Rethinking the Low-Light Video Enhancement: Benchmark Datasets and Methods

Jiaxuan Wang, Huiyuan Fu, *Member, IEEE*, Wenkai Zheng, Xicong Wang,
Xin Wang, *Senior Member, IEEE*, Heng Zhang, and Huadong Ma, *Fellow, IEEE*

***Abstract*—Low-light video enhancement is a critical task in computer vision with a wide range of applications. However, there is a lack of high-quality benchmark datasets in this field. To address this issue, we collect a high-quality low-light video dataset using a well-designed camera system. The videos in our dataset feature apparent camera motion and strict spatial alignment. In order to achieve general low-light video enhancement, we propose a Retinex-based method called Light Adjustable Network (LAN). LAN iteratively adjusts the brightness and adapts to different lighting conditions in various real-world scenarios, producing visually appealing results. We further develop a new dataset capture method and low-light video enhancement method to address the limitation of our previous dataset in capturing dynamic scenes and previous method. The new camera setup and capture method enable the recording of real continuous videos and generate the new dataset. Our new low-light video enhancement method, LAN++, leverages a new inter-frame relationship, difference images. It utilizes the texture information contained in the difference images of dynamic scenes to supplement the high-frequency details of the original features, which produce sharper and more realistic output images. The extensive experiments demonstrate the superiority of our low-light video dataset and enhancement method. Our dataset can be downloaded at https://pan.baidu.com/s/1d3EljvVduVM0wUOvzjWaqA?pwd=p45g.**

***Index Terms*—Computational photography, low-light video enhancement, image decomposition, low-light video dataset.**

I. Introduction

Enhancing videos captured in low-light conditions is a vital area of computer vision that has various applications, including consumer electronics, surveillance, and self-driving cars. The objective is to enhance visibility and improve the visual quality of these videos, which usually suffer from low brightness, low contrast, severe noise, and blur. Although many advancements have been made in this field, low-light video enhancement remains a challenge due to the complex and unpredictable nature of low-light settings.

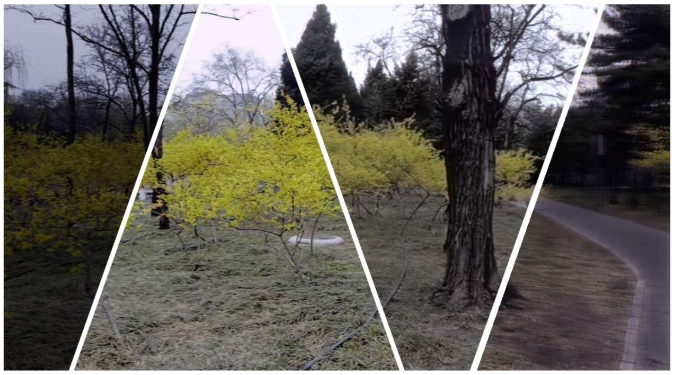


Fig. 1. Visual comparison on a real-captured low-light video. From left to right are low-light input, our LAN++ trained on our DID dataset, our LAN++ trained on R-DID dataset, and our LAN++ trained on SDSD dataset.

Deep learning methods [11], [12], [13], [20], [21], [34], [38] have recently shown promising results in the enhancement of low-light videos. However, the effectiveness of these methods heavily depends on the quality of the training dataset. Acquiring high-quality video pairs of dynamic scenes in both low-light and normal-light conditions is particularly difficult, as it requires capturing videos of the same scene with identical motion. Existing low-light video datasets have limitations that hinder their usefulness. For example, synthetic datasets like SIDGAN [14] lack real-world variability, while static video datasets like SMID [11] fail to capture dynamic scenes. Although SDSD [13] collected paired real-world low-light videos, it still has limitations such as restricted camera motion, imperfect spatial alignment, and inconsistencies across frames. Consequently, constructing a high-quality low-light video dataset remains a challenging task.

Due to the limitations in the current methods for spatiotemporal alignment in low-light video dataset acquisition, we design two camera systems to capture pairs of low-light videos, with one system focusing on spatial alignment and the other on temporal alignment. For spatial alignment, we build a large-scale paired low-light video dataset called "Dancing in the Dark" (DID). DID excels in providing more precise spatial alignment, diverse scenes, and larger camera motion.

Received 1 February 2025; revised 9 July 2025; accepted 22 September 2025. Date of publication 8 October 2025; date of current version 28 October 2025. This work was supported in part by NSFC under Grant 62272059 and Grant U24B20176, in part by Beijing Natural Science Foundation under Grant JQ24020, in part by the National Key Research and Development Program of China under Grant 2023YFF0904800, and in part by Beijing Nova Program under Grant 20230484406. The associate editor coordinating the review of this article and approving it for publication was Dr. Jinli Suo. *(Corresponding author: Huiyuan Fu.)*

Jiaxuan Wang, Huiyuan Fu, Wenkai Zheng, Xicong Wang, and Huadong Ma are with the State Key Laboratory of Networking and Switching Technology and Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: wangjiaxuan2018@bupt.edu.cn; fhy@bupt.edu.cn; ciki@bupt.edu.cn; wangxc@bupt.edu.cn; mhd@bupt.edu.cn).

Xin Wang is with the Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY 11794 USA (e-mail: x.wang@stonybrook.edu).

Heng Zhang is with Xiaomi Company, Beijing 100089, China (e-mail: kazenokizi@live.com).

Digital Object Identifier 10.1109/TIP.2025.3616639

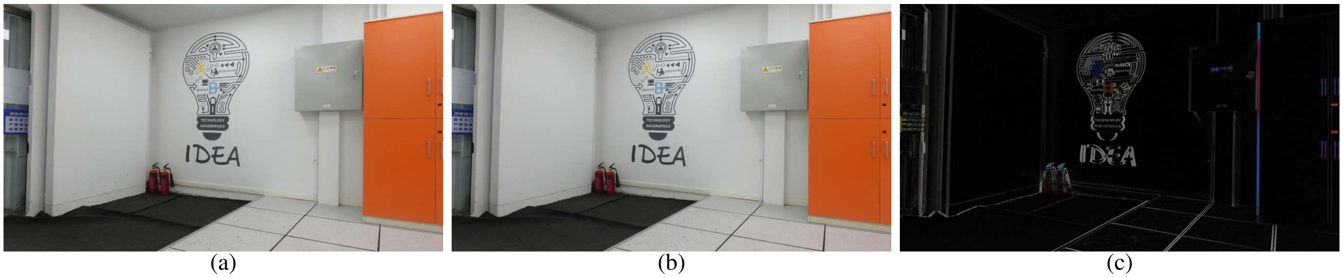


Fig. 2. An example of the difference image. (a) First frame of the dynamic scene. (b) Second frame of the dynamic scene. (c) The difference image of (a) and (b).

Additionally, DID is captured by multiple brands and types of cameras under various lighting conditions, enhancing its generalizability. For temporal alignment, we collect a new version of the dataset called "Real video version of DID" (R-DID). The advantage of R-DID lies in its ability to capture scenes with moving objects and record continuous pairs of videos, which are capabilities that DID lacks.

We propose a Light Adjustable Network (LAN) based on the Retinex theory [1], [39], [40], [41], [42] and an enhanced version of LAN named LAN++ using a novel inter-frame relationship for general low-light video enhancement. By iteratively refining illumination components, LAN generates enhanced results with different luminance levels. Our method is capable of adapting to various scenes by dynamically selecting the degree of light enhancement, which means we have the flexibility to manually adjust the illumination of the results. This approach demonstrates to be good at generalization and can avoid overexposure or underexposure in extreme cases. This is different from previous methods, which were limited to enhancing lighting degradation similar to the training low-light samples. LAN++ adopts a new inter-frame relationship, called difference image, to enhance the details of low-light frames. The difference between two adjacent frames might contain the overall edge and texture of the frame if the scene is dynamic. Fig. 2 shows an example of the difference image. We utilize the two-stage channel attention method to fuse the feature of difference image into the original feature, resulting in sharper and more realistic video frames.

We conducted extensive experiments to demonstrate the generalizability and superiority of our datasets and methods. Fig. 1 shows the results of LAN++ trained on SDSD, DID and R-DID datasets for enhancing low-light videos captured by mobile phones in real scenes.

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

- We build two high-quality paired low-light video datasets: one characterized by pronounced camera motion, strict spatial alignment, and diverse scene content; the other comprises real and continuous videos capable of capturing moving scenes.
- We propose a Retinex-based low-light video enhancement method, named Light Adjustable Network(LAN), which iteratively refines the illumination components and adaptively adjusts the illumination to generate more natural and robust enhanced results. Based on LAN, we propose an improved model named LAN++ to exploit a novel inter-frame relationship through the difference of adjacent images. LAN++ generates sharper and more realistic results with the help of difference images.
- We conduct extensive experiments to validate the effectiveness of our datasets and models compared with state-of-the-art methods.

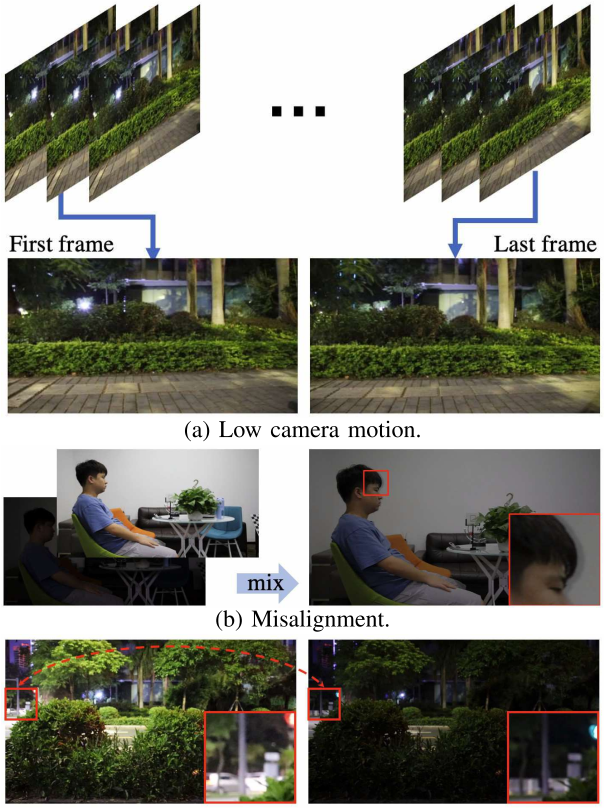
TABLE I
COMPARISON OF OUR DATASET WITH PREVIOUS LOW-LIGHT VIDEO DATASETS

Dataset	Real	Status	Capture Device	Num	Release
SIDGAN [14]	×	Dynamic	-	-	✓
EHSC [10]	✓	Dynamic	Canon 5D Mark III	900	×
SMOID [12]	✓	Dynamic	FLIR GS3-U3-23S6C	35800	×
DRV [11]	✓	Static	Sony RX100 VI	22220	✓
SDSD [13]	✓	Dynamic	Canon 6D Mark II	37500	✓
DID	✓	Dynamic	Sony RX100 M4 Canon EOSR10 Panasonic G9 Fujifilm XT4 Nikon Z5	41038	✓
R-DID	✓	Dynamic	HUAWEI P40 pro	30305	✓

This work is an extension of our previous paper that appears in ICCV2023 [32]. Compared to the conference version, we have introduced a significant amount of new materials: 1) We build a new dataset named R-DID, which is the supplement to the previous DID dataset. Images in DID are captured frame by frame, and can not form continuous videos with scenes of moving objects. R-DID adopts a new capture system with a capturing method different from DID. It shoots low-light and normal-light video pairs simultaneously and uses DeepFlow [30] to align the pair frame by frame. 2) Based on LAN, we build an improved model named LAN++ by utilizing a novel inter-frame relationships, difference images. In dynamic scenes, the difference image contains the overall edge and texture of the frame. LAN++ fuses the information in the difference image into the original frame feature with a two-stage channel attention method. LAN++ outperforms LAN on all datasets. 3) We perform more experiments, analysis and ablation studies on new dataset and model to show the effectiveness of our method over existing state-of-the-art methods.

II. DID AND R-DID DATASET

The effectiveness of low-light enhancement methods depends heavily on the quality of the training dataset. However, acquiring high-quality paired low-light video datasets



(a) Low camera motion.

(b) Misalignment.

(c) Inconsistent scenarios.

Fig. 3. Limited quality of SDSD dataset.

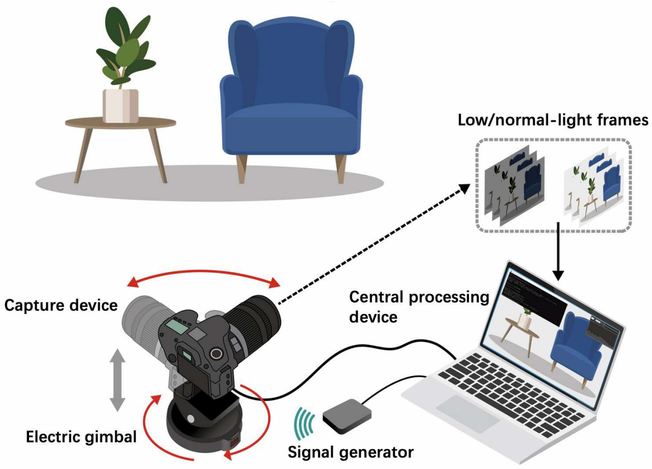


Fig. 4. The camera system of DID.

Algorithm 1 DeepFlow-Based Alignment Procedure

```
Input: normal-light video $V_h$, low-light video $V_l$
Output: Aligned and cropped frame pairs $\{(I_h^{out}, I_l^{out})\}$
Parameters: Target height $h$, target width $w$
for each frame pair $(I_h, I_l)$ in $(V_h, V_l)$ do
    Step 1: Brightness Compensation
    $gray_h \leftarrow \text{RGB2Gray}(I_h)$
    $gray_l \leftarrow \text{RGB2Gray}(I_l)$
    $\mu_h \leftarrow \text{mean}(gray_h)$
    $\mu_l \leftarrow \text{mean}(gray_l)$
    $I_l' \leftarrow \mathcal{C}\left(r^{-1}\left(r(I_l) \cdot (\mu_h/\mu_l)\right)\right)$
    Step 2: DeepFlow Alignment
    $\nabla\psi(I_h) \leftarrow \text{ComputeGradient}(gray_h)$
    Solve for flow field $\mathcal{F}$:
    $\min_{\mathcal{F}} \sum_{ij} \|\nabla\psi(I_h)(i,j) \cdot \mathcal{F}(i,j) +$
    $\partial_t\psi(I_h)(i,j)\|^2 + \lambda\|\nabla\mathcal{F}\|^2$
    $I_l^{align}(x,y) \leftarrow \sum_{m=\lfloor x'\rfloor}^{\lceil x'\rceil} \sum_{n=\lfloor y'\rfloor}^{\lceil y'\rceil} I_l'(m,n)$
    $\prod_{d\in\{x',y'\}}(1-|d-\lfloor d\rfloor|)$
    where $(x',y') = (x,y) + \mathcal{F}(x,y)$
    Step 3: Center Cropping
    $I_h^{out} \leftarrow I_h\left[\frac{H-h}{2} : \frac{H+h}{2}, \frac{W-w}{2} : \frac{W+w}{2}\right]$
    $I_l^{out} \leftarrow I_l^{align}\left[\frac{H-h}{2} : \frac{H+h}{2}, \frac{W-w}{2} : \frac{W+w}{2}\right]$
    Output $(I_h^{out}, I_l^{out})$
end for
```

poses challenges. Even if two videos with different lighting conditions are captured in the same scene, it is difficult to ensure them to follow the same motion trajectory. Although synthetic paired video data can be generated [14], training deep models on such synthetic data may introduce artifacts and color biases when applied to real low-light videos due to the disparity between synthetic and real-world data [9]. As shown in Table I, recently released real-captured paired low-light video datasets like EHSC [10], SMOID [12] and DRV [11] have certain limitations, such as delayed release or static video content, while the video quality of datasets like SDSD [13] may be constrained. Additionally, relying solely on videos from a camera with a consistent non-linear response can substantially degrade model performance when faced with videos from cameras with unknown response functions [15]. Models trained on these datasets often fall short in real-world application scenarios, highlighting the urgent need for a comprehensive, high-quality low-light video dataset.

A. Camera System of DID

To create a comprehensive high-quality low-light video dataset, we have developed a camera system that comprises five brands of capture devices, an electric gimbal, a signal generator, and a central processing device. This system captures paired video datasets by taking frames sequentially. In the process illustrated in Fig. 4, the central processing device adjusts the camera ISO settings to capture a series of low-light frames and corresponding normal-light frames at the same location. Subsequently, these frames are transmitted to the central processing device. The device evaluates the quality of the frames and, if they meet the predefined standards, synthesizes them into final low-light and normal-light frames. If the quality is not satisfactory, a signal is sent to recapture the frames until the desired quality is achieved. After capturing a frame pair, the signal generator activates the electric gimbal to make slight movements. To ensure the continuity between adjacent frames, we impose a constraint where the total sum of horizontal and vertical rotation angles of the electric gimbal is limited to less than 1° each time.

B. Camera System of R-DID

To obtain a real-world paired low-light video datasets, we design a camera system consisting of a smartphone holder and

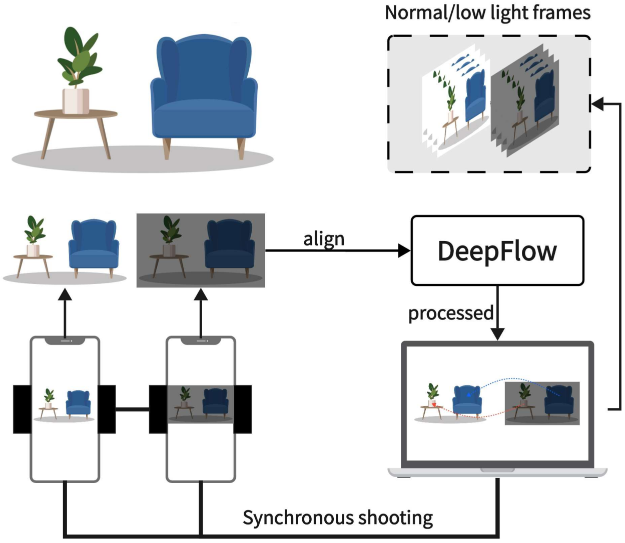


Fig. 5. The camera system of R-DID.

Algorithm 2 Calculate Optical Flow

```
Input: video $V$
Output: Optical flow
Parameters: $H, W$ (The height and width after resize), Farneback optical flow parameters
$meanflow_{norm} \leftarrow 0$;
$count \leftarrow 0$;
for Adjacent frames ($F_{pre}$, $F_{nxt}$) of $V$ do
    $F_{pre} \leftarrow$ RGB2GRAY $\left(\text{Resize}\left(F_{pre}, (H, W)\right)\right)$;
    $F_{nxt} \leftarrow$ RGB2GRAY (Resize $(F_{nxt}, (H, W))$);
    $flow \leftarrow$ calcOpticalFlowFarneback($F_{pre}, F_{nxt}$);
    $flow_{norm} \leftarrow$ norm($flow, axis = -1$);
    $meanflow_{norm} \leftarrow meanflow_{norm}$
$+$mean($flow_{norm}$);
    $count \leftarrow count + 1$;
end for
$meanflow_{norm} \leftarrow meanflow_{norm}/count$;
return $meanflow_{norm}$
```

Algorithm 3 Calculate LOE

```
Input: low-light video $V_l$ and normal-light video $V_h$
Output: LOE
Parameters: $H, W$ (The height and width after resize), $win$ (Window size)
$LOE \leftarrow []$;
for each frame ($F_l$, $F_h$) of $V_l$ and $V_h$ do
    $F_l \leftarrow$ RGB2GRAY (Resize $(F_l, (H, W))$);
    $F_h \leftarrow$ RGB2GRAY (Resize $(F_h, (H, W))$);
    $LOE_{frame} \leftarrow []$;
    for $x \leftarrow 0$to$w - 1$Step$win$ do
        for $y \leftarrow 0$to$h - 1$Step$win$ do
            $RD \leftarrow 0$;
            for $win_x \leftarrow 0$to$win - 1$ do
                for $win_y \leftarrow 0$to$win - 1$ do
                    $E \leftarrow (F_l[x + win_x, y + win_y] > F_l[x : x + win, y : y + win]) \oplus (F_h[x + win_x, y + win_y] > F_h[x : x + win, y : y + win])$;
                    $RD \leftarrow RD +$ sum($E$);
                end for
            end for
            $LOE_{frame}$.append($RD/(win \times win)$);
        end for
    end for
    $LOE$.append(mean($LOE_{frame}$));
end for
return mean($LOE$)
```

two smartphones. Unlike the frame-by-frame shooting method of DID, R-DID takes continuous and real videos. As shown in the Fig. 5, we place two HUAWEI P40 Pro smartphones side by side on a smartphone holder, ensuring that they are in the same plane with minimal distance between their cameras. We use a central controlling device to simultaneously start and stop video recording on both smartphones, adjusting the ISO settings before capturing to ensure that one smartphone records a normal-light video and the other records a low-light video. Despite the close proximity of the cameras, the slight angular deviation in the captured low/normal-light videos requires a further processing for alignment. We perform frame-by-frame alignment for each pair of low/normal-light videos. Algorithm 1 provides the pseudo-code of alignment, mainly consisting of three steps. For each pair, we convert them into grayscale images and increase the average brightness of the low-light frame to match that of the normal-light frame, reducing the impact of brightness differences on subsequent image alignment. Next, we employ DeepFlow [30] to compute the flow mapping from the low-light frame to the normal-light frame. We apply the mapping to the low-light frame and crop the resulting aligned bright-dark frame to remove black borders, yielding a set of aligned low/normal-light frames. The aforementioned alignment operations are applied to every frame of the low/normal-light videos, resulting in aligned low/normal-light video pairs.

C. DID 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. 6. We maintain some reasonable fluctuations in video brightness and illumination intensity within the dataset to enhance the model's generalization capability. In Fig. 7, we give two samples of different scenarios in our dataset.

In order to conduct a quantitative comparison between our dataset and existing low-light video datasets, we have introduced two assessment metrics: lightness-order-error (LOE) [28] and optical flow [29]. LOE evaluates the relative order of lightness in different local areas to assess the alignment of paired low-light and normal-light frames. Optical flow measures pixel motion between consecutive images to capture the dynamics of a video. Algorithm 2 and Algorithm 3 provide the pseudo-code to calculate the optical flow and LOE.

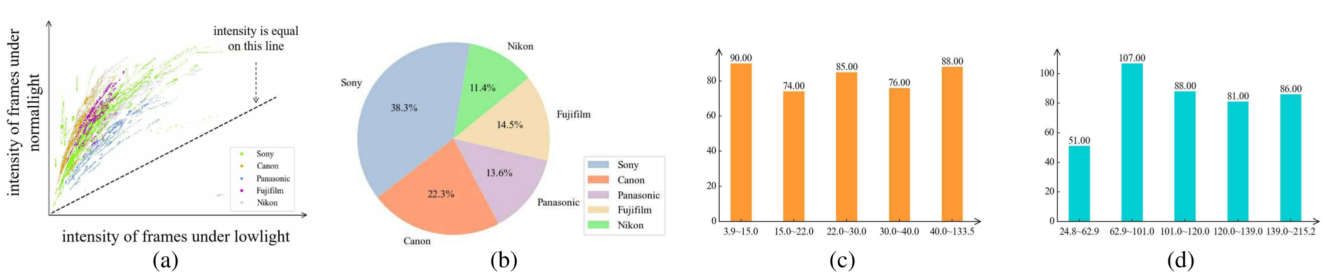


Fig. 6. Statistical indicators for our DID dataset. (a) Intensity distribution for low/normal-light videos. (b) Distribution of videos captured by five different brand cameras. (c) Luminance distribution for low-light videos. (d) Luminance distribution for normal-light videos.

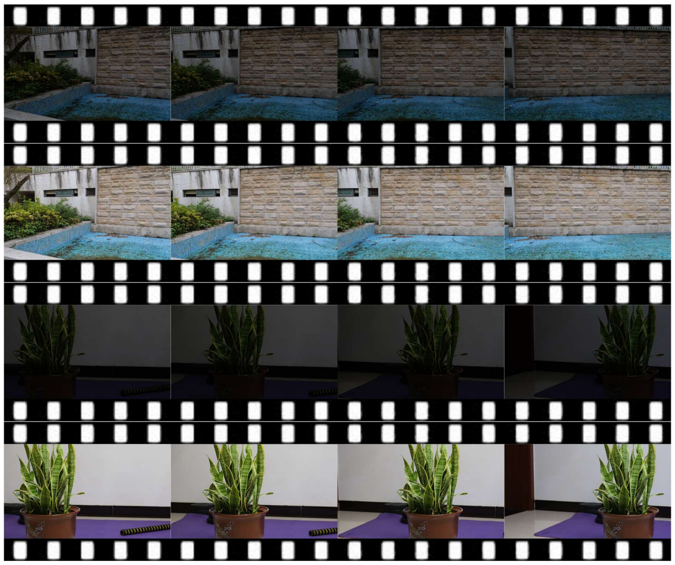


Fig. 7. Two example videos of our DID dataset.

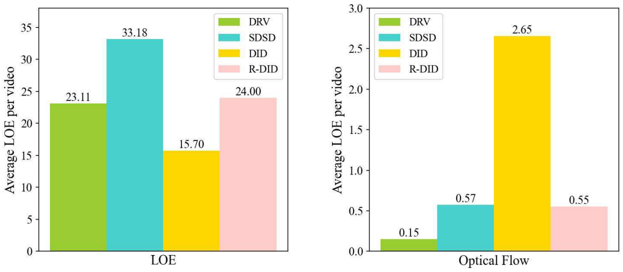


Fig. 8. LOE and optical flow of different datasets.

As shown in the Fig. 8, our dataset exhibits the lowest LOE, indicating excellent alignment in our paired videos. Furthermore, the optical flow measurement of our dataset surpasses that of DRV [11] and SDSD [13], suggesting more pronounced motion in our videos compared to the relatively static content in DRV and SDSD. Therefore, our dataset contains better aligned results and a larger range of scenes.

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.

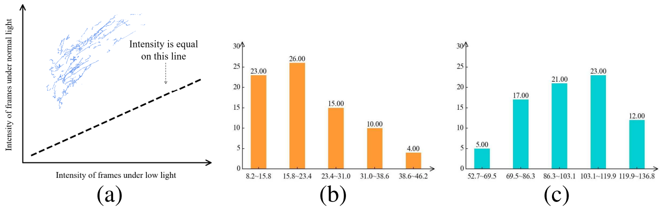


Fig. 9. Statistical indicators for our R-DID dataset. (a) Intensity distribution for low/normal-light videos. (b) Luminance distribution for low-light videos. (c) Luminance distribution for normal-light videos.

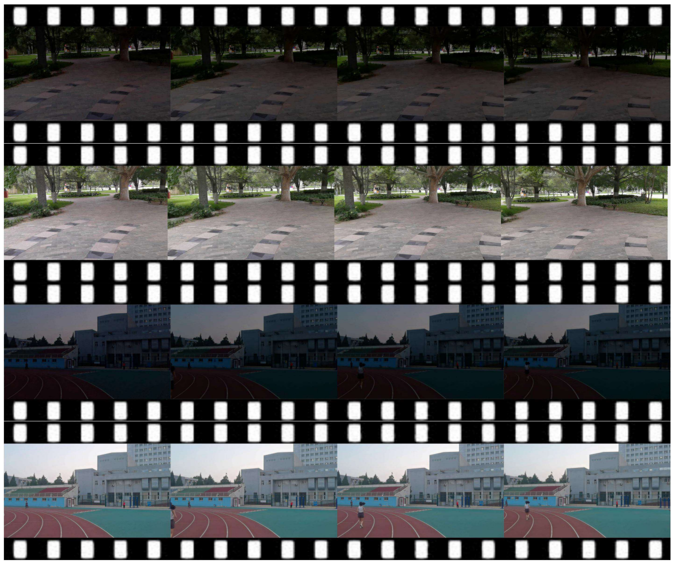


Fig. 10. Two Example Videos Of Our R-Did Dataset.

D. R-DID Data

We collected 78 paired videos with a total of 30305 frames and named them as R-DID dataset ("R" stands for "Real video"). The resolution of our videos is 1280×720, and more statistical indicators of the overall dataset are shown in Fig. 9.

TABLE II
COMPARISONS OF DID AND R-DID

Dataset	Alignment	Video Format	Advantage
DID	Physical alignment	Frame-synthesized video	Low noise. High alignment quality.
R-DID	DeepFlow optical flow method	Real, continuous video	Non-synthetic video. Dynamic scenes.

In Fig. 10, we give two samples of different scenarios in our dataset.

Table II shows the differences of DID and R-DID. DID's alignment quality significantly surpasses all existing video datasets, which is crucial for models to learn the degradation simulation process in low-light enhancement and evaluate model performance. Although DID is a high-quality dataset, it is captured frame-by-frame and therefore does not represent real videos. The scenes in DID are limited to static scenarios, and the synthetic videos do not include motion blur and other phenomena existing in real videos. Our motivation of collecting R-DID is to overcome these limitations of DID. R-DID consists of real videos that contain a large number of dynamic scenes. Models trained on R-DID have better adaptability for applications involving multiple dynamic scenes.

III. METHOD

Enhancing low-light images or videos is a challenging task due to the inherent ill-posed nature of this problem. As with many inverse problems, a single low-light input can correspond to multiple suitable normal-light outputs. Although previous low-light enhancement methods have generated results that are close to ground truth, they typically use one-to-one models to generate fixed outputs for a given input, which limits their generalization performance. In real-world applications, the distribution of low-light samples may differ from existing low-light datasets and may even include extremely dark or slightly dark scenes. Therefore, simply fitting the light degradation of the training data may lead to suboptimal solutions and often results in overexposed or underexposed enhanced images.

To address these challenges, we propose the Light Adjustable Network (LAN) for general low-light video enhancement. Our approach adaptively adjusts the illumination to generate appropriately exposed results and enables users to adjust the light intensity to produce different outputs. This adaptive feature makes our model more robust and versatile, allowing it to perform well on a wider range of low-light scenarios.

In addition, our preliminary studies indicate that the inter-frame relationship across video frames can benefit low-light video enhancement and is overlooked in the literature work. Compared to LAN, we employ a novel inter-frame relationship called the difference image in LAN++. The difference image is the absolute difference of each pixel between adjacent frames. For dynamic scenes, the difference image can clearly display the edge and texture of the current video frame. By incorporating the features of the difference image into the original features, it is beneficial for generating sharper and more realistic normal-light video frames.

A. Light Adjustable Network

Fig. 11 (a) illustrates the framework of our proposed LAN++. According to Retinex theory [1], we decompose the input video frames $X_{t+i,i\in[-k,k]}$, into reflectance components R_t and illumination components I_t [37], and then enhance the illumination components through iterative refinement, and finally synthesize them into normal light frames $\hat{Y}_t$. In the process of estimating reflectance components R_t, we integrate the corresponding difference image feature into the current frame features to supplement the edge and texture information. The definition of the difference image is shown below:

$$D_t\left(x, y, c\right) = \left|F_t(x, y, c) - F_{t+1}(x, y, c)\right| \quad (1)$$

where x, y, c represent the the horizontal, vertical and channel coordinates of the pixel.

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 [2]. Meanwhile, we calculate the difference image $D_{t+i} \in \mathbb{R}^{H\times W\times 3}$ and project them as embedding $D_t^0 \in \mathbb{R}^{H\times W\times C}$. Then, the reflectance estimation module estimates the reflectance R_t from it. The reflectance estimation module is a dual branch hierarchical structure. One branch handles video frame features and the other handles difference image features. Two branches have the same structure with four stages, each consisting of a feature extraction block and a down-sampling layer. The components of the feature extraction block are shown in Fig. 11 (a), and the block design is inspired by [4], [5], and [6]. For the i-th stage, the feature map $F_t^{i-1}, D_t^{i-1} \in \mathbb{R}^{\frac{H}{2^{i-1}}\times\frac{W}{2^{i-1}}\times 2^{i-1}C}$ will be processed to obtain the feature map $F_t^i, D_t^i \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times 2^iC}$. Before the down-sampling layer at each stage, the video frame features and difference image features are input into a fusion module, which will be described in detail in Section III-B. The output feature map of the fusion module has the same size as the input, and the output result is fed into the down-sampling layer of the video frame branch. 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 pre-trained encoder, and then enhances the illumination component with iterative refinement, which will be described in detail in Section III-C.

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 also has a hierarchical architecture with 4 stages, each 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 $\hat{Y}_t$.

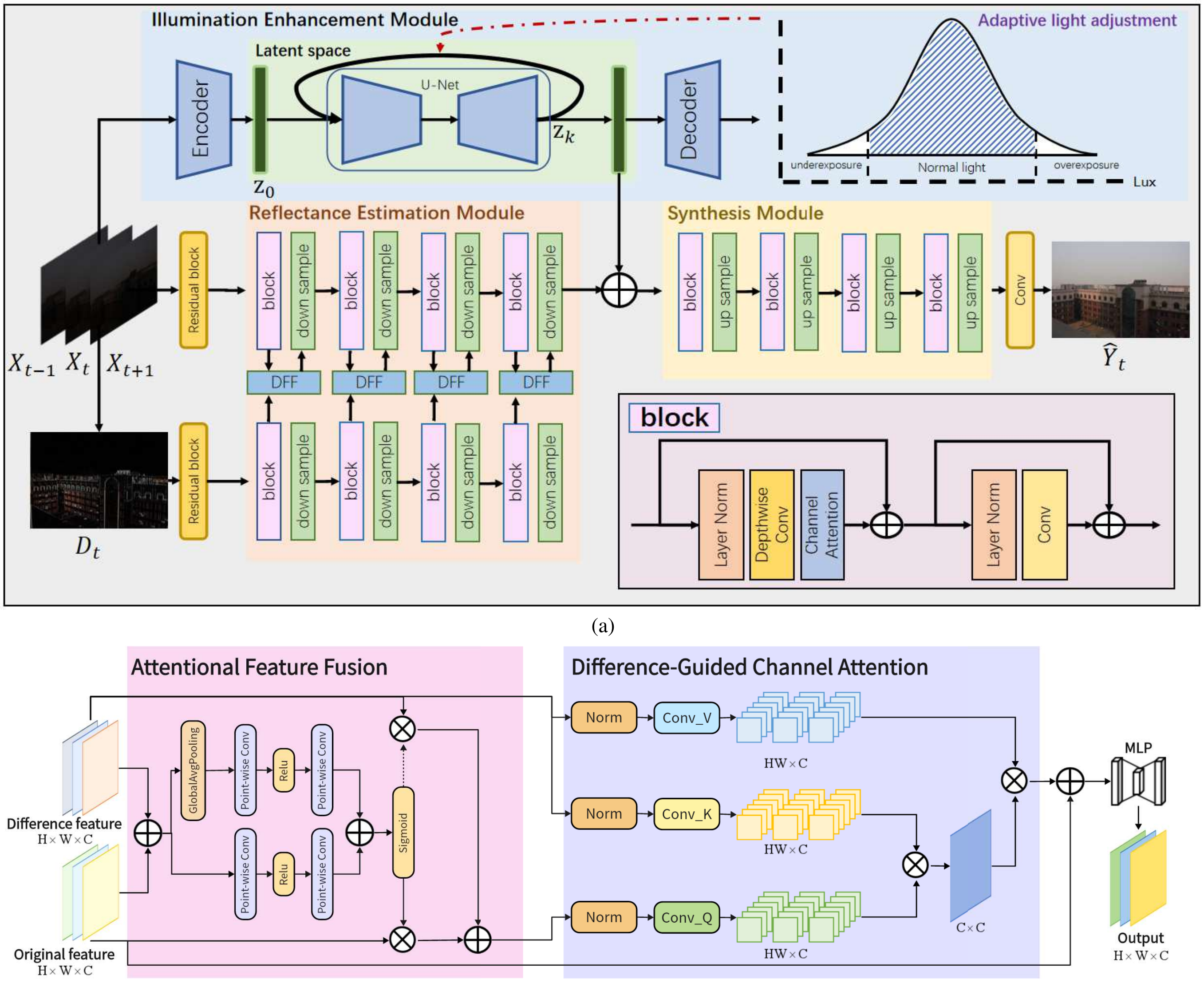


Fig. 11. Overview of our method. (a) The framework of our LAN++ and the structure of block module. (b) The Difference feature fusion (DFF) module.

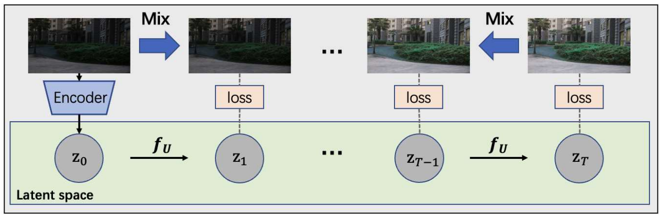


Fig. 12. The process of iterative illumination refinement.

B. Difference Feature Fusion

When enhancing extremely dark video frames, many methods generate blurry and less textured images. To address this issue, we employ a difference feature fusion module that integrates the features of the difference image into the video frame features, leveraging prior features to supplement high-frequency information from the original features. As shown in Fig. 11 (b), the Difference Feature Fusion module (DFF) consists of two stages, with the Attentional Feature Fusion module (AFF) followed by the Difference-Guided Channel Attention module (DGCA).

Inspired by [3] and [31], AFF incorporates the original features $F_t^i \in \mathbb{R}^{H\times W\times C}$ and the differential features $D_t^i \in \mathbb{R}^{H\times W\times C}$ to generate mixed features that enhance the texture. Using point-wise convolutions, we calculate the global channel attention w_g and the local channel attention of the mixed feature w_l to obtain the mixed feature weight w. The weights of the original feature and difference feature are set to w and $1-w$, respectively, and a weighted sum is carried out to produce the output of the first stage $M_t^i \in \mathbb{R}^{H\times W\times C}$. The formula of the M_t^i is shown below:

$$\begin{aligned} w &= w_l + w_g \\ M_t^i &= F_t^i \cdot w + D_t^i \cdot (1-w) \end{aligned} \tag{2}$$

In the second stage, DGCA takes the output from the first stage $M_t^i \in \mathbb{R}^{H\times W\times C}$ and the difference feature $D_t^i \in \mathbb{R}^{H\times W\times C}$ as inputs. DGCA generates Q vector from M_t^i and K,V vectors

TABLE III
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON SDSD DATASET

Methods	Learning	SDSD			
		PSNR ↑	SSIM ↑	LPIPS ↓	MUSIQ ↑
DRBN [26]	Image	22.31	0.65	0.30	37.46
RUAS [27]	Image	15.48	0.64	0.35	47.12
LLFlow [23]	Image	24.90	0.78	0.22	38.08
SNR-Aware [24]	Image	25.27	0.82	0.20	38.17
SCI [25]	Image	16.90	0.64	0.28	42.91
SMG [33]	Image	26.12	0.79	0.15	37.36
MBLLEN [21]	Video	21.79	0.65	0.20	45.98
SMID [11]	Video	24.09	0.69	0.35	34.35
SDSDNet [13]	Video	24.92	0.73	0.16	39.89
Chhirolya et al. [20]	Video	23.46	0.79	0.20	37.58
SGLLIE [34]	Video	23.89	0.70	0.45	35.28
LAN (default T)	Video	26.95	0.85	0.16	47.96
LAN (adaptive)	Video	27.25	0.85	0.14	48.35
LAN++ (default T)	Video	26.64	0.86	0.13	53.60
LAN++ (adaptive)	Video	**28.03**	**0.88**	**0.11**	**54.12**

from D_t^i. The input features are normalized and processed through convolution layers and reshaping operations to obtain Q,K,V vectors [7] of size (hw, c). The Q vector is multiplied by the K vector to compute the difference feature attention matrix. The result of matrix multiplication between the attention matrix and V vector, along with the original features, is fed into an MLP network through skip connections to produce the output result $FD_t^i \in \mathbb{R}^{H \times W \times C}$. The formula of the FD_t^i is shown below:

$$\begin{aligned} FD_t^i &= F_t^i + \Phi(Q, K, V) \\ &= F_t^i + VSoftmax\left(\frac{Q^{\mathrm{T}}K}{\sqrt{c}}\right) \end{aligned} \tag{3}$$

where c is the number of channel dimension.

C. 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. 12, 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 $(z_1, z_2, \ldots, z_{T-1}, z_T)$. 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:

$$\begin{aligned} \alpha_k &= \frac{k}{T} \\ \tilde{z}_k &= \mathcal{E}\left(\alpha_k \cdot Y_t + (1 - \alpha_k) \cdot X_t\right) \end{aligned} \tag{4}$$

where $\mathcal{E}$ denotes the representation of the pre-trained encoder. We use a U-Net [8] 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.

D. 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,

$$l \sim \mathcal{N}\left(\mu, \sigma^2\right) \tag{5}$$

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 $(z_{k-2n}, z_{k-2n+1}, \ldots, z_k)$, where n is a pre-defined interval parameter. We use a one-sided student's test, which compares the mean intensity $\bar{l}_k$ of $(z_{k-2n}, z_{k-2n+1}, \ldots, z_k)$ 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:

$$\begin{aligned} H_0 &: \bar{l}_k \geq \mu \\ H_1 &: \bar{l}_k < \mu \end{aligned} \tag{6}$$

The test statistic is calculated as:

$$t = \frac{\bar{l}_k - \mu}{s_k}\sqrt{2n+1} \tag{7}$$

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

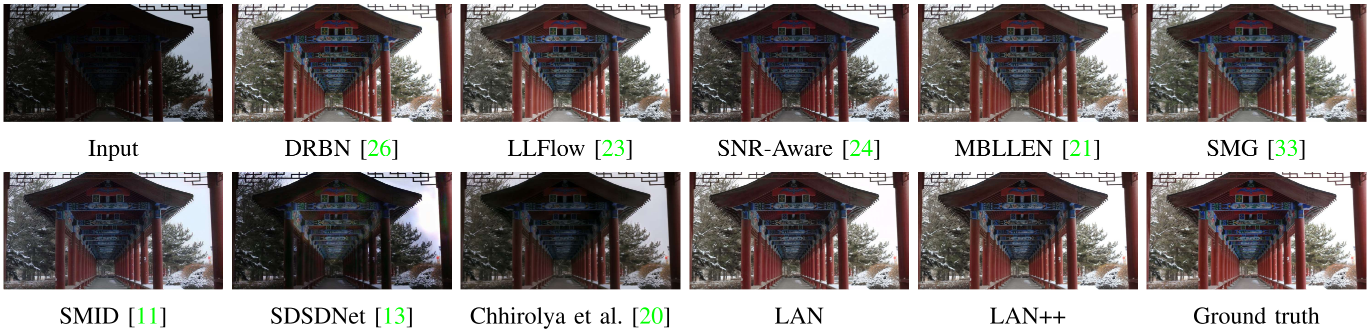
The rejection region is:

$$t < t_{\alpha,2n} \tag{8}$$

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).


Input DRBN [26] LLFlow [23] SNR-Aware [24] MBLLEN [21] SMG [33]

SMID [11] SDSDNet [13] Chhirolya et al. [20] LAN LAN++ Ground truth

Fig. 13. Visual comparison with state-of-the-art low-light enhancement methods on DID dataset.

TABLE IV

QUANTITATIVE RESULTS OF DIFFERENT METHODS ON DID DATASET

Methods	Learning	DID			
		PSNR ↑	SSIM ↑	LPIPS ↓	MUSIQ ↑
DRBN [26]	Image	25.22	0.91	0.09	65.59
RUAS [27]	Image	17.01	0.74	0.24	55.28
LLFlow [23]	Image	25.71	0.92	0.09	66.51
SNR-Aware [24]	Image	24.05	0.90	0.08	66.44
SCI [25]	Image	11.15	0.44	0.53	64.56
SMG [33]	Image	23.58	0.85	0.08	66.22
MBLLEN [21]	Video	24.82	0.91	0.09	66.47
SMID [11]	Video	22.97	0.87	0.18	54.05
SDSDNet [13]	Video	21.88	0.83	0.11	64.31
Chhirolya et al. [20]	Video	22.77	0.88	0.13	62.51
SGLLIE [34]	Video	20.48	0.84	0.19	60.39
LAN (default T)	Video	27.28	0.92	0.08	64.96
LAN (adaptive)	Video	29.01	0.94	0.07	65.05
LAN++ (default T)	Video	27.80	0.93	0.07	69.47
LAN++ (adaptive)	Video	**30.17**	**0.95**	**0.06**	**69.59**

TABLE V

QUANTITATIVE RESULTS OF DIFFERENT METHODS ON R-DID DATASET

Methods	Learning	R-DID			
		PSNR ↑	SSIM ↑	LPIPS ↓	MUSIQ ↑
DRBN [26]	Image	24.33	0.85	0.23	43.87
RUAS [27]	Image	13.23	0.65	0.39	29.63
LLFlow [23]	Image	24.10	0.82	0.29	31.21
SNR-Aware [24]	Image	25.58	0.83	0.25	47.78
SCI [25]	Image	17.22	0.68	0.42	32.14
SMG [33]	Image	25.30	0.80	0.09	63.10
MBLLEN [21]	Video	24.26	0.84	0.22	63.33
SMID [11]	Video	20.58	0.76	0.33	35.02
SDSDNet [13]	Video	19.12	0.57	0.45	26.34
Chhirolya et al. [20]	Video	24.59	0.84	0.18	55.19
SGLLIE [34]	Video	20.49	0.75	0.29	30.93
LAN (default T)	Video	24.83	0.81	0.12	66.58
LAN (adaptive)	Video	26.16	0.81	0.11	66.91
LAN++ (default T)	Video	26.18	0.82	0.09	66.74
LAN++ (adaptive)	Video	**26.73**	**0.82**	**0.09**	**67.06**

TABLE VI

QUANTITATIVE RESULTS OF ABLATION STUDY ON SDSD TESTSET

Model	PSNR	SSIM
No difference feature	27.25	0.85
Only AFF in fusion	27.56	0.87
Only DGCA in fusion	27.75	0.87
AFF and DGSA in fusion	27.76	0.87
No Iterative illumination refinement	26.32	0.87
No adaptive light adjustment	26.64	0.86
LAN++	28.03	0.88

E. Training and Loss Function

First, we follow [16] to train a VAGAN [17], [18] to encode the input frames into latent space. Then we train the illumination enhancement module. The illumination enhancement module functions to adjust the illumination intensity of generated results. During its independent training phase, the training data for the illumination enhancement module is generated through proportional mixing of low-light input and ground truth. Specifically, we create $T-1$ intermediate-state images by blending input and ground truth in different ratios, forming a latent representation sequence with progressively increasing illumination intensity $(\tilde{z}_0, \tilde{z}_1, \ldots, \tilde{z}_{T-1}, \tilde{z}_T)$. We use adjacent pairs $(\tilde{z}_k, \tilde{z}_{k+1})$ in this sequence as input-output pairs for module training, enabling the U-Net within the module to learn the refined low-light enhancement process. During subsequent full-model training, we freeze all parameters in the illumination enhancement module and set the iteration number to T. This configuration allows flexible adjustment of illumination intensity by modifying the U-Net's iteration number during inference. 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:

$$\mathcal{L}_U = \sqrt{\|f_U(\tilde{z}_k) - \tilde{z}_{k+1}\|_F^2 + \epsilon^2} \tag{9}$$

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.

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 a reflectance consistency loss as follows:

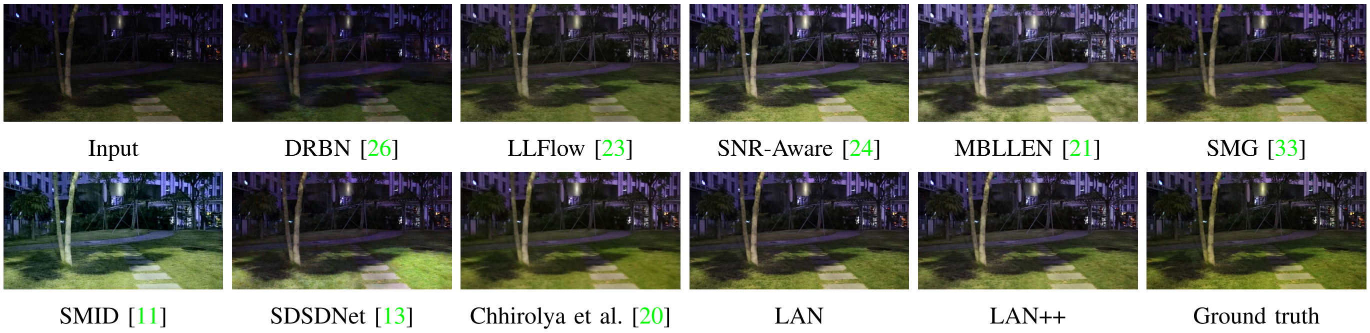
$$\mathcal{L}_R = \|f_R(X_t) - f_R(Y_t)\|_F^2, \tag{10}$$

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

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

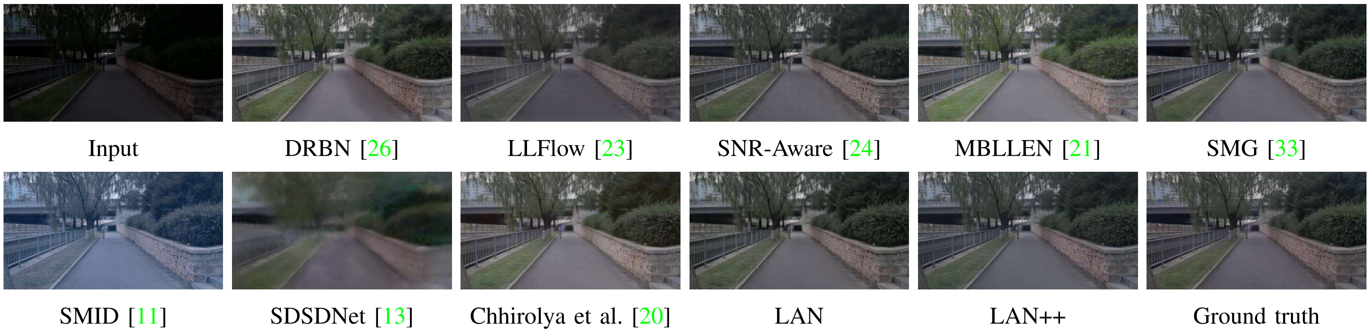
$$\begin{aligned} \mathcal{L} = (1-\lambda)\ &\sqrt{\|\hat{Y}_t - Y_t\|_F^2 + \epsilon^2} \\ &+ \lambda \mathcal{L}_{\text{SSIM}}\left(\hat{Y}_t, Y_t\right) + \frac{\mathcal{L}_R}{\tau} \end{aligned} \tag{11}$$

where λ is a trade-off parameter, $\mathcal{L}_{\text{SSIM}}$ represents the structural similarity loss [19], and τ denotes a temperature parameter.


Input DRBN [26] LLFlow [23] SNR-Aware [24] MBLLEN [21] SMG [33]

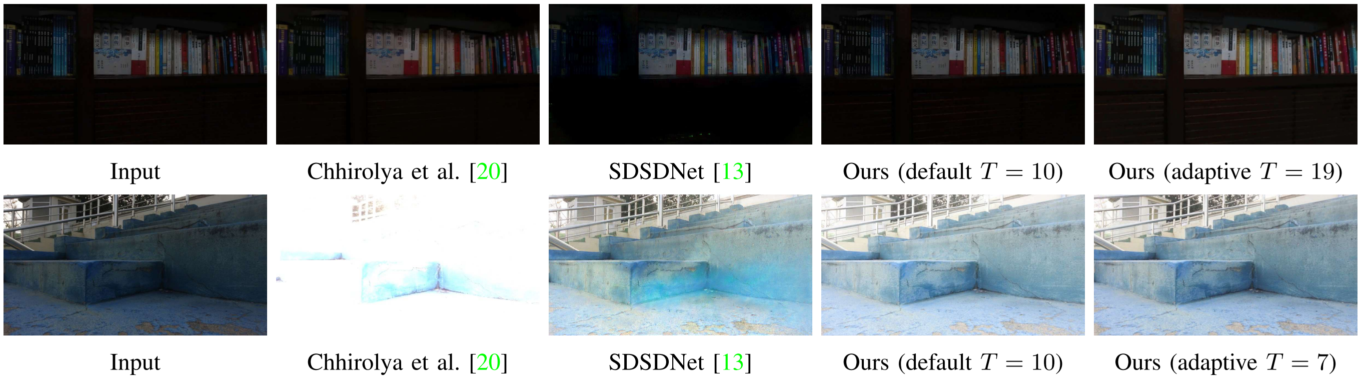
SMID [11] SDSDNet [13] Chhirolya et al. [20] LAN LAN++ Ground truth

Fig. 14. Visual comparison with state-of-the-art low-light enhancement methods on SDSD dataset.


Input DRBN [26] LLFlow [23] SNR-Aware [24] MBLLEN [21] SMG [33]

SMID [11] SDSDNet [13] Chhirolya et al. [20] LAN LAN++ Ground truth

Fig. 15. Visual comparison with state-of-the-art low-light enhancement methods on R-DID dataset.


Input Chhirolya et al. [20] SDSDNet [13] Ours (default $T = 10$) Ours (adaptive $T = 19$)

Input Chhirolya et al. [20] SDSDNet [13] Ours (default $T = 10$) Ours (adaptive $T = 7$)

Fig. 16. Visual comparison on extremely dark and slightly dark videos.

IV. Experiments

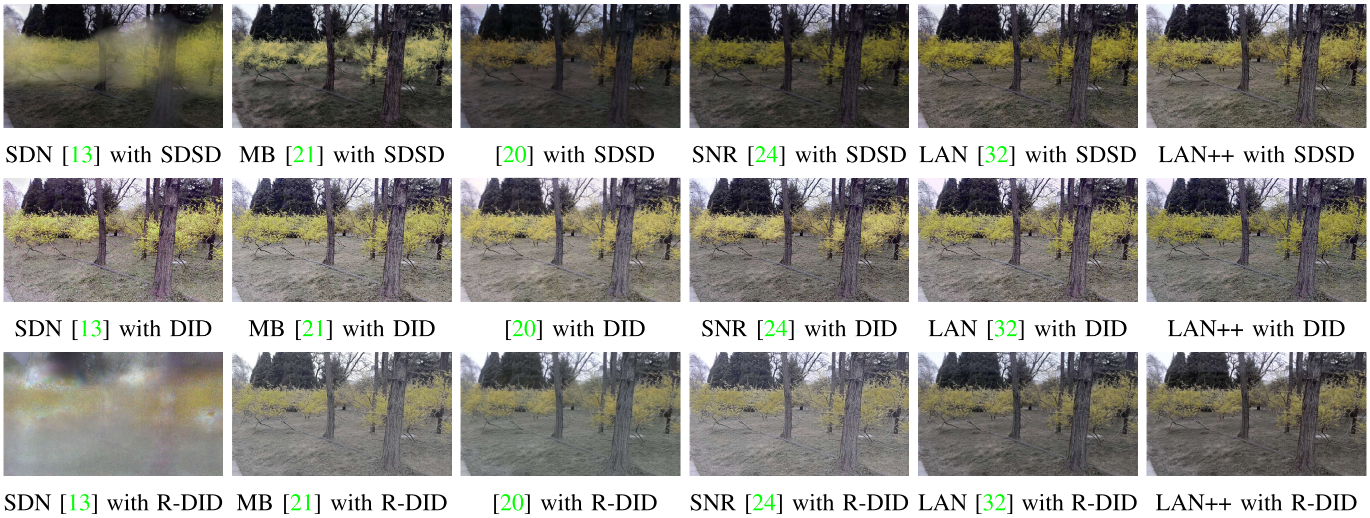
A. Implementation Details

We show the superiority of our proposed approach and the effect of our constructed DID and R-DID through experiments in this section. To evaluate the effect of our method, we retrain 11 previous representative methods on the DID, R-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 and R-DID datasets 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 and LAN++. We augment the data using rotation and horizontal flipping and optimize the network by AdamW optimizer [22] 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 and LAN++ for 200 epochs on a Tesla A100 GPU.

B. Quantitative Evaluation

To comprehensively evaluate the effectiveness of our proposed method, we conduct quantitative experiments on paired video datasets captured under various scenes, including the R-DID, DID and SDSD datasets. Specifically, we evaluate the performance of our method on the test datasets of R-DID and DID, which comprise 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 LAN and LAN++ 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.

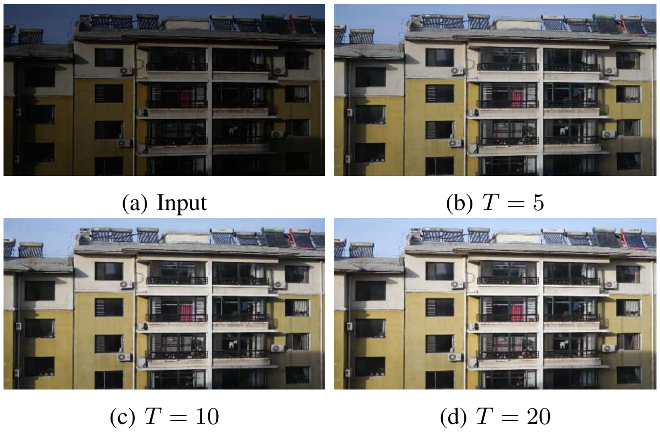


SDN [13] with SDSD MB [21] with SDSD [20] with SDSD SNR [24] with SDSD LAN [32] with SDSD LAN++ with SDSD

SDN [13] with DID MB [21] with DID [20] with DID SNR [24] with DID LAN [32] with DID LAN++ with DID

SDN [13] with R-DID MB [21] with R-DID [20] with R-DID SNR [24] with R-DID LAN [32] with R-DID LAN++ with R-DID

Fig. 17. Visual results in the user study.



(a) Input (b) $T = 5$

(c) $T = 10$ (d) $T = 20$

Fig. 18. Visual comparison of enhanced results of different iterations T. The light becomes stronger as the number of iterations increases.

TABLE VII

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 [21]	33.3%	26.7%	40.0%
SDSDNet [13]	16.7%	13.3%	70.0%
Chhirolya et al. [20]	21.7%	10.0%	68.3%
SNR-Aware [24]	26.7%	15.0%	58.3%
LAN [32]	35.0%	18.3%	46.7%

TABLE VIII

THE RESULTS OF DIFFERENT DATASETS IN THE USER STUDY. "SDSD" IS THE PERCENTAGE THAT THE SDSD DATASET IS PREFERRED, "DID" IS THE PERCENTAGE THAT DID DATASET IS PREFERRED, "R-DID" IS THE PERCENTAGE THAT THE R-DID DATASET IS PREFERRED

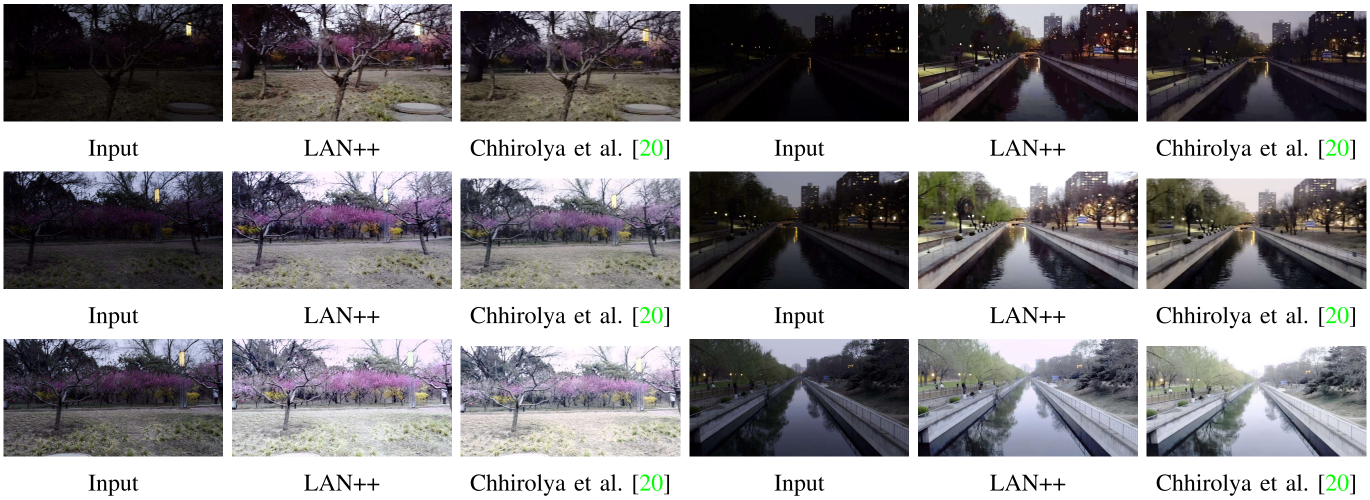
Methods	SDSD	DID	R-DID
MBLLEN [21]	20.0%	42.5%	37.5%
SDSDNet [13]	33.3%	53.3%	13.3%
Chhirolya et al. [20]	21.7%	46.7%	31.7%
SNR-Aware [24]	16.7%	36.7%	46.7%
LAN [32]	25.0%	48.7%	26.3%
LAN++	15.7%	51.0%	33.3%

We adopt four well-known objective evaluation metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM) [19], Learned Perceptual Image Patch Similarity (LPIPS) [35] and Multi-scale Image Quality Transformer (MUSIQ) [36]. 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. LPIPS is a method for measuring image similarity, which evaluates the perceptual difference between two images using deep learning models. Compared to traditional methods, LPIPS is more in line with human perception. MUSIQ is a Transformer-based multi-scale metric for assessing image quality. Unlike the previous three metrics, it is a no-reference image quality assessment method.

Table III, Table IV and Table V present the quantitative evaluation results of different methods on SDSD, DID and R-DID datasets. As shown in the table, our proposed improved version of 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 4dB on DID, more than 1.9dB on SDSD and more than 1dB on R-DID 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.

C. Qualitative Evaluation

We perform thorough qualitative evaluations on the R-DID, DID and SDSD datasets to assess the performance of our proposed method. Fig. 13 presents the results obtained on the DID dataset, where it is observed that SNR-aware, SMG, SMID and the method proposed by Chhirolya et al. produce images with a darker tone, leading to substantial color deviation. Moreover,



Input LAN++ Chhirolya et al. [20] Input LAN++ Chhirolya et al. [20]

Input LAN++ Chhirolya et al. [20] Input LAN++ Chhirolya et al. [20]

Input LAN++ Chhirolya et al. [20] Input LAN++ Chhirolya et al. [20]

Fig. 19. Performance in different degrees of low light conditions.

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. Compared to LAN, LAN++ has a clearer texture.

Fig. 14 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, SMG and SMID result in significant color deviation, which adversely affects the visual quality of the images. Similarly, LAN++ performs better than LAN in terms of texture.

Fig. 15 presents the visual results obtained on the R-DID dataset, which consists of real, continuous video in dynamic scenes containing moving objects. Similar to the previous result, the enhanced images generated from Chhirolya et al., LLFlow, SMG and SDSDNet are blurry, especially SDSDNet, with complete loss of texture. DRBN yields images with unnatural shadows. SNR-aware's result has obvious gridded artifacts. LAN++ and LAN perform relatively well, and the results produced by LAN++ were sharper than LAN.

Fig. 16 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 adaptive light adjustment strategy. Fig. 18 shows the impact of the number of iterations on the enhanced results.

After a comprehensive evaluation of the comparative results of different methods on the three datasets, our proposed method demonstrates excellent visual performance in terms of global brightness, color recovery and details.

D. Ablation Study

LAN++ improves the performance of the low-light video enhancement, to validate the effectiveness of modules, we conduct ablation study on SDSD dataset. The results are shown in Table VI.

To validate the effectiveness of the difference image feature and the difference feature fusion module, we train three models: one without fusing the difference feature, one with only AFF fusing the difference feature, and one with only DGCA fusing the difference feature. The results show that the difference image feature and the difference feature fusion module, which includes AFF and DGCA, greatly contribute to improving the PSNR of the images generated by the model. We also conduct ablation experiments using Difference-Guided Spatial Attention module (DGSA) instead of DGCA. DGSA divides features into 8×8 patches and applies spatial attention mechanism. The results showed slightly inferior performance compared to LAN++, while significantly increasing GPU memory requirements.

To validate the effectiveness of the iterative illumination refinement module, we trained two models: one without the iterative illumination refinement module and one without adaptive light adjustment. The results show that the adaptive iterative illumination refinement module improves the PSNR by more than 1dB, demonstrating its effectiveness.

E. User Study

To compare the subjective visual performance of our methods with other low-light video enhancement methods, we conduct a user study involving 20 participants. We capture dynamic low-light videos using an iPhone 14 Pro in real-world scenarios, segment the videos into frames, and feed them into six different low-light video enhancement methods (MBLLEN, SDSDNet, Chhirolya et al.'s method, SNR-Aware, LAN, and LAN++) trained on three different low-light video datasets (SDSD, DID, R-DID). Subsequently, we group and compare the results to evaluate the performance of various low-light video enhancement models and the quality of different low-light video datasets.

We have each participant perform 5 comparisons, including three sets of model performance comparisons and two sets of dataset generalization comparisons, with each comparison involving the random selection of 3 videos for evaluation. For the model performance comparisons, the DID dataset is used as the training data. Firstly, one method from the four low-light video enhancement methods other than LAN and LAN++ is selected for comparison against LAN and LAN++. In each comparison, participants simultaneously view two videos (referred to as Video A and Video B), and compare them based on factors like realism, brightness, contrast, etc., selecting between "Video A is better," "Video B is better," or "I don't know which is better." Subsequently, one method is chosen from the six methods for the dataset generalization comparison, where participants compare the results of this method when trained on the SDSD, DID, and R-DID datasets, ranking them based on generalization performance into the best, second-best, and worst generalization datasets.

The quantitative results of the user study are shown in Table VII and Table VIII, respectively. Besides, the visual results of the user study can be seen in Fig. 17. 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, DID and R-DID datasets. The results indicate 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 SDSD, DID and R-DID datasets, methods trained with DID datasets can produce more natural and vivid videos, while models trained with R-DID datasets can produce videos with better continuity.

At the same time, we also compared the performance of LAN++ with other methods under different degrees of low-light conditions. As shown in Fig. 19, under extremely low-light conditions, our method adjusts the number of iterations adaptively to produce brighter images. In slightly dim conditions, our method reduces the degree of brightening by adaptively adjusting the number of iterations, avoiding overexposure issues observed in other methods.

V. Conclusion

To help effectively enhance the low-light videos, we create two datasets: a dynamic high-quality paired low-light video dataset called DID, and a real continuous paired low-light video dataset called R-DID. DID is generated with pronounced camera motion and strict spatial alignment, and R-DID is aligned by optical flow algorithm in moving object scenes. 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. Further we incorporate a novel inter-frame relationship, difference image, into the structure of LAN and design an improved Light Adjustable Network called LAN++. Extensive experiments and user studies demonstrate the effectiveness of our proposed datasets and methods, which outperform state-of-the-art approaches. Our work contributes novel resources and methodologies for low-light video enhancement.

References

[1] E. H. Land, "The retinex theory of color vision," *Scientific Amer.*, vol. 237, no. 6, pp. 108–129, Dec. 1977.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[3] Q. Han et al., "Demystifying local vision transformer: Sparse connectivity, weight sharing, and dynamic weight," 2021, *arXiv:2106.04263*.

[4] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11976–11986.

[5] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *Proc. ECCV*, Oct. 2022, pp. 17–33.

[6] W. Yu et al., "MetaFormer is actually what you need for vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 10809–10819.

[7] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[9] C. Li et al., "Low-light image and video enhancement using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9396–9416, Dec. 2022.

[10] W. Wang, X. Chen, C. Yang, X. Li, X. Hu, and T. Yue, "Enhancing low light videos by exploring high sensitivity camera noise," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4111–4119.

[11] C. Chen, Q. Chen, M. Do, and V. Koltun, "Seeing motion in the dark," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3184–3193.

[12] H. Jiang and Y. Zheng, "Learning to see moving objects in the dark," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7324–7333.

[13] R. Wang, X. Xu, C. Fu, J. Lu, B. Yu, and J. Jia, "Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jun. 2021, pp. 9700–9709.

[14] D. Triantafyllidou, S. Moran, S. McDonagh, S. Parisot, and G. Slabaugh, "Low light video enhancement using synthetic data produced with an intermediate domain mapping," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 103–119.

[15] S. Aghajanzadeh and D. Forsyth, "Long scale error control in low light image and video enhancement using equivariance," 2022, *arXiv:2206.01334*.

[16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 10684–10695.

[17] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12873–12883.

[18] J. Yu et al., "Vector-quantized image modeling with improved VQGAN," 2021, *arXiv:2110.04627*.

[19] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[20] S. Chhirolya, S. Malik, and R. Soundararajan, "Low light video enhancement by learning on static videos with cross-frame attention," 2022, *arXiv:2210.04290*.

[21] F. Lv, F. Lu, J. Wu, and C. Lim, "MBLLEN: Low-light image/video enhancement using CNNs," in *Proc. BMVC*, 2018, pp. 1–4.

[22] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[23] Y. Wang, R. Wan, W. Yang, H. Li, L.-P. Chau, and A. C. Kot, "Low-light image enhancement with normalizing flow," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2604–2612.

[24] X. Xu, R. Wang, C. Fu, and J. Jia, "SNR-aware low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 17714–17724.

[25] L. Ma, T. Ma, R. Liu, X. Fan, and Z. Luo, "Toward fast, flexible, and robust low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5637–5646.

[26] W. Yang, S. Wang, Y. Fang, Y. Wang, and J. Liu, "From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3063–3072.

[27] R. Liu, L. Ma, J. Zhang, X. Fan, and Z. Luo, "Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 10556–10565.

[28] S. Wang, J. Zheng, H.-M. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3538–3548, Sep. 2013.

[29] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proc. Image Anal., 13th Scand. Conf.* Cham, Switzerland: Springer, 2003, pp. 363–370.

[30] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "DeepFlow: Large displacement optical flow with deep matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2013, pp. 1385–1392.

[31] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3559–3568.

[32] H. Fu, W. Zheng, X. Wang, J. Wang, H. Zhang, and H. Ma, "Dancing in the dark: A benchmark towards general low-light video enhancement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jun. 2023, pp. 12831–12840.

[33] X. Xu, R. Wang, and J. Lu, "Low-light image enhancement via structure modeling and guidance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 9893–9903.

[34] S. Zheng and G. Gupta, "Semantic-guided zero-shot learning for low-light image/video enhancement," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Waikoloa, HI, USA, Jan. 2022, pp. 581–590.

[35] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[36] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "MUSIQ: Multi-scale image quality transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5128–5137.

[37] H. Fu, W. Zheng, X. Meng, X. Wang, C. Wang, and H. Ma, "You do not need additional priors or regularizers in retinex-based low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18125–18134.

[38] L. Liu et al., "Low-light video enhancement with synthetic event guidance," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 2, pp. 1692–1700.

[39] R. Wang, Q. Zhang, C. Fu, X. Shen, W. Zheng, and J. Jia, "Underexposed photo enhancement using deep illumination estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 6842–6850.

[40] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," 2018, *arXiv:1808.04560*.

[41] X. Fu, D. Zeng, Y. Huang, X.-P. Zhang, and X. Ding, "A weighted variational model for simultaneous reflectance and illumination estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2782–2790.

[42] X. Guo, Y. Li, and H. Ling, "LIME: Low-light image enhancement via illumination map estimation," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 982–993, Feb. 2017.