratio between the maximum possible power of normal light image and the power of the enhanced image and measures the fidelity between them. SSIM is a perceptual approach for predicting the quality of digital images and videos, based on the change of structural information between two images.

Table 2: Quantitative results of different methods on SDSD and our DID datasets.

Methods	Learning	SDSD		DID	
		PSNR	SSIM	PSNR	SSIM
DRBN[26]	Image	22.31	0.65	25.22	0.91
RUAS[12]	Image	15.48	0.64	17.01	0.74
LLFlow[23]	Image	24.90	0.78	25.71	0.92
SNR-Aware[25]	Image	25.27	0.82	24.05	0.90
SCI[15]	Image	16.90	0.64	11.15	0.44
MBLLEN[14]	Video	21.79	0.65	24.82	0.91
SMID[2]	Video	24.09	0.69	22.97	0.87
SDSDNet[20]	Video	24.92	0.73	21.88	0.83
Chhirolya et al.[4]	Video	23.46	0.79	22.77	0.88
Ours (default T)	Video	26.95	0.85	27.28	0.92
Ours (adaptive)	Video	**27.25**	**0.85**	**29.01**	**0.94**

Table 3: The results of different low-light enhancement methods in the user study. "LAN" is the percentage that our result is preferred, "Other" is the percentage that some other approach is preferred, "Same" is the percentage that the users have no preference.

Methods	Other	Same	LAN
MBLLEN[14]	40.7%	14.8%	44.4%
SMID[2]	17.9%	20.5%	61.5%
SDSDNet[20]	0%	0%	100%
Chhirolya et al.[4]	13.9%	27.8%	58.3%

Table 4: The results of different datasets in the user study. "DID" is the percentage that our dataset is preferred, "SDSD" is the percentage that the SDSD dataset is preferred, "Same" is the percentage that the users have no preference.

Methods	SDSD	Same	DID
MBLLEN[14]	0.0%	3.7%	96.3%
SMID[2]	38.0%	0%	62.0%
SDSDNet[20]	38.1%	19.0%	42.9%
Chhirolya et al.[4]	23.8%	4.8%	71.4%
LAN	8.33%	8.33%	83.3%

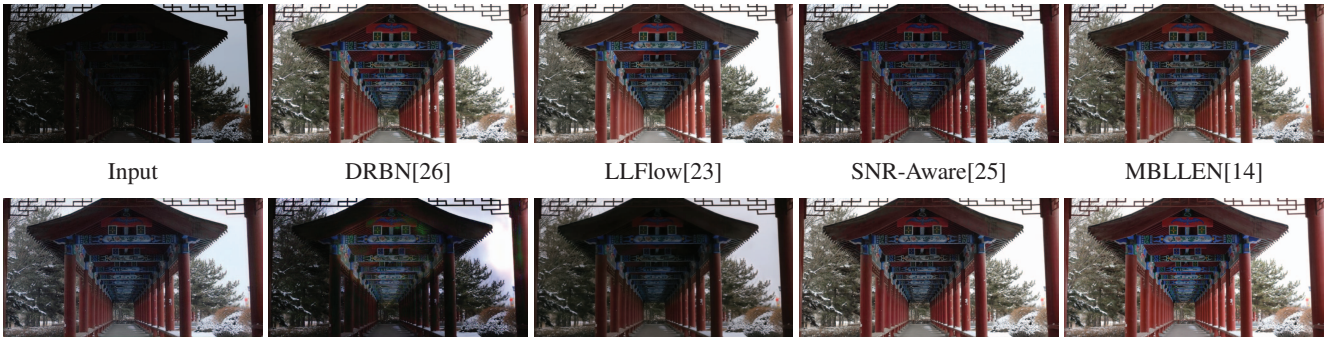


Figure 8: Visual comparison with state-of-the-art low-light enhancement methods on DID dataset.

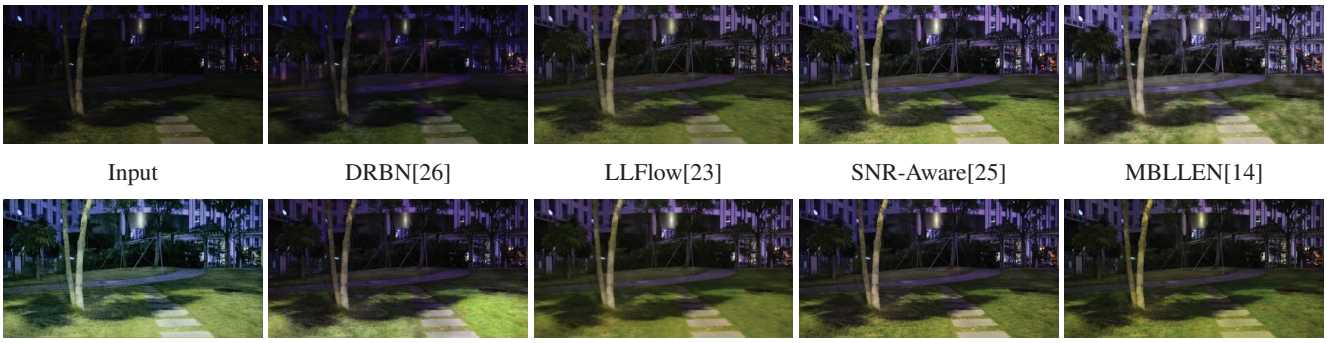


Figure 9: Visual comparison with state-of-the-art low-light enhancement methods on SDSD dataset.

Table 2 presents the quantitative evaluation results of different methods on both DID and SDSD datasets. As shown in the table, our proposed Light Adjustable Network (LAN) outperforms all other methods in all metrics, demonstrating its superior performance in low-light video enhancement. Particularly, our method achieves higher PSNR values than all other methods with a significant margin (more than 3dB on DID and more than 1.9dB on SDSD, respectively). This superiority highlights the effectiveness of our approach in enhancing low-light videos compared to all other methods. Furthermore, our adaptive lighting adjustment strategy is shown to be very effective in improving the performance of the model, especially in datasets with richer scenes.

4.3. Qualitative Evaluation

We perform thorough qualitative evaluations on the DID and SDSD datasets to assess the performance of our proposed method. Fig. 8 presents the results obtained on the DID dataset, where it is observed that SNR-aware, SMID and the method proposed by Chhirolya et al. produce images with a darker tone, leading to substantial color deviation. Moreover, SMID suffers from high levels of noise, while LLFlow exhibits noticeable checkerboard artifacts. DRBN and MBLLEN fail to depict image details effectively, and SDSDNet produces images with severe artifacts.

Fig. 9 presents the visual results obtained on the SDSD dataset, which comprises low-quality videos with significant noise that pose challenges for enhancement. In comparison to the GroundTruth, DRBN yields images with low brightness and weak enhancement. LLFLow exhibits noticeable checkerboard artifacts. Both the Chhirolya et al.'s method and SDSDNet produce images with obvious artifacts or noise, which fail to display specific detailed information. SNR-aware, MBLLEN and SMID result in significant color deviation, which adversely affects the visual quality of the images.

Fig. 10 shows the visual results for the extreme samples. It is evident that for videos with extremely low illumination, most methods produce underexposed outputs, whereas for videos with slightly low illumination, most methods generate overexposed outputs. Our proposed method, with its default iterative settings, has successfully achieved superior results and further improved the brightness with the light adaptive light adjustment strategy. The impact of the number of iterations on the enhanced results is also investigated and presented in the supplementary material.

After a comprehensive evaluation of the comparative results of different methods on the two datasets, our proposed

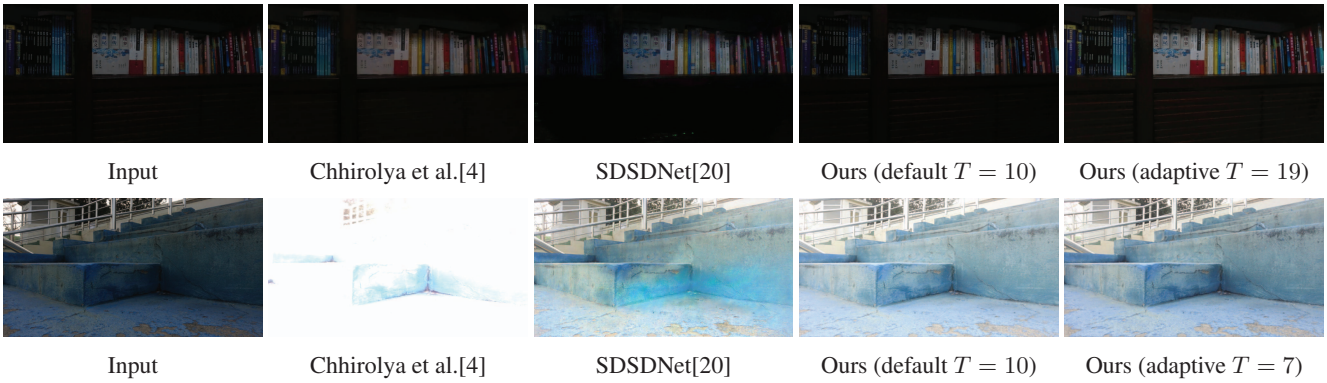


Figure 10: Visual comparison on extremely dark and slightly dark videos.

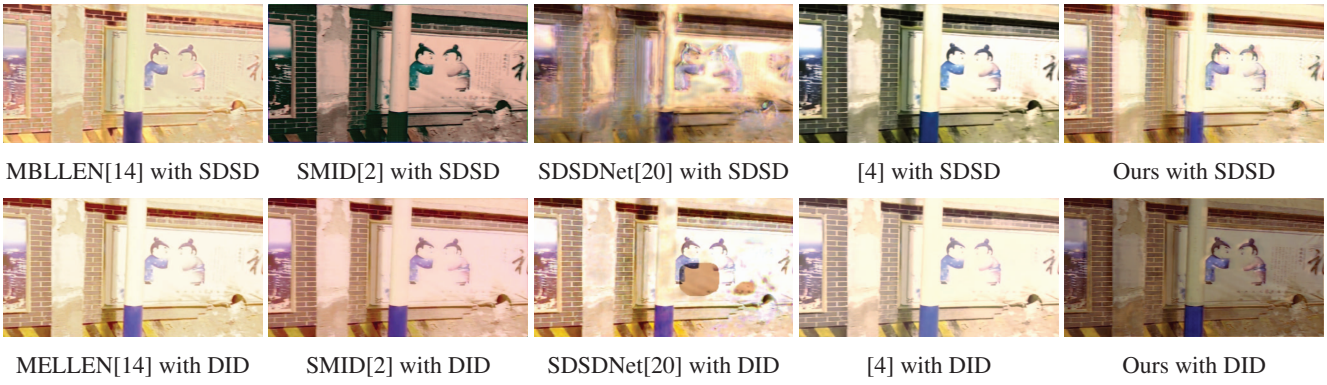


Figure 11: Visual results in the user study.

method demonstrates excellent visual performance in terms of global brightness, color recovery and details.

4.4. User Study

We conduct a user study with 20 participants to compare the subjective visual quality of LAN and other video low-light enhancement methods. We use an iPhone 14 Pro to capture 20 videos in real-world scenarios using camera motion and local subject motion for the user study. The videos are then enhanced using five video low-light enhancement methods (MBLLEN, SDSDNet, Chhirolya et al.'s method, SMID and LAN), each trained on both DID and SDSD to compare the performance of different models and the generalization of models trained on these two datasets.

Each participant underwent five sets of tests, including three sets of model performance tests and two sets of model generalization tests, with three videos randomly selected for comparison in each set. The model performance tests use models trained on the DID dataset. The participants first randomly select three out of the four video enhancement methods other than LAN and compare their results with those of LAN. Then, two methods are randomly selected from the five methods for dataset generalization testing, where the participants compare the results of models trained on both SDSD and DID for each method. In each comparison, participants simultaneously view two videos (referred to as video A and video B) and compare them in terms of photo realism, brightness, contrast etc., making a choice among three options: "Video A is better", "Video B is better", or "I cannot determine which is better".

The quantitative results in the user study are shown in Table 3 and Table 4, respectively. Besides, the visual results in the user study can be seen in Fig. 11. The tables and the figure present the comparison results between our method and other methods as well as the comparison results for generalization between the SDSD and DID datasets. The data indicates that our method is more appealing to users in all comparisons with other methods, suggesting that our results are more natural and realistic. In addition, in the comparison of generalization between the DID and SDSD datasets, our DID dataset is found to have better generalization across all methods according to user feedback. It is evident that models trained on SDSD exhibit blurry enhanced results for real-captured low-light videos. This blurriness primarily arises from imprecise spatial alignment in the paired training data. In contrast, models trained on DID produce enhancement results without such blurriness.

5. Conclusion

In this paper, we present a dynamic high-quality paired low-light video dataset, called DID ("Dancing in the Dark"), captured using our designed camera system, with pronounced camera motion and strict spatial alignment. Based on the Retinex theory, we propose a Light Adjustable Network (LAN) for general low-light video enhancement, which adaptively adjusts the illumination to generate natural and robust enhanced results. Extensive experiments and user studies demonstrate the effectiveness of our proposed dataset and method, which outperform state-of-the-art approaches. Our work provides a valuable resource and a novel method for low-light video enhancement.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 62272059, the Funds for Creative Research Groups of China under Grant 61921003, the Beijing Nova Program under Grant Z201100006820124, and the 111 Project (B18008).

References

[1] Sara Aghajanzadeh and David Forsyth. Long scale error control in low light image and video enhancement using equivariance. *arXiv preprint arXiv:2206.01334*, 2022.

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

[3] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 17–33. Springer, 2022.

[4] Shivam Chhirolya, Sameer Malik, and Rajiv Soundararajan. Low light video enhancement by learning on static videos with cross-frame attention. *arXiv preprint arXiv:2210.04290*, 2022.

[5] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[6] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*, pages 363–370. Springer, 2003.

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

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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

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

[11] Chongyi Li, Chunle Guo, Linghao Han, Jun Jiang, Ming-Ming Cheng, Jinwei Gu, and Chen Change Loy. Low-light image and video enhancement using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):9396–9416, 2021.

[12] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10561–10570, 2021.

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[14] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. Mbllen: Low-light image/video enhancement using cnns. In *BMVC*, volume 220, page 4, 2018.

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

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[18] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.

[19] Danai Triantafyllidou, Sean Moran, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. Low light video enhancement using synthetic data produced with an intermediate domain mapping. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 103–119. Springer, 2020.

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

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

[22] Wei Wang, Xin Chen, Cheng Yang, Xiang Li, Xuemei Hu, and Tao Yue. Enhancing low light videos by exploring high

sensitivity camera noise. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4111–4119, 2019.

[23] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex Kot. Low-light image enhancement with normalizing flow. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2604–2612, 2022.

[24] 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(4):600–612, 2004.

[25] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Snr-aware low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17714–17724, 2022.

[26] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3063–3072, 2020.

[27] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.

[28] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022.