Learning Temporal Consistency for Low Light Video Enhancement from Single Images

Fan Zhang[1] Yu Li[2] Shaodi You[3] Ying Fu[1*]
[1]Beijing Institute of Technology [2]Applied Research Center (ARC), Tencent PCG
[3]University of Amsterdam

Abstract

Single image low light enhancement is an important task and it has many practical applications. Most existing methods adopt a single image approach. Although their performance is satisfying on a static single image, we found, however, they suffer serious temporal instability when handling low light videos. We notice the problem is because existing data-driven methods are trained from single image pairs where no temporal information is available. Unfortunately, training from real temporally consistent data is also problematic because it is impossible to collect pixel-wisely paired low and normal light videos under controlled environments in large scale and diversities with noise of identical statistics. In this paper, we propose a novel method to enforce the temporal stability in low light video enhancement with only static images. The key idea is to learn and infer motion field (optical flow) from a single image and synthesize short range video sequences. Our strategy is general and can extend to large scale datasets directly. Based on this idea, we propose our method which can infer motion prior for single image low light video enhancement and enforce temporal consistency. Rigorous experiments and user study demonstrate the state-of-the-art performance of our proposed method. Our code and model will be publicly available at https://github.com/zkawfanx/StableLLVE.

1. Introduction

Illumination in sunny day and low light night can vary more than 10 orders of magnitude. In low light scenes, sensor noise is not negligible due to the low signal-to-noise ratio (SNR). Therefore, low light image enhancement is an important task which improves the SNR and enhances the image after modeling the noise and the signal. It enables various computer vision algorithms to perform properly.

Rather than explicitly modeling the noise and the signal, recent data-driven methods [3, 4, 9, 25, 26, 27] implicitly learn such models from image data and get satisfying results on a single static image. And they require pixel-wisely paired images of low and high SNR for training. However, we notice that it is impossible to collect pixel-wisely paired low and normal light videos under controlled environments in large scale and diversities with noise of identical statistics. Therefore, existing single image methods use either synthetic data or temporally inconsistent single image data for training. Thus, no temporal consistency can be learned through existing data. One can perceive serious artifacts and flickering from existing single image methods when handling low light videos.

In this paper, we aim to enforce temporal consistency even when training from static images. We propose a novel method to enforce the temporal stability in low light video enhancement with only static images. The key idea is to learn and infer motion field (optical flow) from a single image and synthesize short range video sequences. Our strategy is general and can extend to large scale datasets directly.

Based on this idea, we propose our method which can infer motion prior for single image low light video enhancement and enforce temporal consistency. In particular, we present an image-based method to achieve low light video enhancement and tackle temporal inconsistency problem by imposing consistency on the network. Specifically, we choose optical flow to mimic motions of dynamic scenes. It is more capable of representing both global and local motions. We first predict plausible optical flow from static images. Then we warp images with optical flow to be adjacent frames and impose consistency on deep model.

We conduct rigorous experiments to validate the effectiveness of our method. Experimental results on both synthetic and real data show that our method outperforms the state-of-the-art single image methods and achieves comparable results to video-based ones, which means our method can alleviate flickering problem without the need of videos. Furthermore, we also conduct a user study on 26 volunteers, of whom 78.9% prefer our method, suggesting the better temporal stability of our method.

*Corresponding author: fuying@bit.edu.cn

Our main contributions are summarized as follows:

- We present a novel solution to solving temporal inconsistency problem of low light video enhancement when using only single image data.
- We propose to use optical flow prior to indicate potential motion from single image and thus enable us to model the temporal consistency.
- We demonstrate the state-of-the-art performance of our method from rigorous experiments and user study.

2. Related works

Low light video enhancement is closely related to low light image enhancement. In this section, we briefly review some typical methods of this two tasks.

Low Light Image Enhancement Traditional low light image enhancement methods can be divided into two categories, *i.e*., histogram equalization based methods and Retinex theory based methods. Histogram equalization [22] is a simple yet effective method to stretch the histogram of images and improve the contrast. Many methods [1, 2, 7, 15, 20] extend it using more complex priors. Arici *et al*. [1] propose WAHE to adjust the level of contrast enhancement while alleviating unnatural artifacts by introducing specially designed penalties. Celik and Tjahjadi [2] propose CVC to enhance the contrast of an input image using interpixel contextual information. Lee *et al*. [15] propose LDR to enhance image contrast by amplifying the gray-level differences between adjacent pixels based on the layered difference representation of 2D histograms. On the other hand, Retinex theory [14] assumes that an image is composed of reflection and illumination. Jobson *et al*. [11] propose the best placement of the logarithmic function and Gaussian form to define a specific retinex called SSR to handle gray-world violations. They [10] also extend it to multiscale version and define a method of color restoration. Lee *et al*. [16] adaptively compute the weights of each SSR output according to the content of input. Wang *et al*. [28] propose NPE to enhance image details while preserving naturalness. Guo *et al*. [6] propose LIME to refine the initial illumination map of each pixel by imposing a structure prior and get final enhancement from it.

Deep learning based methods are recently introduced into low light image enhancement task. Lore *et al*. [17] propose a multi-autoencoder framework called LLNet for enhancing low light images and denoising. Wei *et al*. [30] propose RetinexNet based on the Retinex theory [14] to decompose images into reflectance and illumination and enhance the illumination to get normal light images. Lv *et al*. [19] design a multibranch network called MBLLEN to handle low light image enhancement and denoising simultaneously. They also [18] extend it by adding attention module and provide a large scale synthetic dataset. Wang *et al*. [27] propose a network called DeepUPE to model image-to-image illumination and collect an expert-retouched dataset. Zhang *et al*. [33] propose a network called KinD based on retinex theory and design a restoration module to handle noise. Chen *et al*. [4] collect a dataset named SID and train a U-Net [24] to estimate enhanced sRGB images from raw low light images. These models do not fully consider temporal consistency and may face flickering problem if applied to videos directly.

Low Light Video Enhancement Unlike low light image enhancement, low light video enhancement is still open and challenging. Common solution is to extend low light image enhancement models to their 3D version. Lv *et al*. [19] substitute the 2D convolution layers of MBLLEN [19] with 3D ones to handle image sequences and train the model on synthetic low light video data. Jiang *et al*. [9] propose a novel setup to collect dark and bright video pairs and train a modified 3D U-Net on them, which gets promising results thanks to this dataset. However, this specialized equipment is unavailable to the public yet, which consequently limits the diversity and scale of collected video dataset.

Other attempts have also been made to utilize image-based methods to enhance low light videos and alleviate flickering problem. Self-consistency is often utilized to improve the performance and stability of deep models by imposing similarity of data pairs. Chen *et al*. [3] collect a video dataset containing low light image sequences and their long exposure ground truths of static scenes and train their model with randomly sampled frame pairs from the same sequence. With the help of self-consistency loss, the model learns to tolerate minor differences of inputs caused by noise and keeps its outputs stable. Eilertsen *et al*. [5] propose more general strategies to learn temporal stability in which they apply random disturbances like noise or global affine transformation including rotation, translation to images and feed them into networks. By enforcing consistency between warped outputs, they help model keep stable when processing video frames. However, simple transformations such as rotation and translation are not enough to represent motions between video frames, since they can not describe complex motions such as irregular motion of non-linear objects and ego-motion of cameras. In contrast, we consider optical flow as descriptor of motions in dynamic scenes which is complement to representing motions and well exploited in the past decades.

Lai *et al*. [13] also propose a deep network with ConvLSTM module to learn temporal consistency from video sequences explicitly utilizing optical flow estimated by FlowNet2 [8] at the training stage, which serves as a general

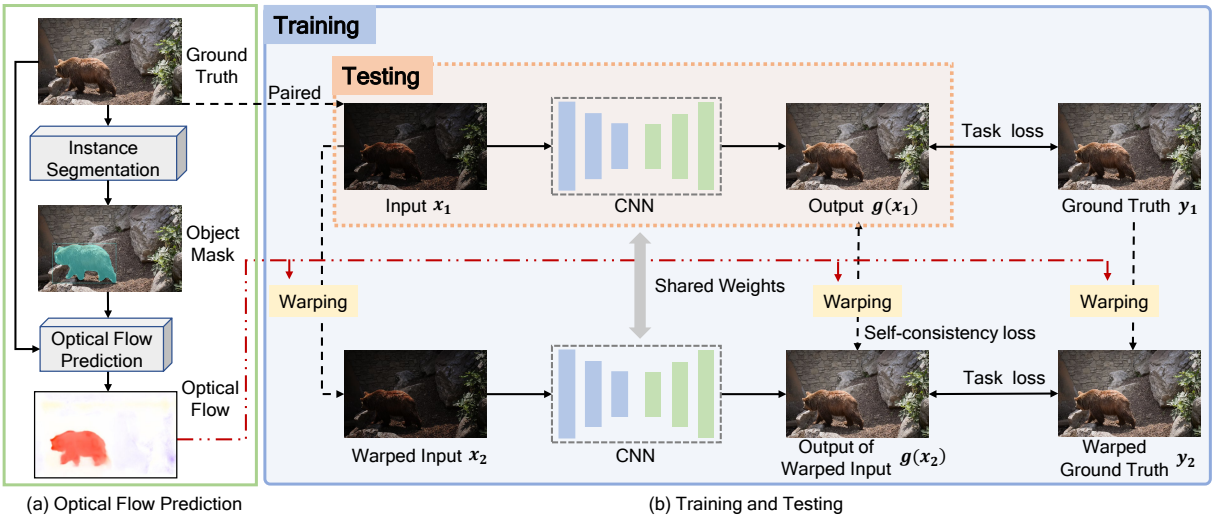
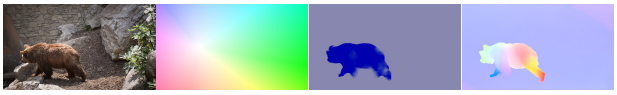


Figure 1. Overview of our full pipeline that consists of two steps. (a) Optical Flow Prediction. We first utilize instance segmentation [31] to detach objects from background and randomly sample 10 guidance motion vectors on each object region. With bright image and vectors fed into optical flow prediction network [32], we can vary the directions and magnitudes to get diverse optical flow. This prediction step can be computed offline before training. (b) Training and Testing. Our method consists of two branches of which the upper one works in both training and testing phase while the other one only works during training as an auxiliary branch to impose temporal consistency on the network. Images in the second branch are warped from images in the main branch with the same optical flow. During inference stage, our network directly take the input and predict the output.



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

Figure 2. Examples of optical flow results. (a) Normal light video frame. (b) Optical flow from global affine transformation. (c) Optical flow prediction by our instance aware optical flow simulation. (d) Optical flow estimated from adjacent normal light video frames. The predicted optical flow is instance-aware and more similar to the real optical flow between adjacent video frames than that of global transformation.

post-processing method to alleviating flickering regardless of specific task. Different from their work, we train image-based models with image data and embed temporal consistency implicitly into them with optical flow generated from single images.

3. Method

We propose a novel method to achieve low light video enhancement via image-based methods and tackle flickering problem by stabilizing the network. More concretely, we utilize optical flow to represent motion occurred between video frames of dynamic scenes. We mimic adjacent frames of images by warping them with corresponding optical flow. Given image pairs of original image and warped image, we train our network in a siamese way in which we feed them one by one to the network. By imposing consistency between output pairs we can help network be temporally stable. We focus on both noise-free and noisy cases and we test our model on real noisy data to show the robustness and flexibility of our network.

In this section, we first introduce the whole work flow and architecture and then provide implementation details.

3.1. Architecture

An ideal temporally stable model should be transform-invariant. In other word, the model should get outputs of transformed inputs with the same transformations as if the operations are applied to outputs directly. Only a model behaving in this way can process videos frame by frame without flickering problem. Holding this assumption, we try to generate motion with optical flow to mimic actual video sequences. By feeding such image pairs into network and enforcing consistency between outputs before and after warping, we can help the network learn temporal stability.

As shown in Figure 1, we first predict plausible optical flow from well illuminated ground truth before training the network. We utilize an pretrained instance segmentation model from open-source toolkit Detectron2 [31] to detach object regions from backgrounds, where local motions usually occur. Given estimated object masks, it is easy to get optical flow predictions with an unsupervised model called CMP [32]. Figure 2 shows a comparison of our predicted

Figure 3. Visual results of clean case. Ours is much cleaner than the baseline in terms of artifacts and comparable to video-based methods.

optical flow and that of global transformation with real optical flow estimated from adjacent ground truth video frames, where our prediction is instance-aware and more similar to the real case. With necessary optical flow ready, we train our image-based model in a siamese way. The upper part of the pipeline is the first pass of network and is the same as common training procedure. A low light image x_1 from training dataset is fed into the network $g(\cdot)$ and it predicts an enhanced result $g(x_1)$. The network learns to recover normal light images with the help of supervision from corresponding well illuminated ground truth y_1. To provide more temporal information, we warp the input image x_1 with random optical flow f which is predicted based on ground truth. The warped image x_2 serves as input for the second pass. The output $g(x_2)$ is also compared to corresponding warped ground truth y_2 for supervision. Finally, the output $g(x_1)$ is warped with the same optical flow f to $W\big(g(x_1), f\big)$ and compared with output $g(x_2)$.

Previous works [3, 4, 9] have collected their low light datasets and simply train a U-Net [24] on their data. Here, we also choose this simple yet effective model to validate the effectiveness of our method and follow the implementation in SID [4]. We adopt l_1 loss for all losses and the loss used to train the network can be defined as a combination of enhancement loss $\mathcal{L}_e$ and consistency loss $\mathcal{L}_c$:

$$\mathcal{L} = \mathcal{L}_e + \lambda \mathcal{L}_c, \tag{1}$$

where λ is the weight which balances the constraints of two loss parts. Specifically, $\mathcal{L}_e$ and $\mathcal{L}_c$ are formulated as:

$$\mathcal{L}_e = \sum_{i=1,2} \|g(x_i) - y_i\|_1, \tag{2}$$

and

$$\mathcal{L}_c = \|W\big(g(x_1), f\big) - g(x_2)\|_1, \tag{3}$$

where $g(\cdot)$ represents the network forwarding operation. x_i and y_i denote the input and the ground truth in the ith pass. f is the optical flow we generate for motion simulation.

For generality, we take both noise-free and noisy cases into consideration. For noise-free case, we train deep model directly on low light and normal light image pairs following the procedure described above. For noisy case, we first sample noise from Gaussian and Poisson distributions and add it to low light images before being fed into network.

Table 1. Quantitative comparison on clean cases. The three groups from top to bottom are image-based methods, video-based methods, and single image methods utilizing self-consistency including SFR and ours.

Method	PSNR↑	SSIM↑	AB(Var)↓	MABD↓	WE (×1e-3)↓
LIME [6]	17.36	0.7386	9.65	0.37	3.420
MBLLEN [19]	18.41	0.8100	77.24	1.95	1.700
RetinexNet [30]	19.78	0.8353	1.32	0.09	1.372
SID [4]	22.95	0.9428	4.93	0.43	1.182
MBLLVEN [19]	24.50	0.9482	1.79	0.80	0.999
SMOID [9]	**24.85**	0.9472	**1.30**	0.17	1.077
SFR [5]	23.81	0.9413	2.14	0.11	1.097
BLIND [13]	22.87	0.9344	8.66	0.43	**0.977**
Ours	24.07	**0.9483**	1.96	**0.05**	1.061

3.2. Implementation Details

Our training is implemented on Pytorch [21]. We apply random cropping, horizontal flipping and rotation for data augmentation. Cropping size is 512×512 and rotation angles include 90, 180 and 270 degrees. The learning rate is set to 1×10^{-4}, and the model is trained by Adam optimizer [12] with default parameters for 50 epochs on single GTX 1080ti. For the stability of training, we stop the gradients of $\mathcal{L}_c$ propogating to the warped output $W\big(g(x_1), f\big)$.

4. Data Preparation

In this section, we detail the data preparation procedure for training our model.

4.1. Optical Flow Prediction

Instead of global affine transformation, we choose optical flow to represent motion for its ability to represent both global and local motions. Thus, we need to acquire predicted optical flow of images first. Unlike most optical flow methods which concentrate on estimating optical flow

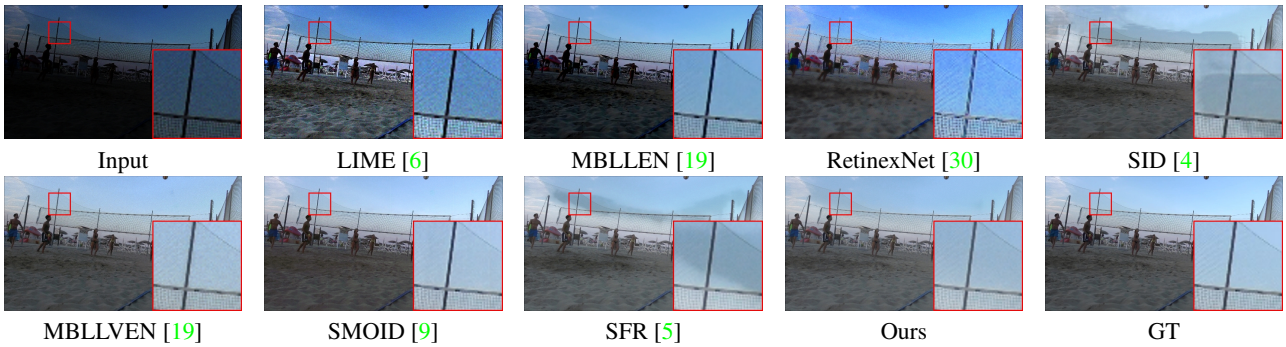


Figure 4. Visual results of noisy case. Results are similar to clean case where ours performs better than image-based methods and comparably to video-based methods

of two different images, we try to predict random optical flow from a single image. Specifically, Conditional Motion Propagation Network (CMP) [32] is adopted. We utilize its pretrained model to predict plausible optical flow of ground truths. In the generation step, it needs to be initialized with some guidance motion vectors on objects and we use instance segmentation [31] to help obtain these vectors. We first segment images and get rough object masks representing objects regions. Then, we sample several motion vectors on each region and predict optical flow based on them:

$$f = \mathcal{CMP}(y, V), \tag{4}$$

where $\mathcal{CMP}$ represents the optical flow prediction model, y and V denote ground truth images and guidance motion vectors, respectively.

We randomly sample 10 vectors for each object in images to get final predictions. Notice that the randomly sampled guidance vectors can not ensure the quality of predicted optical flow but we believe that failures in optical flow predictions can also be helpful for training by introducing disturbances. The predicted optical flow serves as initial start point that generates various optical flow cases via augmentation. With the predicted optical flow results, we can get the warped image by

$$x_2 = W(x_1, f), \tag{5}$$

where f represents the predicted optical flow, x_1 and x_2 are the original and warped images respectively.

Visualizations of optical flow we predict are included in the supplementary material.

4.2. Low Light Image Synthesis

To investigate the effectiveness of our method, we need to compare our models with both image-based and video-based methods but low light video datasets are rare. In this paper, we choose DAVIS dataset [23] as our ground truth data. It is a large scale dataset for video segmentation tasks. We exclude badly illuminated videos and synthesize low light videos. Following [18], we darken these bright images using gamma correction and linear scaling:

$$x = \beta \times (\alpha \times y)^{\gamma}, \tag{6}$$

where γ is gamma correction which is sampled in a uniform distribution $U(2, 3.5)$. α and β denote linear scaling factors and are sampled from $U(0.9, 1)$ and $U(0.5, 1)$, respectively.

DAVIS [23] contains two resolutions, full resolution and 480P. We use all full resolution videos, including training set, test and validation sets for 2017 challenge and 2019 challenge. After excluding badly illuminated videos, we keep all videos with 1920×1080 resolution and get 107 videos containing 7179 frames in total. We randomly split these videos into training set and test set, 87 videos in training set and 20 videos in test set specifically. The same image augmentation is applied to corresponding optical flow. Standalone augmentation is performed to them after that to get various plausible optical flow. They are randomly rotated by 2 degrees, randomly flipped, and random global offset is added in horizontal or vertical direction or both.

Table 2. Quantitative results for noisy case. Our method is more stable than image-based methods and comparably stable to video-based methods.

Method	PSNR↑	SSIM↑	AB(Var)↓	MABD↓	WE(×1e-3)↓
LIME [6]	16.83	0.4567	8.29	0.33	5.545
MBLLEN [19]	18.38	0.7982	78.76	1.93	1.719
RetinexNet [30]	19.56	0.7475	1.45	**0.09**	1.769
SID [4]	22.93	0.9253	4.03	0.39	1.303
MBLLVEN [19]	23.08	0.8839	2.81	1.02	1.221
SMOID [9]	23.42	0.9212	**0.82**	0.17	1.184
SFR [5]	22.82	0.9299	2.29	0.12	1.200
BLIND [13]	22.94	0.9174	7.86	0.33	1.031
Ours	**24.01**	**0.9305**	3.00	0.10	**1.024**

4.3. Noise

Noise is another matter we want to take care of. Aside from optical flow prediction and low light image generation, we use Gaussian and Poisson noise for noise simulation. We

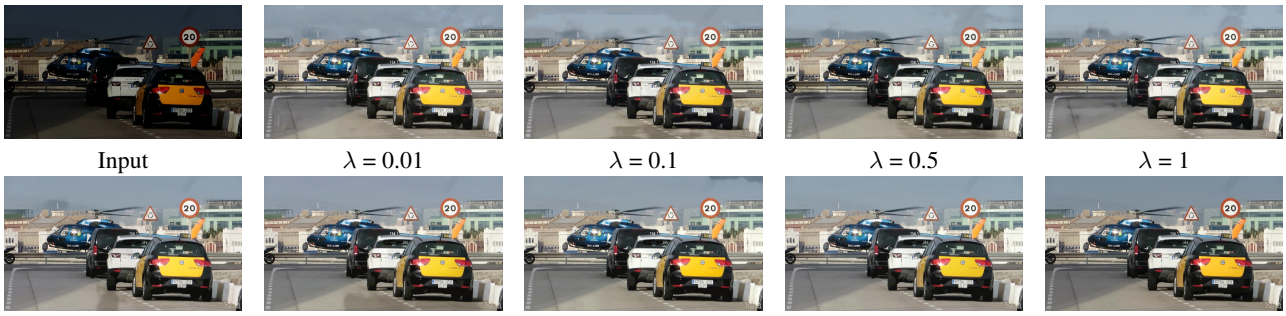


Figure 5. Visual results of different weights. With the increase of weights, model outputs becomes smoother.

believe that our pipeline is robust and can be applied to real noisy images. The noise sampling can be described as

$$n = \mathcal{P}(x, \sigma_p) + \mathcal{N}(\sigma_g), \tag{7}$$

where σ_p, σ_g denote parameters of Poisson noise and Gaussian noise, respectively. They are both sampled from $U(0.01, 0.04)$.

5. Experiments

We conduct quantitative and qualitative experiments to verify the effectiveness of our method. First we compare our method with other methods under noise-free and noisy setting respectively. Then we make comparisons using real low light videos. After that, we conduct ablation study to figure out optimal weight of our self-consistency branch and to show how it behaves with existence of different noise components and under different low light levels. Finally, we conduct a user study and inference speed test for further comparison.

Due to the limited space, we only provide several typical visual results here. More qualitative results, results of another real scene, results of ablation study on different light levels and results of inference speed test can be found in our supplementary material.

5.1. Experiment Setup

We compare our method with three kinds of enhancement methods, including image-based methods, video-based methods and methods utilizing self-consistency. Seven methods are selected from these categories. In the first group, LIME [6] is a traditional method while MBLLEN [19], RetinexNet [30] and SID [4] are deep learning methods. Two video-based methods MBLLVEN [19] and SMOID [9] are also learning based methods. The last method proposed by Eilertsen *et al*. [5] imposes consistency between global transformed image pairs and we denote it as SFR here for short. In addition, we include the post-processing method from Lai *et al*. [13] which is denoted as BLIND in quantitative evaluation to further complement our experiments in spite that post-processing is not the focus of our discussion.

Table 3. Ablation Study of branch weights. With the increase of weights, model becomes more stable temporally and PSNR and SSIM increase.

Weight	PSNR↑	SSIM↑	AB(Var)↓	MABD↓	WE(×1e-3)↓
$\lambda = 0.01$	22.26	0.9381	5.25	0.55	1.356
$\lambda = 0.1$	22.40	0.9442	4.13	0.40	1.348
$\lambda = 0.5$	22.54	0.9433	4.66	0.47	1.298
$\lambda = 1$	22.44	0.9476	4.84	0.53	1.415
$\lambda = 5$	22.82	0.9478	3.83	0.41	1.250
$\lambda = 10$	23.37	0.9548	1.89	0.25	1.231
$\lambda = 15$	**24.05**	**0.9571**	1.73	0.08	1.171
$\lambda = 20$	24.04	0.9545	**1.30**	**0.04**	**1.147**

We evaluate their performances with two common metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural SIMilarity (SSIM) [29]. Furthermore, we choose AB(Var) from [19], Mean Absolute Brightness Difference (MABD) from [9] and Warping Error (WE) from [13] to validate temporal stability of models. Warping Error takes use of optical flow and reflects differences among adjacent frames mainly in dynamic areas. The lower values in the three metrics stand for better temporal stability.

5.2. Synthetic Data

Here we conduct two experiments for both noise-free and noisy cases. Quantitative results are provided in Tables 1 and 2.

As can be seen in Table 1, image-based methods LIME [6], MBLLEN [19] and RetinexNet [30] get comparable results on PSNR and SSIM under noise-free setting. MBLLEN and RetinexNet are better than LIME in terms of PSNR and SSIM and more stable according to Warping Error. While MBLLEN gets large values in AB(Var) due to its under-exposed and over-exposed enhancements. But they are all worse than our baseline SID. Video-based methods MBLLVEN [19] and SMOID [9] both have better PSNR and SSIM and smaller Warping Error. For SFR [5] and our method, we can see that both methods achieve comparable results as video-based methods while ours are better

Table 4. Ablation Study of different noise distributions. Ours gets comparable results in cases of different noise distributions, which shows robustness of our method.

Noise	PSNR↑	SSIM↑	AB(Var)↓	MABD↓	WE(×1e-3)↓
G	23.29	0.8681	2.38	0.1208	1.208
P	**24.04**	**0.9374**	1.96	**0.1176**	**1.176**
G+P	23.27	0.8648	**1.95**	0.1243	1.243

than the other. As for the post-processing method BLIND [13], it only improves the results of the baseline on Warping Error and has no help to PSNR and SSIM.

Several typical enhancements are shown for visual comparison in Figure 3. We can see that LIME suffers from over saturation and RetinexNet gets unreal results. MBLLEN performs poorly in recovering brightness. SID [4] suffers from checkerboard artifacts due to deconvolution. SFR [5] and ours are more stable temporally.

Experimental results of all compared methods under Gaussian and Poisson noise are provided in Table 4. We can see that all methods decrease slightly in their PSNR and SSIM while their temporal stability keep tight with their clean cases. Besides, our method and SFR [5] can achieve comparable performance and temporal stability as video-based methods. Also our method surpasses all compared image-based methods and the post-processing method BLIND [13].

As shown in Figure 4, we can see that LIME, MBLLEN and RetinexNet all fail to recover correct low light video. Brightness of MBLLEN is much lower than ground truth which results in the large value of AB(Var). RetinexNet enhances images with unreal color and too much smoothness which results in better temporal stability. SID actually gets heavier artifacts due to the existence of noise. SFR and ours perform better while ours is still better than the other. Video-based methods all get pleasant visual quality compared to aforementioned ones.

Both quantitative metric results and visual quality show that our method can improve temporal stability of deep model and alleviate flickering problem without the need of video training data.

5.3. Real Data

To further verify the robustness of the proposed method, we collect real low light videos. All tested methods except LIME are trained on synthetic noisy data. As shown in Figure 7, traditional method LIME actually performs well for it does not get influenced by data distribution but suffers from over-exposure and over-saturation. Learning based methods all show somehow differences from real data but we can still discriminate out their temporal stability on real videos. Among these results, MBLLEN enhances its results similarly to LIME and faces over-saturation too. RetinexNet gets unreal color and over-exposed. SID suffers from artifacts. In contrast, SFR and our method evidently alleviate this problem. But we can still find out some artifacts in outputs from SFR.

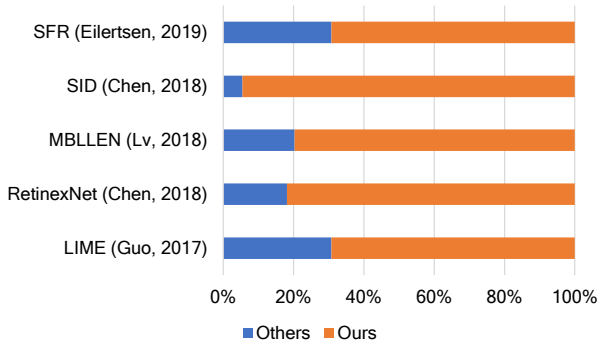


Figure 6. Preference distribution of the user study. Our method is compared with seven methods in seven blind A/B tests. Participants vote for methods that are more stable and visually pleasant.

5.4. Ablation Study

Training a temporally stable image-based model is actually a compromise between visual quality and temporal stability. The optimal result lies in the balance of them. To show the influence of different weight on consistency branch and generality of our model, we conduct two ablation studies of weight parameter and noise distributions.

We conduct ablation study to investigate optimal weight for our method. With different parameter settings, our method behaves accordingly. As we can see in Table 3, with the increase of branch weight, the network becomes more temporally stable compared to that with smaller weight and improves its PSNR and SSIM. When the weight arrives at a certain point, the benefit of improving enhancement quality disappears and the network starts to drop in PSNR and SSIM for more improvements on temporal stability. And we can find out the best parameter setting is around $\lambda = 20$. Visual results are provided in Figure 5.

We compare our method on different noise distributions including Gaussian noise only, Poisson noise only and mixed noise. Quantitative results are provided in Table 4. We can find that with various noise components, our pipeline all work properly.

5.5. User Study

We conduct a user study on video stability with 26 participants. The experiment consists of 5 groups of blind A/B tests between our method and other image-based methods. 7 test videos are randomly selected for each group. Only two enhanced videos are provided to users at a time. Figure 6 shows that our method surpasses all image-based methods by a large margin.

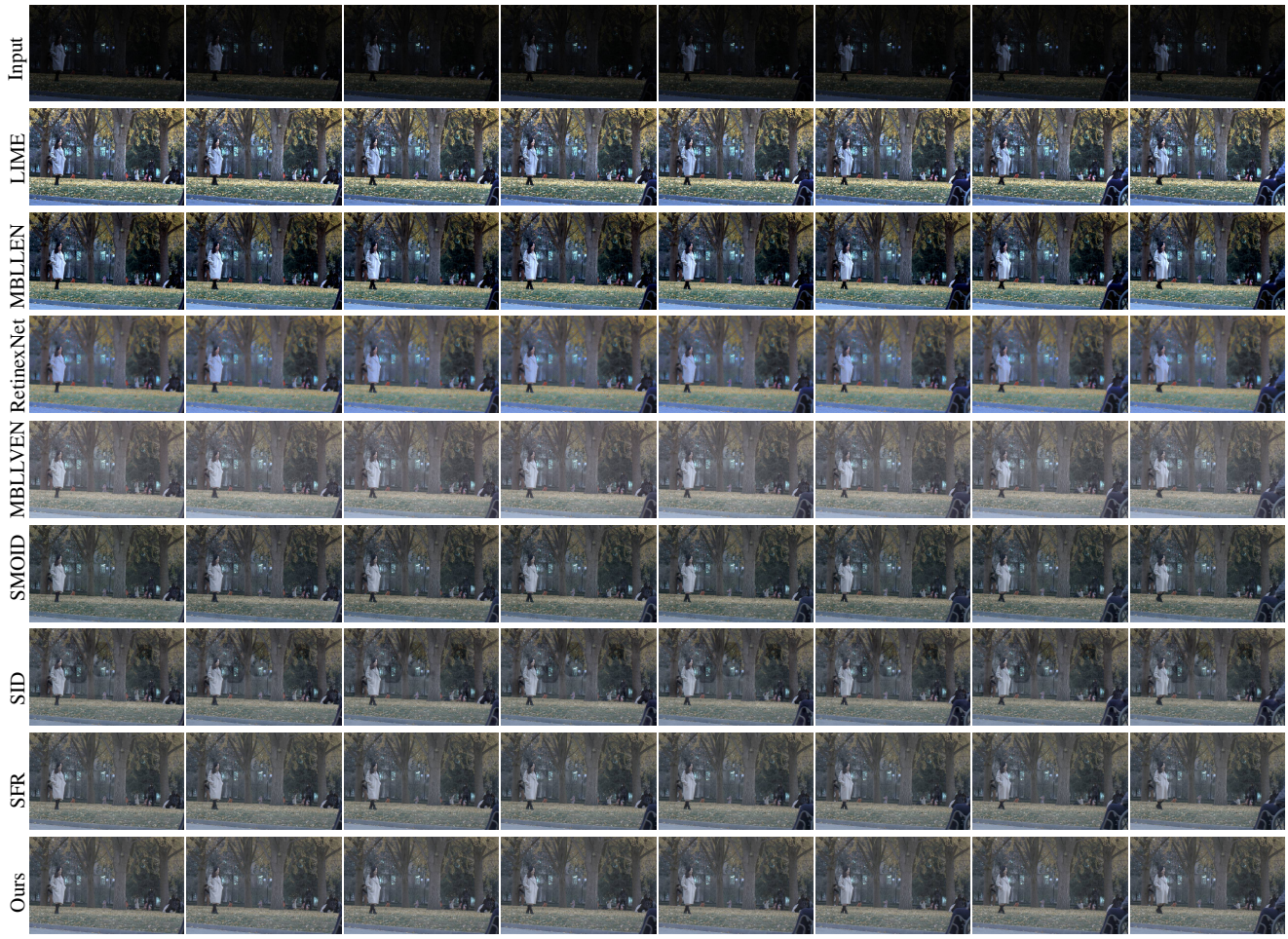


Figure 7. Several frames for real data test. From top to bottom: 1. Input; 2. LIME [6]; 3. MBLLEN [19]; 4. RetinexNet [30]; 5. MBLLVEN [19]; 6. SMOID [9]; 7. SID [4]; 8. SFR [5]; 9. Our results.

6. Conclusion

In this paper, we propose a novel method for low light video enhancement with image-based model and alleviate flickering by temporally stabilizing it. With the help of generated optical flow, we guide the model to learn temporal stability by enforcing consistency on warped outputs. Quantitative and qualitative results show the good balance of enhancement quality and temporal stability of the trained model. Our method can effectively work for the video recovery by single frames. In the future, we are planning to investigate how to extend our model for deraining, dehazing, intrinsic decomposition and other tasks where temporal consistency are important.

Acknowledgments This work was supported by the National Natural Science Foundation of China under Grants No. 61827901, No. 62088101, and No. 61936011.

References

[1] Tarik Arici, Salih Dikbas, and Yucel Altunbasak. A histogram modification framework and its application for image contrast enhancement. *IEEE Transactions on Image Processing*, 18(9):1921–1935, 2009. 2

[2] Turgay Celik and Tardi Tjahjadi. Contextual and variational contrast enhancement. *IEEE Transactions on Image Processing*, 20(12):3431–3441, 2011. 2

[3] Chen Chen, Qifeng Chen, Minh N Do, and Vladlen Koltun. Seeing motion in the dark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 4

[4] Chen Chen, Qifeng Chen Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018. 1, 2, 4, 5, 6, 7, 8

[5] Gabriel Eilertsen, Rafal K Mantiuk, and Jonas Unger. Single-frame regularization for temporally stable cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2019. 2, 4, 5, 6, 7, 8

[6] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26(2):982–993, 2017. 2, 4, 5, 6, 8

[7] Haidi Ibrahim and Nicholas Sia Pik Kong. Brightness preserving dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics*, 53(4):1752–1758, 2007. 2

[8] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017. 2

[9] Haiyang Jiang and Yinqiang Zheng. Learning to see moving objects in the dark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 4, 5, 6, 8

[10] Daniel J Jobson, Zia-ur Rahman, and Glenn A Woodell. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image Processing*, 6(7):965–976, 1997. 2

[11] Daniel J Jobson, Zia-ur Rahman, and Glenn A Woodell. Properties and performance of a center/surround retinex. *IEEE Transactions on Image Processing*, 6(3):451–462, 1997. 2

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[13] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2, 4, 5, 6, 7

[14] Edwin H Land. The retinex theory of color vision. *Scientific American*, 237(6):108–129, 1977. 2

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

[16] Chang-Hsing Lee, Jau-Ling Shih, Cheng-Chang Lien, and Chin-Chuan Han. Adaptive multiscale retinex for image contrast enhancement. In *Proceedings of the International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, 2013. 2

[17] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61:650–662, 2017. 2

[18] Feifan Lv, Yu Li, and Feng Lu. Attention guided low-light image enhancement with a large scale low-light simulation dataset. *arXiv: 1908.00682*, 2019. 2, 5

[19] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. Mbllen: Low-light image/video enhancement using cnns. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 2, 4, 5, 6, 8

[20] Keita Nakai, Yoshikatsu Hoshi, and Akira Taguchi. Color image contrast enhacement method based on differential intensity/saturation gray-levels histograms. In *Proceedings of the International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, 2013. 2

[21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the Advances in neural information processing systems (NIPS)*, 2019. 4

[22] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 39(3):355–368, 1987. 2

[23] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 5

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical image computing and computer-assisted intervention (MICCAI)*. Springer, 2015. 2, 4

[25] Li Tao, Chuang Zhu, Jiawen Song, Tao Lu, Huizhu Jia, and Xiaodong Xie. Low-light image enhancement using cnn and bright channel prior. In *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017. 1

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

[27] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2019. 1, 2

[28] 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. 2

[29] 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. 6

[30] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 2, 4, 5, 6, 8

[31] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 3, 5

[32] Xiaohang Zhan, Xingang Pan, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised learning via conditional motion propagation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2019. 3, 5

[33] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *Pro-*

ceedings of the ACM International Conference on Multimedia (MM), 2019. 2