Low-Light Image and Video Enhancement Using Deep Learning: A Survey

Chongyi Li, Chunle Guo, Linghao Han, Jun Jiang, Ming-Ming Cheng, *Senior Member, IEEE*, Jinwei Gu, *Senior Member, IEEE*, and Chen Change Loy, *Senior Member, IEEE*

Abstract— Low-light image enhancement (LLIE) aims at improving the perception or interpretability of an image captured in an environment with poor illumination. Recent advances in this area are dominated by deep learning-based solutions, where many learning strategies, network structures, loss functions, training data, etc. have been employed. In this paper, we provide a comprehensive survey to cover various aspects ranging from algorithm taxonomy to unsolved open issues. To examine the generalization of existing methods, we propose a low-light image and video dataset, in which the images and videos are taken by different mobile phones' cameras under diverse illumination conditions. Besides, for the first time, we provide a unified online platform that covers many popular LLIE methods, of which the results can be produced through a user-friendly web interface. In addition to qualitative and quantitative evaluation of existing methods on publicly available and our proposed datasets, we also validate their performance in face detection in the dark. This survey together with the proposed dataset and online platform could serve as a reference source for future study and promote the development of this research field. The proposed platform and dataset as well as the collected methods, datasets, and evaluation metrics are publicly available and will be regularly updated. Project page: https://www.mmlab-ntu.com/project/lliv_survey/index.html.

Index Terms—image and video restoration, low-light image dataset, low-light image enhancement platform, computational photography.

1 Introduction

Images are often taken under sub-optimal lighting conditions, under the influence of backlit, uneven light, and dim light, due to inevitable environmental and/or technical constraints such as insufficient illumination and limited exposure time. Such images suffer from the compromised aesthetic quality and unsatisfactory transmission of information for high-level tasks such as object tracking, recognition, and detection. Figure 1 shows some examples of the degradations induced by sub-optimal lighting conditions.

Low-light enhancement enjoys a wide range of applications in different areas, including visual surveillance, autonomous driving, and computational photography. In particular, smartphone photography has become ubiquitous and prominent. Limited by the size of the camera aperture, the requirement of real-time processing, and the constraint of memory, taking photographs with a smartphone's camera in a dim environment is especially challenging. There is an exciting research arena of enhancing low-light images and videos in such applications.

Traditional methods for low-light enhancement include Histogram Equalization-based methods [35], [36] and Retinex model-based methods [37], [38], [39], [40], [41], [42], [43], [44]. The latter received relatively more attention. A typical Retinex model-based approach decomposes a low-light image into a reflection component and an illumination component by priors or regularizations. The estimated reflection component is treated as the enhanced result. Such methods have some limitations: **1)** the ideal assumption that treats the reflection component as the enhanced result does not always hold, especially given various illumination properties, which could lead to unrealistic enhancement such as loss of details and distorted colors, **2)** the noise is usually ignored in the Retinex model, thus it is remained or amplified in the enhanced results, **3)** finding an effective prior or regularization is challenging. Inaccurate prior or regularization may result in artifacts and color deviations in the enhanced results, and **4)** the runtime is relatively long because of their complicated optimization process.

(a) back lit (b) uneven light (c) dim light

(d) extremely low (e) colored light (f) boosted noise

Fig. 1. Examples of images taken under sub-optimal lighting conditions. These images suffer from the buried scene content, reduced contrast, boosted noise, and inaccurate color.

Recent years have witnessed the compelling success of deep learning-based LLIE since the first seminal work [1]. Deep learning-based solutions enjoy better accuracy, robustness, and speed over conventional methods, thus attracting

C. Li and C. C. Loy are with the S-Lab, Nanyang Technological University (NTU), Singapore (e-mail: chongyi.li@ntu.edu.sg and ccloy@ntu.edu.sg).
C. Guo, L. Han, and M. M. Cheng are with the College of Computer Science, Nankai University, Tianjin, China (e-mail: guochunle@nankai.edu.cn, lhhan@mail.nankai.edu.cn, and cmm@nankai.edu.cn).
J. Jiang and J. Gu are with the SenseTime (e-mail: jiangjun@sensebrain.site and gujinwei@sensebrain.site).
C. Li and C. Guo contribute equally.
C. C. Loy is the corresponding author.

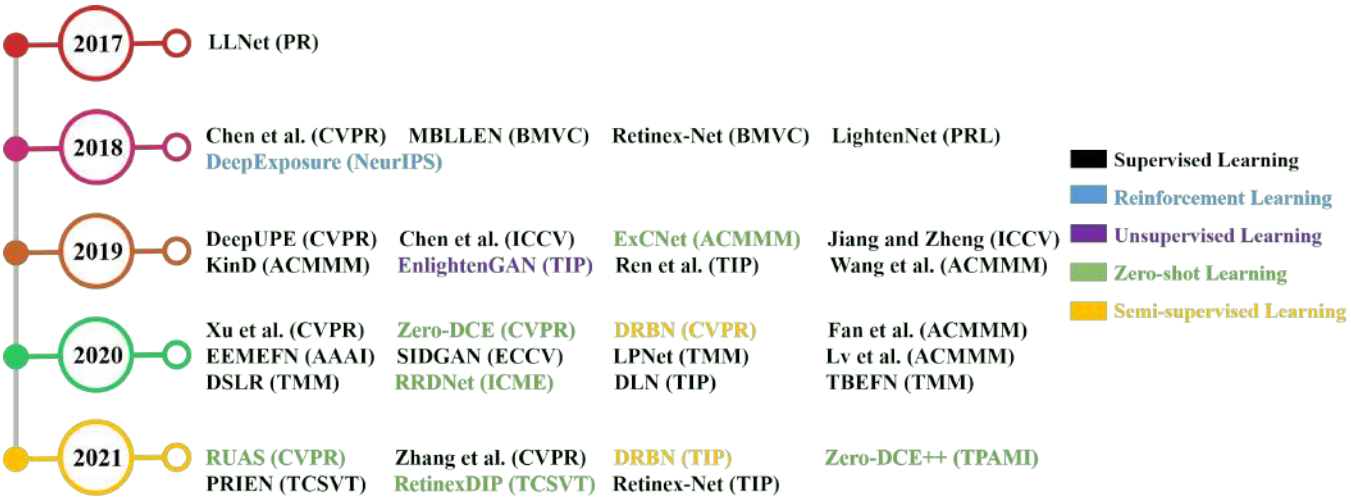


Fig. 2. A concise milestone of deep learning-based low-light image and video enhancement methods. **Supervised learning-based methods**: LLNet [1], Chen et al. [2], MBLLEN [3], Retinex-Net [4], LightenNet [5], SCIE [6], DeepUPE [7], Chen et al. [8], Jiang and Zheng [9], Wang et al. [10], KinD [11], Ren et al. [12], Xu et al. [13], Fan et al. [14], Lv et al. [15], EEMEFN [16], SIDGAN. [17], LPNet [18], DLN [19], TBEFN [20], DSLR [21], Zhang et al. [22], PRIEN [23], and Retinex-Net [24]. **Reinforcement learning-based method**: DeepExposure [25]. **Unsupervised learning-based method**: EnlightenGAN [26]. **Zero-shot learning-based methods**: ExCNet [27], Zero-DCE [28], RRDNet [29], Zero-DCE++ [30], RetinexDIP [31], and RUAS [32]. **Semi-supervised learning-based method**: DRBN [33] and DRBN [34].

increasing attention. A concise milestone of deep learning-based LLIE methods is shown in Figure 2. As shown, since 2017, the number of deep learning-based solutions has grown year by year. Learning strategies used in these solutions cover Supervised Learning (SL), Reinforcement Learning (RL), Unsupervised Learning (UL), Zero-Shot Learning (ZSL), and Semi-Supervised Learning (SSL). Note that we only report some representative methods in Figure 2. In fact, there are more than 100 papers on deep learning-based methods from 2017 to 2021. Moreover, although some general photo enhancement methods [45], [46], [47], [48], [49], [50], [51], [52], [53] can improve the brightness of images to some extent, we omit them in this survey as they are not designed to handle diverse low-light conditions. We concentrate on deep learning-based solutions that are specially developed for low-light image and video enhancement.

Despite deep learning has dominated the research of LLIE, an in-depth and comprehensive survey on deep learning-based solutions is lacking. There are two reviews of LLIE [54], [55]. Wang et al. [54] mainly reviews conventional LLIE methods while our work systematically and comprehensively reviews recent advances of deep learning-based LLIE. In comparison to Liu et al. [55] that reviews existing LLIE algorithms, measures the machine vision performance of different methods, provides a low-light image dataset serving both low-level and high-level vision enhancement, and develops an enhanced face detector, our survey reviews the low-light image and video enhancement from different aspects and has the following unique characteristics. **1)** Our work mainly focuses on recent advances of deep learning-based low-light image and video enhancement, where we provide in-depth analysis and discussion in various aspects, covering learning strategies, network structures, loss functions, training datasets, test datasets, evaluation metrics, model sizes, inference speed, enhancement performance, etc. Thus, this survey centers on deep learning and its applications in low-light image and video enhancement. **2)** We propose a dataset that contains images and videos captured by different mobile phones' cameras under diverse illumination conditions to evaluate the generalization of existing methods. This new and challenging dataset is a supplement of existing low-light image and video enhancement datasets as such a dataset is lacking in this research area. Besides, we are the first, to the best of our knowledge, to compare the performance of deep learning-based low-light image enhancement methods on this kind of data. **3)** We provide an online platform that covers many popular deep learning-based low-light image enhancement methods, where the results can be produced by a user-friendly web interface. With our platform, one without any GPUs can assess the results of different methods for any input images online, which speeds up the development of this research field and helps to create new research. We hope that our survey could provide novel insights and inspiration to facilitate the understanding of deep learning-based LLIE, foster research on the raised open issues, and speed up the development of this research field.

2 Deep Learning-Based LLIE

2.1 Problem Definition

We first give a common formulation of the deep learning-based LLIE problem. For a low-light image $I \in \mathbb{R}^{W \times H \times 3}$ of width W and height H, the process can be modeled as:

$$\widehat{R} = \mathcal{F}(I; \theta), \tag{1}$$

where $\widehat{R} \in \mathbb{R}^{W \times H \times 3}$ is the enhanced result and $\mathcal{F}$ represents the network with trainable parameters θ. The purpose of deep learning is to find optimal network parameters $\widehat{\theta}$ that minimizes the error:

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}(\widehat{R}, R), \tag{2}$$

where $R \in \mathbb{R}^{W \times H \times 3}$ is the ground truth, and the loss function $\mathcal{L}(\widehat{R}, R)$ drives the optimization of network. Various

loss functions such as supervised loss and unsupervised loss can be used. More details will be presented in Section 3.

2.2 Learning Strategies

According to different learning strategies, we categorize existing LLIE methods into supervised learning, reinforcement learning, unsupervised learning, zero-shot learning, and semi-supervised learning. A statistic analysis from different perspectives is presented in Figure 3. In what follows, we review some representative methods of each strategy.

Supervised Learning. For supervised learning-based LLIE methods, they are further divided into end-to-end, deep Retinex-based, and realistic data-driven methods.

The first deep learning-based LLIE method LLNet [1] employs a variant of stacked-sparse denoising autoencoder [56] to brighten and denoise low-light images simultaneously. This pioneering work inspires the usage of end-to-end networks in LLIE. Lv et al. [3] propose an end-to-end multi-branch enhancement network (MBLLEN). The MBLLEN improves the performance of LLIE via extracting effective feature representations by a feature extraction module, an enhancement module, and a fusion module. The same authors [15] propose other three subnetworks including an Illumination-Net, a Fusion-Net, and a Restoration-Net to further improve the performance. Ren et al. [12] design a more complex end-to-end network that comprises an encoder-decoder network for image content enhancement and a recurrent neural network for image edge enhancement. Similar to Ren et al. [12], Zhu et al. [16] propose a method called EEMEFN. The EEMEFN consists of two stages: multi-exposure fusion and edge enhancement. A multi-exposure fusion network, TBEFN [20], is proposed for LLIE. The TBEFN estimates a transfer function in two branches, of which two enhancement results can be obtained. At last, a simple average scheme is employed to fuse these two images and further refine the result via a refinement unit. In addition, pyramid network (LPNet) [18], residual network [19], and Laplacian pyramid [21] (DSLR) are introduced into LLIE. These methods learn to effectively and efficiently integrate feature representations via commonly used end-to-end network structures for LLIE. Based on the observation that noise exhibits different levels of contrast in different frequency layers, Xu et al. [57] proposed a frequency-based decomposition-and-enhancement network. This network recovers image contents with noise suppression in the low-frequency layer while inferring the details in the high-frequency layer. Recently, a progressive-recursive low-light image enhancement network [23] is proposed, which uses a recursive unit to gradually enhance the input image. To solve temporal instability when handling low-light videos, Zhang et al. [22] propose to learn and infer motion field from a single image then enforce temporal consistency.

In comparison to directly learning an enhanced result in an end-to-end network, deep Retinex-based methods enjoy better enhancement performance in most cases owing to the physically explicable Retinex theory [58], [59]. Deep Retinex-based methods usually separately enhance the illuminance component and the reflectance components via specialized subnetworks. A Retinex-Net [4] is proposed, which includes a Decom-Net that splits the input image into light-independent reflectance and structure-aware smooth illumination and an Enhance-Net that adjusts the illumination map for low-light enhancement. Recently, the Retinex-Net [4] is extended by adding new constraints and advanced network designs for better enhancement performance [24]. To reduce the computational burden, Li et al. [5] propose a lightweight LightenNet for weakly illuminated image enhancement, which only consists of four layers. The LightenNet takes a weakly illuminated image as the input and then estimates its illumination map. Based on the Retinex theory [58], [59], the enhanced image is obtained by dividing the input image by the illumination map. To accurately estimate the illumination map, Wang et al. [60] extract the global and local features to learn an image-to-illumination mapping by their proposed DeepUPE network. Zhang et al. [11] separately develop three subnetworks for layer decomposition, reflectance restoration, and illumination adjustment, called KinD. Furthermore, the authors alleviate the visual defects left in the results of KinD [11] by a multi-scale illumination attention module. The improved KinD is called KinD++ [61]. To solve the issue that the noise is omitted in the deep Retinex-based methods, Wang et al. [10] propose a progressive Retinex network, where an IM-Net estimates the illumination and a NM-Net estimates the noise level. These two subnetworks work in a progressive mechanism until obtaining stable results. Fan et al. [14] integrate semantic segmentation and Retinex model for further improving the enhancement performance in real cases. The core idea is to use semantic prior to guide the enhancement of both the illumination component and the reflectance component.

Although some methods can achieve decent performance, they show poor generalization capability in real low-light cases due to the usage of synthetic training data. To solve this issue, some works attempt to generate more realistic training data or capture real data. Cai et al. [6] build a multi-exposure image dataset, where the low-contrast images of different exposure levels have their corresponding high-quality reference images. Each high-quality reference image is obtained by subjectively selecting the best output from 13 results enhanced by different methods. Moreover, a frequency decomposition network is trained on the built dataset and separately enhances the high-frequency layer and the low-frequency layer via a two-stage structure. Chen et al. [2] collect a real low-light image dataset (SID) and train the U-Net [62] to learn a mapping from low-light raw data to the corresponding long-exposure high-quality reference image. Further, Chen et al. [8] extend the SID dataset to low-light videos (DRV). The DRV contains static videos with the corresponding long-exposure ground truths. To ensure the generalization capability of processing the videos of dynamic scenes, a siamese network is proposed. To enhance the moving objects in the dark, Jiang and Zheng [9] design a co-axis optical system to capture temporally synchronized and spatially aligned low-light and well-lighted video pairs (SMOID). Unlike the DRV video dataset [8], the SMOID video dataset contains dynamic scenes. To learn the mapping from raw low-light video to well-lighted video, a 3D U-Net-based network is proposed. Considering the limitations of previous low-light video datasets such as DRV dataset [8] only containing statistic videos and SMOID dataset [9] only

having 179 video pairs, Triantafyllidou et al. [17] propose a low-light video synthesis pipeline, dubbed SIDGAN. The SIDGAN can produce dynamic video data (raw-to-RGB) by a semi-supervised dual CycleGAN with intermediate domain mapping. To train this pipeline, the real-world videos are collected from Vimeo-90K dataset [63]. The low-light raw video data and the corresponding long-exposure images are sampled from DRV dataset [8]. With the synthesized training data, this work adopts the same U-Net network as Chen et al. [2] for low-light video enhancement.

Reinforcement Learning. Without paired training data, Yu et al. [25] learn to expose photos with reinforcement adversarial learning, named DeepExposure. Specifically, an input image is first segmented into sub-images according to exposures. For each sub-image, local exposure is learned by the policy network sequentially based on reinforcement learning. The reward evaluation function is approximated by adversarial learning. At last, each local exposure is employed to retouch the input, thus obtaining multiple retouched images under different exposures. The final result is achieved by fusing these images.

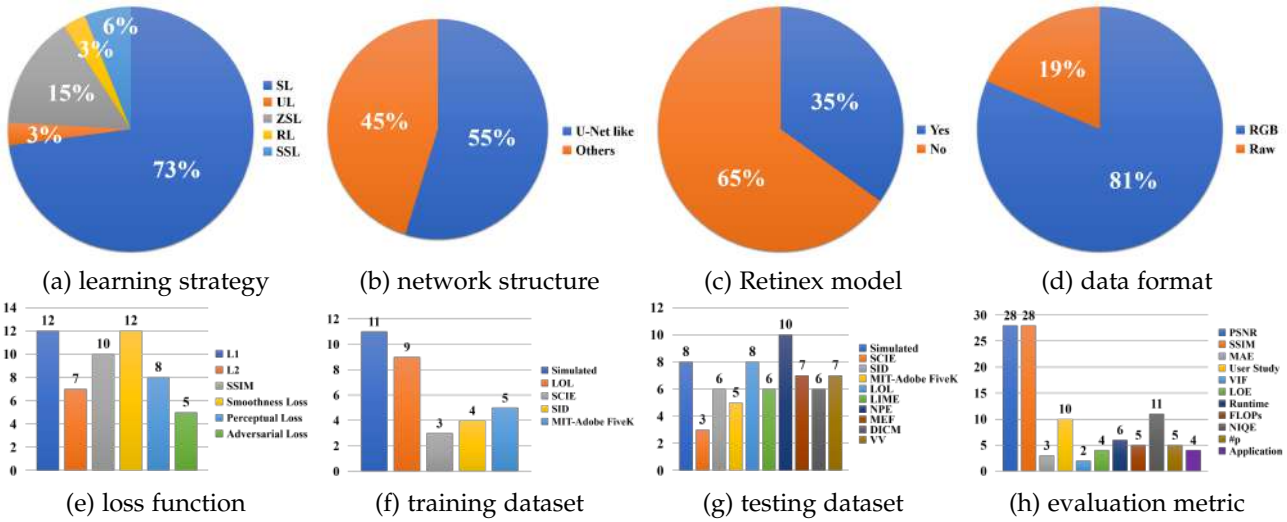
Unsupervised Learning. Training a deep model on paired data may result in overfitting and limited generalization capability. To solve this issue, an unsupervised learning method named EnligthenGAN [26] is proposed. The EnlightenGAN adopts an attention-guided U-Net [62] as the generator and uses the global-local discriminators to ensure the enhanced results look like realistic normal-light images. In addition to global and local adversarial losses, the global and local self feature preserving losses are proposed to preserve the image content before and after the enhancement. This is a key point for the stable training of such a one-path Generative Adversarial Network (GAN) structure.

Zero-Shot Learning. The supervised learning, reinforcement learning, and unsupervised learning methods either have limited generalization capability or suffer from unstable training. To remedy these issues, zero-shot learning is proposed to learn the enhancement solely from the testing images. Note that the concept of zero-shot learning in the low-level vision tasks is used to emphasize that the method does not require paired or unpaired training data, which is different from its definition in high-level visual tasks. Zhang et al. [27] propose a zero-shot learning method, called ExCNet, for back-lit image restoration. A network is first used to estimate the S-curve that best fits the input image. Once the S-curve is estimated, the input image is separated into a base layer and a detail layer using the guided filter [64]. Then the base layer is adjusted by the estimated S-curve. Finally, the Weber contrast [65] is used to fuse the detailed layer and the adjusted base layer. To train the ExCNet, the authors formulate the loss function as a block-based energy minimization problem. Zhu et al. [29] propose a three-branch CNN, called RRDNet, for underexposed images restoration. The RRDNet decomposes an input image into illumination, reflectance, and noise via iteratively minimizing specially designed loss functions. To drive the zero-shot learning, a combination of Retinex reconstruction loss, texture enhancement loss, and illumination-guided noise estimation loss is proposed. Zhao et al. [31] perform Retinex decomposition via neural networks and then enhance the low-light image based on the Retinex model, called RetinexDIP. Inspired by Deep Image Prior (DIP) [66], RetinexDIP generates the reflectance component and illumination component of an input image by randomly sampled white noise, in which the component characteristics-related losses such as illumination smoothness are used for training. Liu et al. [32] propose a Retinex-inspired unrolling method for LLIE, in which the cooperative architecture search is used to discover lightweight prior architectures of basic blocks and non-reference losses are used to train the network. Different from the image reconstruction-based methods [1], [3], [4], [11], [12], [21], [61], a deep curve estimation network, Zero-DCE [28], is proposed. Zero-DCE formulates the light enhancement as a task of image-specific curve estimation, which takes a low-light image as input and produces high-order curves as its output. These curves are used for pixel-wise adjustment on the dynamic range of the input to obtain an enhanced image. Further, an accelerated and lightweight version is proposed, called Zero-DCE++ [30]. Such curve-based methods do not require any paired or unpaired data during training. They achieve zero-reference learning via a set of non-reference loss functions. Besides, unlike the image reconstruction-based methods that need high computational resources, the image-to-curve mapping only requires lightweight networks, thus achieving a fast inference speed.

Semi-Supervised Learning. To combine the strengths of supervised learning and unsupervised learning, semi-supervised learning has been proposed in recent years. Yang et al. [33] propose a semi-supervised deep recursive band network (DRBN). The DRBN first recovers a linear band representation of an enhanced image under supervised learning, and then obtains an improved one by recomposing the given bands via a learnable linear transformation based on unsupervised adversarial learning. The DRBN is extended by introducing Long Short Term Memory (LSTM) networks and an image quality assessment network pre-trained on an aesthetic visual analysis dataset, which achieves better enhancement performance [34].

Observing Figure 3(a), we can find that supervised learning is the mainstream among deep learning-based LLIE methods, of which the percentage reaches 73%. This is because supervised learning is relatively easy when paired training data such as LOL [4], SID [2] and diverse low-/normal-light image synthesis approaches are used. However, supervised learning-based methods suffer from some challenges: **1)** collecting a large-scale paired dataset that covers diverse real-world low-light conditions is difficult, **2)** synthetic low-light images do not accurately represent real-world illuminance conditions such as spatially varying lighting and different levels of noise, and **3)** training a deep model on paired data may result in limited generalization to real-world images of diverse illumination properties.

Therefore, some methods adopt unsupervised learning, reinforcement learning, semi-supervised learning, and zero-shot learning to bypass the challenges in supervised learning. Although these methods achieve competing performance, they still suffer from some limitations: **1)** for unsupervised learning/semi-supervised learning methods, how to implement stable training, avoid color deviations, and build the relations of cross-domain information challenges current methods, **2)** for reinforcement learning methods, designing an effective reward mechanism and implementing



(a) learning strategy (b) network structure (c) Retinex model (d) data format

(e) loss function (f) training dataset (g) testing dataset (h) evaluation metric

Fig. 3. A statictic analysis of deep learning-based LLIE methods, including learning strategy, network characteristic, Retinex model, data format, loss function, training dataset, testing dataset, and evaluation metric. Best viewed by zooming in.

efficient and stable training are intricate, and **3)** for zero-shot learning methods, the design of non-reference losses is non-trivial when the color preservation, artifact removal, and gradient back-propagation should be taken into account.

3 Technical Review and Discussion

In this section, we first summarize the representative deep learning-based LLIE methods in Table 1, then analyze and discuss their technical characteristics.

3.1 Network Structure

Diverse network structures and designs have been used in the existing models, spanning from the basic U-Net, pyramid network, multi-stage network to frequency decomposition network. After analyzing Figure 3(b), it can be observed that the U-Net and U-Net-like networks are mainly adopted network structures in LLIE. This is because U-Net can effectively integrate multi-scale features and employ both low-level and high-level features. Such characteristics are essential for achieving satisfactory low-light enhancement.

Nevertheless, some key issues may be ignored in the current LLIE network structures: **1)** after going through several convolutional layers, the gradients of an extremely low light image may vanish during the gradient back-propagation due to its small pixel values. This would degrade the enhancement performance and affect the convergence of network training, **2)** the skip-connections used in the U-Net-like networks might introduce noise and redundant features into the final results. How to effectively filter out the noise and integrate both low-level and high-level features should be carefully considered, and **3)** although some designs and components are proposed for LLIE, most of them are borrowed or modified from related low-level visual tasks. The characteristics of low-light data should be considered when designing the network structures.

3.2 Combination of Deep Model and Retinex Theory

As presented in Figure 3(c), almost 1/3 of methods combine the designs of deep networks with the Retinex theory, e.g., designing different subnetworks to estimate the components of the Retinex model and estimating the illumination map to guide the learning of networks. Despite such a combination can bridge deep learning-based and model-based methods, their respective weaknesses may be introduced into the final models: **1)** the ideal assumption that the reflectance is the final enhanced result used in Retinex-based LLIE methods would still affect the final results, and **2)** the risk of overfitting in deep networks still exists despite the use of Retinex theory. How to cream off the best and filter out the impurities should be carefully considered when researchers combine deep learning with the Retinex theory.

3.3 Data Format

As shown in Figure 3(d), RGB data format dominates most methods as it is commonly found as the final imagery form produced by smartphone cameras, Go-Pro cameras, and drone cameras. Although raw data are limited to specific sensors such as those based on Bayer patterns, the data cover wider color gamut and higher dynamic range. Hence, deep models trained on raw data usually recover clear details and high contrast, obtain vivid color, reduce the effects of noises and artifacts, and improve the brightness of extremely low-light images. In future research, a smooth transformation from raw data of different patterns to RGB format would have the potentials to combine the convenience of RGB data and the advantage of high-quality enhancement of raw data for LLIE.

3.4 Loss Function

In Figure3(e), the commonly adopted loss functions in LLIE models include reconstruction loss (L_1, L_2, SSIM), perceptual loss, and smoothness loss. Besides, according to different demands and formulations, color loss, exposure loss, adversarial loss, etc are also adopted. We detail representative loss functions as follows.

Reconstruction Loss. Different reconstruction losses have their advantages and disadvantages. L_2 loss tends to penalize larger errors, but is tolerant to small errors. L_1 loss preserves colors and luminance well since an error is weighted

TABLE 1
Summary of essential characteristics of representative deep learning-based methods. "Retinex" indicates whether the models are Retinex-based or not. "simulated" means the testing data are simulated by the same approach as the synthetic training data. "self-selected" stands for the real-world images selected by the authors. "#P"represents the number of trainable parameters. "-" means this item is not available or not indicated in the paper.

	Method	Learning	Network Structure	Loss Function	Training Data	Testing Data	Evaluation Metric	Format	Platform	Retinex
2017	LLNet [1]	SL	SSDA	SRR loss	simulated by Gamma Correction & Gaussian Noise	simulated self-selected	PSNR SSIM	RGB	Theano	
2018	LightenNet [5]	SL	four layers	L_2 loss	simulated by random illumination values	simulated self-selected	PSNR MAE SSIM User Study	RGB	Caffe MATLAB	✓
	Retinex-Net [4]	SL	multi-scale network	L_1 loss smoothness loss invariable reflectance loss	LOL simulated by adjusting histogram	self-selected	-	RGB	TensorFlow	✓
	MBLLEN [3]	SL	multi-branch fusion	SSIM loss region loss perceptual loss	simulated by Gamma Correction & Poisson Noise	simulated self-selected	PSNR SSIM AB VIF LOE TOMI	RGB	TensorFlow	
	SCIE [6]	SL	frequency decomposition	L_2 loss L_1 loss SSIM loss	SCIE	SCIE	PSNR FSIM Runtime FLOPs	RGB	Caffe MATLAB	
	Chen et al. [2]	SL	U-Net	L_1 loss	SID	SID	PSNR SSIM	raw	TensorFlow	
	Deepexposure [25]	RL	policy network GAN	deterministic policy gradient adversarial loss	MIT-Adobe FiveK	MIT-Adobe FiveK	PSNR SSIM	raw	TensorFlow	
2019	Chen et al. [8]	SL	siamese network	L_1 loss self-consistency loss	DRV	DRV	PSNR SSIM MAE	raw	TensorFlow	
	Jiang and Zheng [9]	SL	3D U-Net	L_1 loss	SMOID	SMOID	PSNR SSIM MSE	raw	TensorFlow	
	DeepUPE [60]	SL	illumination map	L_1 loss color loss smoothness loss	retouched image pairs	MIT-Adobe FiveK	PSNR SSIM User Study	RGB	TensorFlow	✓
	KinD [11]	SL	three subnetworks U-Net	reflectance similarity loss illumination smoothness loss mutual consistency loss L_1 loss L_2 loss SSIM loss texture similarity loss illumination adjustment loss	LOL	LOL LIME NPE MEF	PSNR SSIM LOE NIQE	RGB	TensorFlow	✓
	Wang et al. [10]	SL	two subnetworks pointwise Conv	L_1 loss	simulated by camera imaging model	IP100 FNF38 MPI LOL NPE	PSNR SSIM NIQE	RGB	Caffe	✓
	Ren et al. [12]	SL	U-Net like network RNN dilated Conv	L_2 loss perceptual loss adversarial loss	MIT-Adobe FiveK with Gamma correction & Gaussion noise	simulated self-selected DPED	PSNR SSIM Runtime	RGB	Caffe	
	EnlightenGAN [26]	UL	U-Net like network	adversarial loss self feature preserving loss	unpaired real images	NPE LIME MEF DICM VV BBD-100K ExDARK	User Study NIQE Classification	RGB	PyTorch	
	ExCNet. [27]	ZSL	fully connected layers	energy minimization loss	real images	$IE_{ps}D$	User Study CDIQA LOD	RGB	PyTorch	
2020	Zero-DCE [28]	ZSL	U-Net like network	spatial consistency loss exposure control loss color constancy loss illumination smoothness loss	SICE	SICE NPE LIME MEF DICM VV DARK FACE	User Study PI PNSR SSIM MAE Runtime Face detection	RGB	PyTorch	
	DRBN [33]	SSL	recursive network	SSIM loss perceptual loss adversarial loss	LOL images selected by MOS	LOL	PSNR SSIM SSIM-GC	RGB	PyTorch	
	Lv et al. [15]	SL	U-Net like network	Huber loss SSIM loss perceptual loss illumination smoothness loss	simulated by a retouching module	LOL SICE DeepUPE	User Study PSNR SSIM VIF LOE NIQE #P Runtime Face detection	RGB	TensorFlow	✓
	Fan et al. [14]	SL	four subnetworks U-Net like network feature modulation	mutual smoothness loss reconstruction loss illumination smoothness loss cross entropy loss consistency loss SSIM loss gradient loss ratio learning loss	simulated by illumination adjustment, slight color distortion, and noise simulation	simulated self-selected	PSNR SSIM NIQE	RGB	-	✓
	Xu et al. [57]	SL	frequency decomposition U-Net like network	L_2 loss perceptual loss	SID in RGB	SID in RGB self-selected	PSNR SSIM	RGB	PyTorch	
	EEMEFN [16]	SL	U-Net like network edge detection network	L_1 loss weighted cross-entropy loss	SID	SID	PSNR SSIM	raw	TensorFlow PaddlePaddle	
	DLN [19]	SL	residual learning interactive factor back projection network	SSIM loss total variation loss	simulated by illumination adjustment, slight color distortion, and noise simulation	simulated LOL	User Study PSNR SSIM NIQE	RGB	PyTorch	
	LPNet [18]	SL	pyramid network	L_1 loss perceptual loss luminance loss	LOL SID in RGB MIT-Adobe FiveK	LOL SID in RGB MIT-Adobe FiveK MEF NPE DICM VV	PSNR SSIM NIQE #P FLOPs Runtime	RGB	PyTorch	
	SIDGAN [17]	SL	U-Net	CycleGAN loss	SIDGAN	SIDGAN	PSNR SSIM TPSNR TSSIM ATWE	raw	TensorFlow	
	RRDNet [29]	ZSL	three subnetworks	retinex reconstruction loss texture enhancement loss noise estimation loss	-	NPE LIME MEF DICM	NIQE CPCQI	RGB	PyTorch	✓
	TBEFN [20]	SL	three stages U-Net like network	SSIM loss perceptual loss smoothness loss	SCIE LOL	SCIE LOL DICM MEF NPE VV	PSNR SSIM NIQE Runtime #P FLOPs	RGB	TensorFlow	✓
	DSLR [21]	SL	Laplacian pyramid U-Net like network	L_2 loss Laplacian loss color loss	MIT-Adobe FiveK	MIT-Adobe FiveK self-selected	PSNR SSIM NIQMC NIQE BTMQI CaHDC	RGB	PyTorch	
2021	RUAS [32]	ZSL	neural architecture search	cooperative loss similar loss total variation loss	LOL MIT-Adobe FiveK	LOL MIT-Adobe FiveK	PSNR SSIM Runtime #P FLOPs	RGB	PyTorch	✓
	Zhang et al. [22]	SL	U-Net	L_1 loss consistency loss	simulated by illumination adjustmentand noise simulation	simulated self-selected	User Study PSNR SSIM AB MABD WE	RGB	PyTorch	
	Zero-DCE++ [30]	ZSL	U-Net like network	spatial consistency loss exposure control loss color constancy loss illumination smoothness loss	SICE	SICE NPE LIME MEF DICM VV DARK FACE	User Study PI PNSR SSIM #P MAE Runtime Face detection FLOPs	RGB	PyTorch	
	DRBN [34]	SSL	recursive network	perceptual loss detail loss quality loss	LOL	LOL	PSNR SSIM SSIM-GC	RGB	PyTorch	
	Retinex-Net [24]	SL	three subnetworks	L_1 loss L_2 loss SSIM loss total variation loss	LOL simulated by adjusting histogram	LOL simulated NPE DICM VV	PNSR SSIM UQI OSS User Study	RGB	PyTorch	✓
	RetinexDIP [31]	ZSL	encoder-decoder networks	reconstruction loss illumination-consistency loss reflectnce loss illumination smoothness loss	-	DICM, ExDark Fusion LIME NASA NPE VV	NIQE NIQMC CPCQI	RGB	PyTorch	✓
	PRIEN [23]	SL	recursive network	SSIM loss	MEF LOL simulated by adjusting histogram	LOL LIME NPE MEF VV	PNSR SSIM LOE TMQI	RGB	PyTorch	

equally regardless of the local structure. SSIM loss preserves the structure and texture well. Refer to this research paper [67] for detailed analysis.

Perceptual Loss. Perceptual loss [68], particularly the feature reconstruction loss, is proposed to constrain the results similar to the ground truth in the feature space. The loss improves the visual quality of results. It is defined as the Euclidean distance between the feature representations of an enhanced result and those of corresponding ground truth. The feature representations are typically extracted from the VGG network [69] pre-trained on ImageNet dataset [70].

Smoothness Loss. To remove noise in the enhanced results or preserve the relationship of neighboring pixels, smoothness loss (TV loss) is often used to constrain the enhanced result or the estimated illumination map.

Adversarial Loss. To encourage enhanced results to be indistinguishable from reference images, adversarial learning solves a max-min optimization problem [71], [72].

Exposure Loss. As one of key non-reference losses, exposure loss measures the exposure levels of enhanced results without paired or unpaired images as reference images.

The commonly used loss functions in LLIE networks are also employed in image reconstruction networks for image super-resolution [73], image denoising [74], image detraining [75], [76], [77], and image deblurring [78]. Different from these versatile losses, the specially designed exposure loss for LLIE inspires the design of non-reference losses. A non-reference loss makes a model enjoying better generalization capability. It is an on-going research to consider image characteristics for the design of loss functions.

3.5 Training Datasets

Figure 3(f) reports the usage of a variety of paired training datasets for training low-light enhancement networks. These datasets include real-world captured datasets and synthetic datasets. We list them in Table 2.

Simulated by Gamma Correction. Owing to its nonlinearity and simplicity, Gamma correction is used to adjust the luminance or tristimulus values in video or still image systems. It is defined by a power-law expression:

$$V_{\text{out}} = AV_{\text{in}}^{\gamma}, \tag{3}$$

where the input V_{in} and output V_{out} are typically in the range of [0,1]. The constant A is set to 1 in the common case. The power γ controls the luminance of the output. Intuitively, the input is brightened when γ <1 while the input is darkened when γ >1. The input can be the three RGB channels of an image or the luminance-related channels such as L channel in the CIELab color space and Y channel in the YCbCr color space. After adjusting the luminance-related channel using Gamma correction, the corresponding channels in the color space are adjusted by equal proportion to avoid producing artifacts and color deviations.

To simulate images taken in real-world low-light scenes, Gaussian noise, Poisson noise, or realistic noise is added to the Gamma corrected images. The low-light image synthesized using Gamma correction can be expressed as:

$$I_{\text{low}} = n(g(I_{\text{in}}; \gamma)), \tag{4}$$

where n represents the noise model, $g(I_{\text{in}}; \gamma)$ represents the Gamma correction function with Gamma value γ, I_{in} is a normal-light and high-quality image or luminance-related channel. Although this function produces low-light images of different lighting levels by changing the Gamma value γ, it tends to introduce artifacts and color deviations into the synthetic low-light images due to the nonlinear adjustment.

TABLE 2
Summary of paired training datasets. 'Syn' represents Synthetic.

Name	Number	Format	Real/Syn	Video
Gamma Correction	$+\infty$	RGB	Syn	
Random Illumination	$+\infty$	RGB	Syn	
LOL [4]	500	RGB	Real	
SCIE [6]	4,413	RGB	Real	
VE-LOL-L [55]	2,500	RGB	Real+Syn	
MIT-Adobe FiveK [79]	5,000	raw	Real	
SID [4]	5,094	raw	Real	
DRV [8]	202	raw	Real	✓
SMOID [9]	179	raw	Real	✓

Simulated by Random Illumination. According to the Retinex model, an image can be decomposed into a reflectance component and an illumination component. Assuming image content is independent of illumination component and local region in the illumination component have the same intensity, a low-light image can be obtained by

$$I_{\text{low}} = I_{\text{in}}L, \tag{5}$$

where L is a random illumination value in the range of [0,1]. Noises can be added to the synthetic image. Such a linear function avoids artifacts, but the strong assumption requires the synthesis to operate only on image patches where local regions have the same brightness. A deep model trained on such image patches may lead to sub-optimal performance due to the negligence of context information.

LOL. LOL [4] is the first paired low-/normal-light image dataset taken in real scenes. The low-light images are collected by changing the exposure time and ISO. LOL contains 500 pairs of low-/normal-light images of size 400×600 saved in RGB format.

SCIE. SCIE is a multi-exposure image dataset of low-contrast and good-contrast image pairs. It includes multi-exposure sequences of 589 indoor and outdoor scenes. Each sequence has 3 to 18 low-contrast images of different exposure levels, thus containing 4,413 multi-exposure images in total. The 589 high-quality reference images are obtained by selecting from the results of 13 representative enhancement algorithms. That is many multi-exposure images have the same high-contrast reference image. The image resolutions are between 3,000×2,000 and 6,000×4,000. The images in SCIE are saved in RGB format.

MIT-Adobe FiveK. MIT-Adobe FiveK [79] was collected for global tone adjustment but has been used in LLIE. This is because the input images have low light and low contrast. MIT-Adobe FiveK contains 5,000 images, each of which is retouched by 5 trained photographers towards visually pleasing renditions, akin to a postcard. The images are all in raw format. To train the networks that can handle images of RGB format, one needs to use Adobe Lightroom to preprocess the images and save them as RGB format following a dedicated pipeline[1]. The images are commonly resized to have a long edge of 500 pixels.

1. https://github.com/nothinglo/Deep-Photo-Enhancer/issues/38#issuecomment-449786636

SID. SID [2] contains 5,094 raw short-exposure images, each with a corresponding long-exposure reference image. The number of distinct long-exposure reference images is 424. In other words, multiple short-exposure images correspond to the same long-exposure reference image. The images were taken using two cameras: Sony α7S II and Fujifilm X-T2 in both indoor and outdoor scenes. Thus, the images have different sensor patterns (Sony camera' Bayer sensor and Fuji camera's APS-C X-Trans sensor). The resolution is 4,240×2,832 for Sony and 6,000×4,000 for Fuji. Usually, the long-exposure images are processed by libraw (a raw image processing library) and saved in the RGB color space, and randomly cropped 512×512 patches for training.

VE-LOL. VE-LOL [55] consists of two subsets: paired VE-LOL-L that is used for training and evaluating LLIE methods and unpaired VE-LOL-H that is used for evaluating the effect of LLIE methods on face detection. Specifically, VE-LOL-L includes 2,500 paired images. Among them, 1,000 pairs are synthetic, while 1,500 pairs are real. VE-LOL-H includes 10,940 unpaired images, where human faces are manually annotated with bounding boxes.

DRV. DRV [8] contains 202 static raw videos, each of which has a corresponding long-exposure ground truth. Each video was taken at approximately 16 to 18 frames per second in a continuous shooting mode and is with up to 110 frames. The images were taken by a Sony RX100 VI camera in both indoor and outdoor scenes, thus all in raw format of Bayer pattern. The resolution is 3,672×5,496.

SMOID. SMOID [9] contains 179 pairs of videos taken by a co-axis optical system, each of which has 200 frames. Thus, SMOID includes 35,800 extremely low light raw data of Bayer pattern and their corresponding well-lightened RGB counterparts. SMOID consists of moving vehicles and pedestrians under different illumination conditions.

Some issues challenge the aforementioned paired training datasets: **1)** deep models trained on synthetic data may introduce artifacts and color deviations when processing real-world images and videos due to the gap between synthetic data and real data, **2)** the scale and diversity of real training data are unsatisfactory, thus some methods incorporate synthetic data to augment the training data. This may lead to sub-optimal enhancement, and **3)** the input images and corresponding ground truths may exist misalignment due to the effects of motion, hardware, and environment. This would affect the performance of deep networks trained using pixel-wise loss functions.

3.6 Testing Datasets

In addition to the testing subsets in the paired datasets [2], [4], [6], [8], [9], [55], [79], there are several testing data collected from related works or commonly used for experimental comparisons. Besides, some datasets such as face detection in the dark [80] and detection and recognition in low-light images [81] are employed to test the effects of LLIE on high-level visual tasks. We summarize the commonly used testing datasets in Table 3 and introduce the representative testing datasets as follows.

BBD-100K. BBD-100K [84] is the largest driving video dataset with 10,000 videos taken over 1,100-hour driving experience across many different times in the day, weather conditions, and driving scenarios, and 10 tasks annotations. The videos taken at nighttime in BBD-100K are used to validate the effects of LLIE on high-level visual tasks and the enhancement performance in real scenarios.

2. https://sites.google.com/site/vonikakis/datasets

TABLE 3
Summary of testing datasets.

Name	Number	Format	Application	Video
LIME [39]	10	RGB		
NPE [37]	84	RGB		
MEF [82]	17	RGB		
DICM [83]	64	RGB		
VV[2]	24	RGB		
BBD-100K [84]	10,000	RGB	✓	✓
ExDARK [81]	7,363	RGB	✓	
DARK FACE [80]	6,000	RGB	✓	
VE-LOL-H [55]	10,940	RGB	✓	

ExDARK. ExDARK [81] dataset is built for object detection and recognition in low-light images. ExDARK dataset contains 7,363 low-light images from extremely low-light environments to twilight with 12 object classes annotated with image class labels and local object bounding boxes.

DARK FACE. DARK FACE [80] dataset contains 6,000 low-light images captured during the nighttime, each of which is labeled with bounding boxes of the human face.

From Figure 3(g) and Table 1, we can observe that one prefers using the self-collected testing data in the experiments. The main reasons lie into three-fold: **1)** besides the test partition of paired datasets, there is no acknowledged benchmark for evaluations, **2)** the commonly used test sets suffer from some shortcomings such as small scale (some test sets contain 10 images only), repeated content and illumination properties, and unknown experimental settings, and **3)** some of the commonly used testing data are not originally collected for evaluating LLIE. In general, current testing datasets may lead to bias and unfair comparisons.

3.7 Evaluation Metrics

Besides human perception-based subjective evaluations, image quality assessment (IQA) metrics, including both full-reference and non-reference IQA metrics, are able to evaluate image quality objectively. In addition, user study, number of trainable parameters, FLOPs, runtime, and applications also reflect the performance of LLIE models, as shown in Fig. 3(h). We will detail them as follows.

PSNR and MSE. PSNR and MSE are widely used IQA metrics. They are always non-negative, and values closer to infinite (PSNR) and zero (MSE) are better. Nevertheless, the pixel-wise PSNR and MSE may provide an inaccurate indication of the visual perception of image quality since they neglect the relation of neighboring pixels.

MAE. MAE represents the mean absolute error, serving as a measure of errors between paired observations. The smaller the MAE value is, the better similarity is.

SSIM. SSIM is used to measure the similarity between two images. It is a perception-based model that considers image degradation as perceived change in structural information. The value 1 is only reachable in the case of two identical sets of data, indicating perfect structural similarity.

LOE. LOE represents the lightness order error that reflects the naturalness of an enhanced image. For LOE, the smaller the LOE value is, the better the lightness order is preserved.

Application. Besides improving the visual quality, one of the purposes of image enhancement is to serve high-level visual tasks. Thus, the effects of LLIE on high-level visual applications are commonly examined to validate the performance of different methods.

The current evaluation approaches used in LLIE need to be improved in several aspects: **1)** although the PSNR, MSE, MAE, and SSIM are classic and popular metrics, they are still far from capturing real visual perception of human, **2)** some metrics are not originally designed for low-light images. They are used for assessing the fidelity of image information and contrast. Using these metrics may reflect the image quality, but they are far from the real purpose of low-light enhancement, **3)** metrics especially designed for low-light images are lacking, except for the LOE metric. Moreover, there is no metric for evaluating low-light video enhancement, and **4)** a metric that can balance both the human vision and the machine perception is expected.

4 Benchmarking and Empirical Analysis

This section provides empirical analysis and highlights some key challenges in deep learning-based LLIE. To facilitate the analysis, we propose a low-light image and video dataset to examine the performance of different solutions. We also develop the first online platform, where the results of LLIE models can be produced via a user-friendly web interface. In this section, we conduct extensive evaluations on several benchmarks and our proposed dataset.

In the experiments, we compare 13 representative RGB format-based methods, including eight supervised learning-based methods (LLNet [1], LightenNet [5], Retinex-Net [4], MBLLEN [3], KinD [11], KinD++ [61], TBEFN [20], DSLR [21]), one unsupervised learning-based method (EnlightenGAN [26]), one semi-supervised learning-based method (DRBN [33]), and three zero-shot learning-based methods (ExCNet [27], Zero-DCE [28], RRDNet [29]). Besides, we also compare two raw format-based methods, including SID [85] and EEMEFN [16]. Note that RGB format-based methods dominate LLIE. Moreover, most raw format-based methods do not release their code. Thus, we choose two representative methods to provide empirical analysis and insights. For all compared methods, we use the publicly available code to produce their results for fair comparisons.

4.1 A New Low-Light Image and Video Dataset

We propose a Low-Light Image and Video dataset, called LLIV-Phone, to comprehensively and thoroughly validate the performance of LLIE methods. LLIV-Phone is the largest and most challenging real-world testing dataset of its kind. In particular, the dataset contains 120 videos (45,148 images) taken by 18 different mobile phones' cameras including iPhone 6s, iPhone 7, iPhone7 Plus, iPhone8 Plus, iPhone 11, iPhone 11 Pro, iPhone XS, iPhone XR, iPhone SE, Xiaomi Mi 9, Xiaomi Mi Mix 3, Pixel 3, Pixel 4, Oppo R17, Vivo Nex, LG M322, OnePlus 5T, Huawei Mate 20 Pro under diverse illumination conditions (e.g., weak lighting, underexposure, moonlight, twilight, dark, extremely dark, backlit, non-uniform light, and colored light.) in both indoor and outdoor scenes. A summary of the LLIV-Phone dataset is provided in Table 4. We present several samples of LLIV-Phone dataset in Figure 4. The LLIV-Phone dataset is available at the project page.

TABLE 4
Summary of LLIV-Phone dataset. LLIV-Phone dataset contains 120 videos (45,148 images) taken by 18 different mobile phones' cameras. "#Video" and "#Image" represent the number of videos and images, respectively.

Phone's Brand	#Video	#Image	Resolution
iPhone 6s	4	1,029	1920×1080
iPhone 7	13	6,081	1920×1080
iPhone7 Plus	2	900	1920×1080
iPhone8 Plus	1	489	1280×720
iPhone 11	7	2,200	1920×1080
iPhone 11 Pro	17	7,739	1920×1080
iPhone XS	11	2,470	1920×1080
iPhone XR	16	4,997	1920×1080
iPhone SE	1	455	1920×1080
Xiaomi Mi 9	2	1,145	1920×1080
Xiaomi Mi Mix 3	6	2,972	1920×1080
Pixel 3	4	1,311	1920×1080
Pixel 4	3	1,923	1920×1080
Oppo R17	6	2,126	1920×1080
Vivo Nex	12	4,097	1280×720
LG M322	2	761	1920×1080
OnePlus 5T	1	293	1920×1080
Huawei Mate 20 Pro	12	4,160	1920×1080

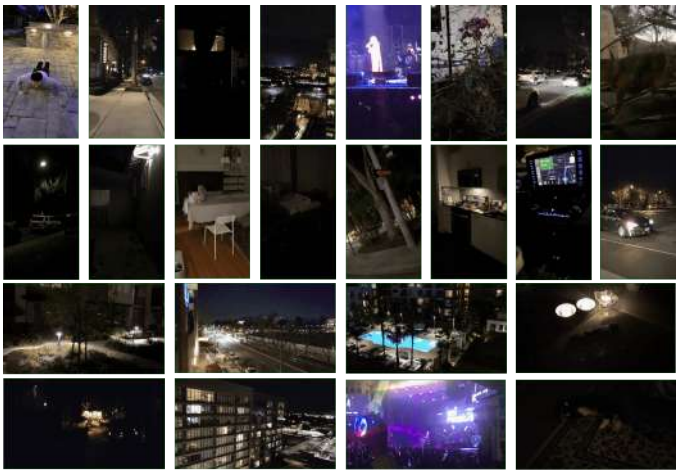


Fig. 4. Several images sampled from the proposed LLIV-Phone dataset. The images and videos are taken by different devices under diverse lighting conditions and scenes.

This challenging dataset is collected in real scenes and contains diverse low-light images and videos. Consequently, it is suitable for evaluating the generalization capability of different low-light image and video enhancement models. Notably, the dataset can be used as the training dataset for unsupervised learning and the reference dataset for synthesis methods to generate realistic low-light data.

4.2 Online Evaluation Platform

Different deep models may be implemented in different platforms such as Caffe, Theano, TensorFlow, and PyTorch. As a result, different algorithms demand different configurations, GPU versions, and hardware specifications. Such

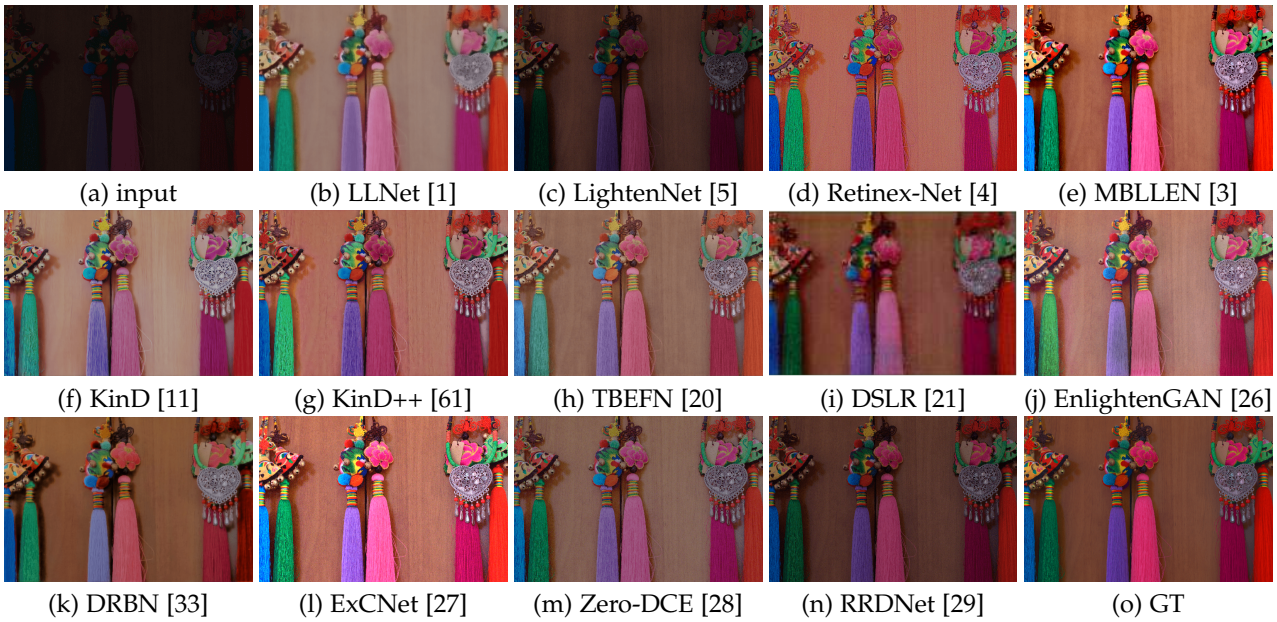


Fig. 5. Visual results of different methods on a low-light image sampled from LOL-test dataset.

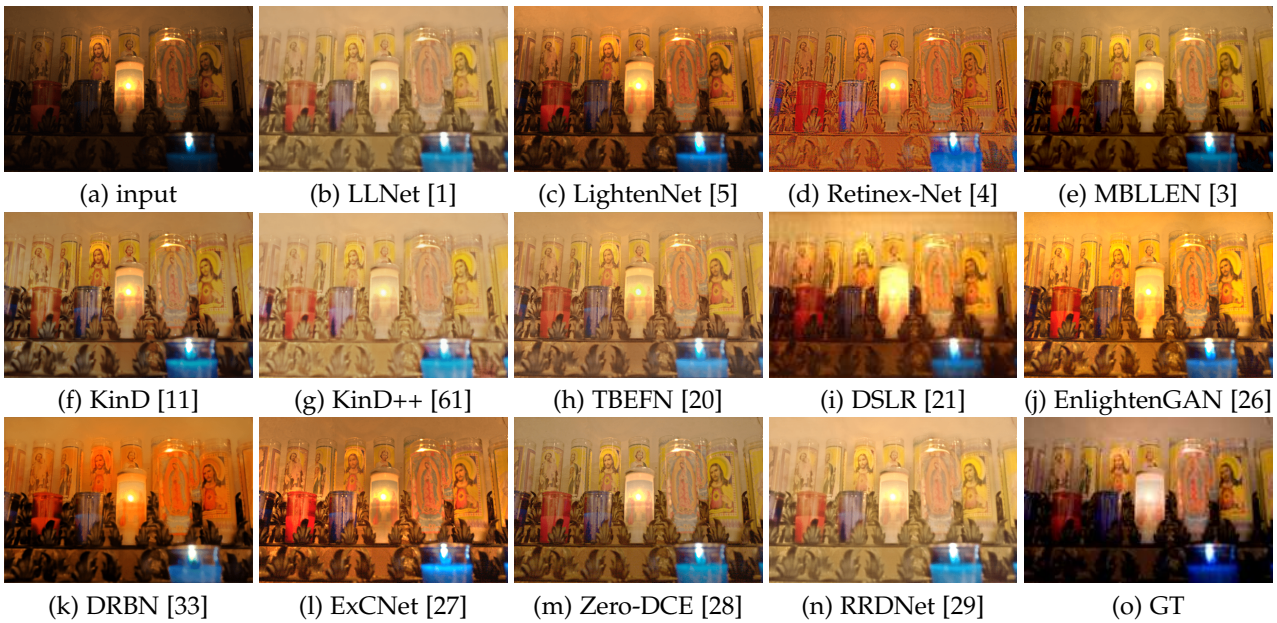


Fig. 6. Visual results of different methods on a low-light image sampled from MIT-Adobe FiveK-test dataset.

requirements are prohibitive to many researchers, especially for beginners who are new to this area and may not even have GPU resources. To resolve these problems, we develop an LLIE online platform, called LLIE-Platform, which is available at http://mc.nankai.edu.cn/ll/.

To the date of this submission, the LLIE-Platform covers 14 popular deep learning-based LLIE methods including LLNet [1], LightenNet [5], Retinex-Net [4], EnlightenGAN [26], MBLLEN [3], KinD [11], KinD++ [61], TBEFN [20], DSLR [21], DRBN [33], ExCNet [27], Zero-DCE [28], Zero-DCE++ [30], and RRDNet [29], where the results of any input can be produced through a user-friendly web interface. We will regularly offer new methods on this platform. We wish that this LLIE-Platform could serve the growing research community by providing users a flexible interface to run existing deep learning-based LLIE methods and develop their own new LLIE methods.

4.3 Benchmarking Results

To qualitatively and quantitatively evaluate different methods, in addition to the proposed LLIV-Phone dataset, we also adopt the commonly used LOL [4] and MIT-Adobe FiveK [79] datasets for RGB format-based methods, and SID [85] dataset for raw format-based methods. More visual results can be found in the supplementary material. The comparative results on the real low-light videos taken by different mobile phones' cameras can be found at YouTube https://www.youtube.com/watch?v=Elo9TkrG5Oo&t=6s.

We select five images on average from each video of the LLIV-Phone dataset, forming an image testing dataset

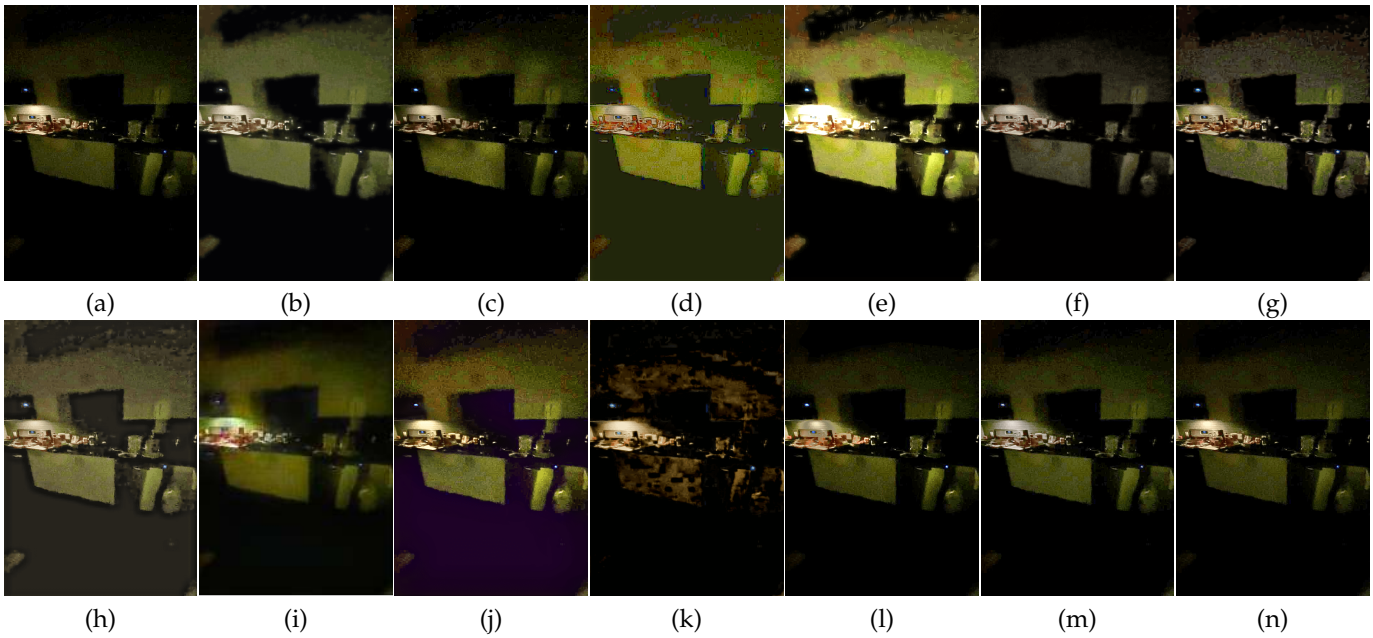


Fig. 7. Visual results of different methods on a low-light image sampled from LLIV-Phone-imgT dataset. (a) input. (b) LLNet [1]. (c) LightenNet [5]. (d) Retinex-Net [4]. (e) MBLLEN [3]. (f) KinD [11]. (g) KinD++ [61]. (h) TBEFN [20]. (i) DSLR [21]. (j) EnlightenGAN [26]. (k) DRBN [33]. (l) ExCNet [27]. (m) Zero-DCE [28]. (n) RRDNet [29].

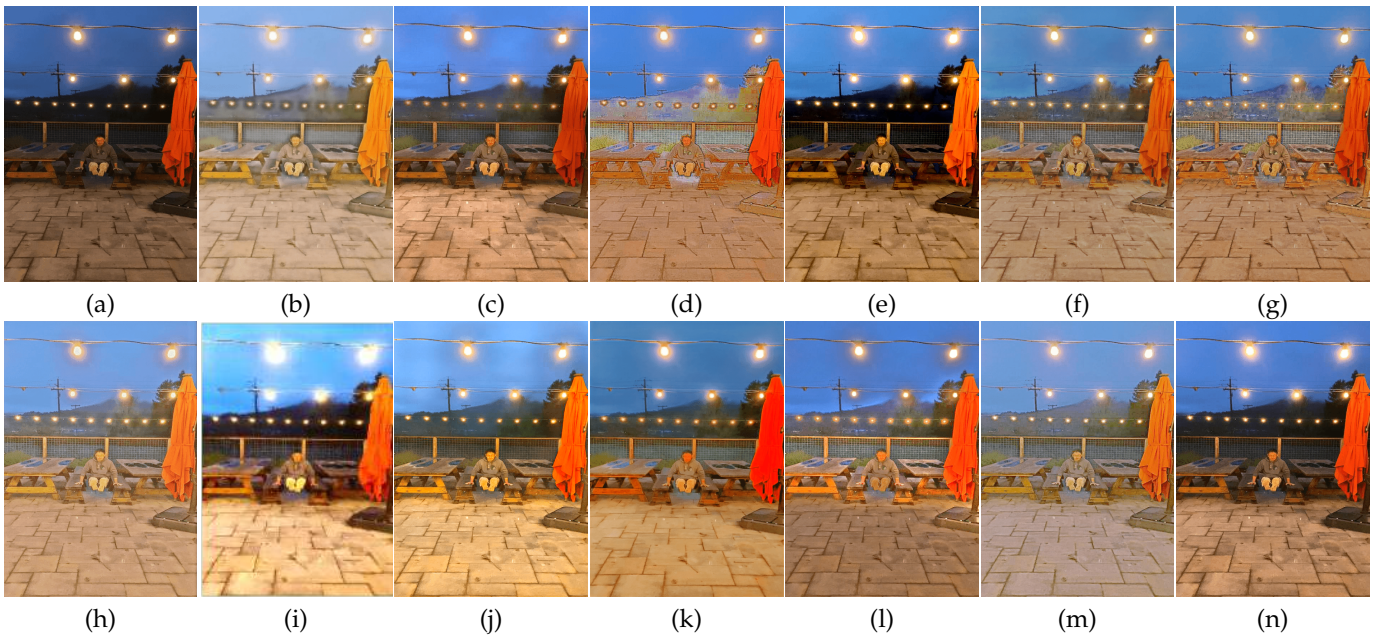


Fig. 8. Visual results of different methods on a low-light image sampled from LLIV-Phone-imgT dataset. (a) input. (b) LLNet [1]. (c) LightenNet [5]. (d) Retinex-Net [4]. (e) MBLLEN [3]. (f) KinD [11]. (g) KinD++ [61]. (h) TBEFN [20]. (i) DSLR [21]. (j) EnlightenGAN [26]. (k) DRBN [33]. (l) ExCNet [27]. (m) Zero-DCE [28]. (n) RRDNet [29].

with a total of 600 images (denoted as LLIV-Phone-imgT). Furthermore, we randomly select one video from the videos of each phone's brand of LLIV-Phone dataset, forming a video testing dataset with a total of 18 videos (denoted as LLIV-Phone-vidT). We half the resolutions of the frames in both LLIV-Phone-imgT and LLIV-Phone-vidT because some deep learning-based methods cannot process the full resolution of test images and videos. For the LOL dataset, we adopt the original test set including 15 low-light images captured in real scenes for testing, denoted as LOL-test. For the MIT-Adobe FiveK dataset, we follow the protocol in Chen et al. [47] to decode the images into PNG format and resize them to have a long edge of 512 pixels using Lightroom. We adopt the same testing dataset as Chen et al. [47], MIT-Adobe FiveK-test, including 500 images with the retouching results by expert C as the corresponding ground truths. For the SID dataset, we use the default test set used in EEMEFN [16] for fair comparisons, denoted as SID-test (SID-test-Bayer and SID-test-X-Trans), which is a partial test set of SID [85]. The SID-test-Bayer includes 93 images of the Bayer pattern while the SID-test-X-Trans includes 94 images of the APS-C X-Trans pattern.

Qualitative Comparison. We first present the results of different methods on the images sampled from LOL-test and

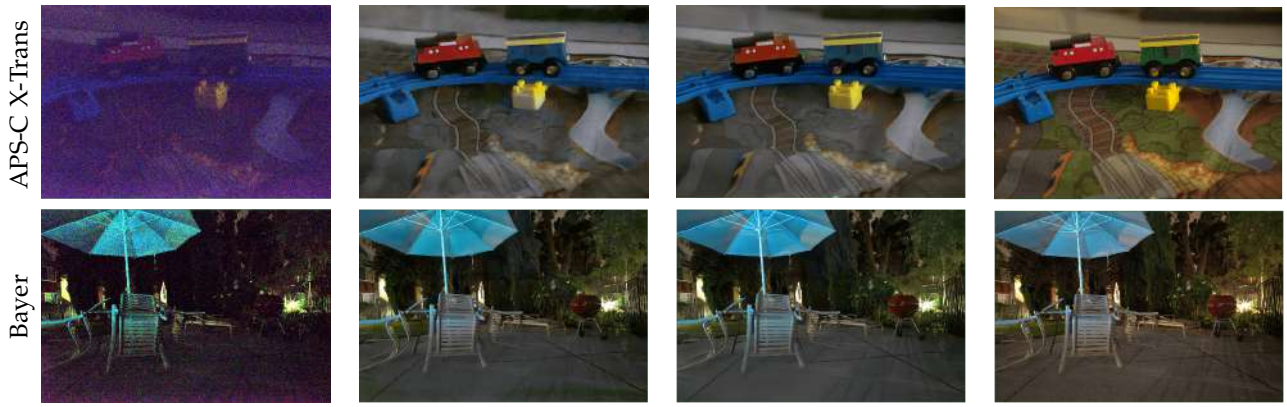


Fig. 9. Visual results of different methods on two raw low-light images sampled from SID-test-Bayer and SID-test-X-Trans test datasets. The inputs are amplified for visualization.

TABLE 5
Quantitative comparisons on LOL-test and MIT-Adobe FiveK-test test datasets in terms of MSE ($\times 10^3$), PSNR (in dB), SSIM [86], and LPIPS [87]. The best result is in red whereas the second and third best results are in blue and purple under each case, respectively.

Learning	Method	LOL-test				MIT-Adobe FiveK-test			
		MSE↓	PSNR ↑	SSIM↑	LPIPS↓	MSE↓	PSNR ↑	SSIM↑	LPIPS↓
	input	12.613	7.773	0.181	0.560	1.670	17.824	0.779	0.148
SL	LLNet [1]	1.290	17.959	0.713	0.360	4.465	12.177	0.645	0.292
	LightenNet [5]	7.614	10.301	0.402	0.394	4.127	13.579	0.744	0.166
	Retinex-Net [4]	1.651	16.774	0.462	0.474	4.406	12.310	0.671	0.239
	MBLLEN [3]	1.444	17.902	0.715	0.247	1.296	19.781	0.825	0.108
	KinD [11]	1.431	17.648	0.779	0.175	2.675	14.535	0.741	0.177
	KinD++ [61]	1.298	17.752	0.760	0.198	7.582	9.732	0.568	0.336
	TBEFN [20]	1.764	17.351	0.786	0.210	3.865	12.769	0.704	0.178
	DSLR [21]	3.536	15.050	0.597	0.337	1.925	16.632	0.782	0.167
UL	EnlightenGAN [26]	1.998	17.483	0.677	0.322	3.628	13.260	0.745	0.170
SSL	DRBN [33]	2.359	15.125	0.472	0.316	3.314	13.355	0.378	0.281
ZSL	ExCNet [27]	2.292	15.783	0.515	0.373	2.927	13.978	0.710	0.187
	Zero-DCE [28]	3.282	14.861	0.589	0.335	3.476	13.199	0.709	0.203
	RRDNet [29]	6.313	11.392	0.468	0.361	7.057	10.135	0.620	0.303

TABLE 6
Quantitative comparisons on SID-test test dataset in terms of MSE ($\times 10^3$), PSNR (in dB), SSIM [86], and LPIPS [87]. The best result is in red under each case. To compute the quantitative scores of input raw data, we use the corresponding camera ISP pipelines provided by Chen et al. [85] to transfer raw data to RGB format.

Learning	Method	SID-test–Bayer				SID-test–X-Trans			
		MSE↓	PSNR ↑	SSIM↑	LPIPS↓	MSE↓	PSNR ↑	SSIM↑	LPIPS↓
	input	5.378	11.840	0.063	0.711	4.803	11.880	0.075	0.796
SL	SID [85]	0.140	28.614	0.757	0.465	0.235	26.663	0.680	0.586
	EEMEFN [16]	0.126	29.212	0.768	0.448	0.191	27.423	0.695	0.546

MIT-Adobe FiveK-test datasets in Figures 5 and 6.

As shown in Figure 5, all methods improve the brightness and contrast of the input image. However, none of them successfully recovers the accurate color of the input image when the results are compared with the ground truth. In particular, LLNet [1] produces blurring result. LightenNet [5] and RRDNet [29] produce under-exposed results while MBLLEN [3] and ExCNet [27] over-expose the image. KinD [11], KinD++ [61], TBEFN [20], DSLR [21], EnlightenGAN [26], and DRBN [33] introduce obvious artifacts. In Figure 6, LLNet [5], KinD++ [61], TBEFN [20], and RRDNet [29] produce over-exposed results. Retinex-Net [4], KinD++ [61], and RRDNet [29] yield artifacts and blurring in the results.

We found that the ground truths of MIT-Adobe FiveK dataset still contain some dark regions. This is because the dataset is originally designed for global image retouching, where restoring low light regions is not the main priority in this task. We also observed that the input images in LOL dataset and MIT-Adobe FiveK dataset are relatively clean from noise, which is different from real low-light scenes. Although some methods [18], [21], [60] take the MIT-Adobe FiveK dataset as the training or testing dataset, we argue that this dataset is not appropriate for the task of LLIE due to its mismatched/unsatisfactory ground truth for LLIE.

To examine the generalization capability of different methods, we conduct comparisons on the images sampled from our LLIV-Phone-imgT dataset. The visual results of different methods are shown in Figures 7 and 8. As presented in Figure 7, all methods cannot effectively improve the brightness and remove the noise of the input low-light

image. Moreover, Retinex-Net [4], MBLLEN [3], and DRBN [33] produce obvious artifacts. In Figure 8, all methods enhance the brightness of this input image. However, only MBLLEN [3] and RRDNet [29] obtain visually pleasing enhancement without color deviation, artifacts, and over-/under-exposure. Notably, for regions with a light source, none of the methods can brighten the image without amplifying the noise around these regions. Taking light sources into account for LLIE would be an interesting direction to explore. The results suggest the difficulty of enhancing the images of the LLIV-Phone-imgT dataset. Real low-light images fail most existing LLIE methods due to the limited generalization capability of these methods. The potential reasons are the use of synthetic training data, small-scaled training data, or unrealistic assumptions such as the local illumination consistency and treating the reflectance component as the final result in the Retinex model in these methods.

We further present the visual comparisons of raw format-based methods in Figure 9. As shown, the input raw data have obvious noises. Both SID [2] and EEMEFN [16] can effectively remove the effects of noises. In comparison to the simple U-Net structure used in SID [2], the more complex structure of EEMEFN [16] obtains better brightness recovery. However, their results are far from the corresponding GT, especially for the input of APS-C X-Trans pattern.

Quantitative Comparison. For test sets with ground truth i.e., LOL-test, MIT-Adobe FiveK-test, and SID-test, we adopt MSE, PSNR, SSIM [86], and LPIPS [87] metrics to quantitatively compare different methods. LPIPS [87] is a deep learning-based image quality assessment metric that measures the perceptual similarity between a result and its corresponding ground truth by deep visual representations. For LPIPS, we employ the AlexNet-based model to compute the perceptual similarity. A lower LPIPS value suggests a result that is closer to the corresponding ground truth in terms of perceptual similarity. In Table 5 and Table 6, we show the quantitative results of RGB format-based methods and raw format-based methods, respectively.

As presented in Table 5, the quantitative scores of supervised learning-based methods are better than those of unsupervised learning-based, semi-supervised learning-based, and zero-shot learning-based methods on LOL-test and MIT-Adobe FiveK-test datasets. Among them, LLNet [1] obtains the best MSE and PSNR values on the LOL-test dataset; however, its performance drops on the MIT-Adobe FiveK-test dataset. This may be caused by the bias of LLNet [1] towards the LOL dataset since it was trained using the LOL training dataset. For the LOL-test dataset, TBEFN [20] obtains the highest SSIM value while KinD [11] achieves the lowest LPIPS value. There is no winner across these four evaluation metrics on the LOL-test dataset despite the fact that some methods were trained on the LOL training dataset. For the MIT-Adobe FiveK-test dataset, MBLLEN [3] outperforms all compared methods under the four evaluation metrics in spite of being trained on synthetic training data. Nevertheless, MBLLEN [3] still cannot obtain the best performance on both two test datasets.

As presented in Table 6, both SID [85] and EEMEFN [16] improve the quality of input raw data. Compared with the quantitative scores of SID [85], EEMEFN [16] achieves consistently better performance across different raw data patterns and evaluation metrics.

For LLIV-Phone-imgT test set, we use the non-reference IQA metrics, i.e., NIQE [88], perceptual index (PI) [88], [89], [90], LOE [37], and SPAQ [91] to quantitatively compare different methods. In terms of LOE, the smaller the LOE value is, the better the lightness order is preserved. For NIQE, the smaller the NIQE value is, the better the visual quality is. A lower PI value indicates better perceptual quality. SPAQ is devised for the perceptual quality assessment of smartphone photography. A larger SPAQ value suggests better perceptual quality of smartphone photography. The quantitative results are provided in Table 7.

TABLE 7
Quantitative comparisons on LLIV-Phone-imgT dataset in terms of NIQE [88], LOE [37], PI [88], [89], [90], and SPAQ [91]. The best result is in red whereas the second and third best results are in blue and purple under each case, respectively.

Learning	Method	LoLi-Phone-imgT			
		NIQE↓	LOE ↓	PI↓	SPAQ↑
	input	6.99	0.00	5.86	44.45
SL	LLNet [1]	5.86	5.86	5.66	40.56
	LightenNet [5]	5.34	952.33	4.58	45.74
	Retinex-Net [4]	5.01	790.21	3.48	50.95
	MBLLEN [3]	5.08	220.63	4.27	42.50
	KinD [11]	4.97	405.88	4.37	44.79
	KinD++ [61]	4.73	681.97	3.99	46.89
	TBEFN [20]	4.81	552.91	4.30	44.14
	DSLR [21]	4.77	447.98	4.31	41.08
UL	EnlightenGAN [26]	4.79	821.87	4.19	45.48
SSL	DRBN [33]	5.80	885.75	5.54	42.74
ZSL	ExCNet [27]	5.55	723.56	4.38	46.74
	Zero-DCE [28]	5.82	307.09	4.76	46.85
	RRDNet [29]	5.97	142.89	4.84	45.31

Observing Table 7, we can find that the performance of Retinex-Net [4], KinD++ [61], and EnlightenGAN [26] is relatively better than the other methods. Retinex-Net [4] achieves the best PI and SPAQ scores. The scores suggest the good perceptual quality of the results enhanced by Retinex-Net [4]. However, from Figure 7(d) and Figure 8(d), the results of Retinex-Net [4] evidently suffer from artifacts and color deviations. Moreover, KinD++ [61] attains the lowest NIQE score while the original input achieves the lowest LOE score. For the de-facto standard LOE metric, we question if the lightness order can effectively reflect the enhancement performance. Overall, the non-reference IQA metrics experience biases on the evaluations of the quality of enhanced low-light images in some cases.

To prepare videos in the LLIV-vidT testing set, we first discard videos without obvious objects in consecutive frames. A total of 10 videos are chosen. For each video, we select one object that appears in all frames. We then use a tracker [92] to track the object in consecutive frames of the input video and ensure the same object appears in the bounding boxes. We discard the frames with inaccurate object tracking. The coordinates of the bounding box in each frame are collected. We employ these coordinates to crop the corresponding regions in the results enhanced by different methods and compute the average luminance variance (ALV) scores of the object in the consecutive frames as: $ALV = \frac{1}{N}\sum_{i=1}^{N}(L_i - L_{\text{avg}})^2$, where N is the number of frames

TABLE 8
Quantitative comparisons on LLIV-Phone-vidT dataset in terms of average luminance variance (ALV) score. The best result is in red whereas the second and third best results are in blue and purple.

Learning	Method	LoLi-Phone-vidT
		ALV↓
	input	185.60
SL	LLNet [1]	85.72
	LightenNet [5]	643.93
	Retinex-Net [4]	94.05
	MBLLEN [3]	113.18
	KinD [11]	98.05
	KinD++ [61]	115.21
	TBEFN [20]	58.69
	DSLR [21]	175.35
UL	EnlightenGAN [26]	90.69
SSL	DRBN [33]	115.04
ZSL	ExCNet [27]	1375.29
	Zero-DCE [28]	117.22
	RRDNet [29]	147.11

of a video, L_i represents the average luminance value of the region of bounding box in the ith frame, and L_{avg} denotes the average luminance value of all bounding box regions in the video. A lower ALV value suggests better temporal coherence of the enhanced video. The ALV values of different methods averaged over the 10 videos of the LLIV-vidT testing set are shown in Table 8. The ALV values of different methods on each video can be found in the supplementary material. Besides, we follow Jiang and Zheng [9] to plot their luminance curves in the supplementary material.

As shown in Table 8, TBEFN [20] obtains the best temporal coherence in terms of ALV value whereas LLNet [1] and EnlightenGAN [26] rank the second and third best, respectively. In contrast, the ALV value of ExCNet [27], as the worst performer, reaches 1375.29. This is because the performance of the zero-reference learning-based ExCNet [27] is unstable for the enhancement of consecutive frames. ExCNet [27] can effectively improve the brightness of some frames while it does not work well on other frames.

4.4 Computational Complexity

In Table 9, we compare the computational complexity of RGB format-based methods, including runtime, trainable parameters, and FLOPs averaged over 32 images of size 1200×900×3 using an NVIDIA 1080Ti GPU. We omit LightenNet [5] for fair comparisons because only the CPU version of its code is publicly available. Besides, we do not report the FLOPs of ExCNet [27] and RRDNet [29] as the number depends on the input images (different inputs require different numbers of iterations).

As presented in Table 9, Zero-DCE [28] has the shortest runtime because it only estimates several curve parameters via a lightweight network. As a result, its number of trainable parameters and FLOPs are much fewer. Moreover, the number of trainable parameters and FLOPs of LightenNet [5] are the least among the compared methods. This is because LightenNet [5] estimates the illumination map of input image via a tiny network of four convolutional layers. In contrast, the FLOPs of LLNet [1] and KinD++ [61] are extremely large, reaching 4124.177G and 12238.026G, respectively. The runtime of SSL-based ExCNet [27] and RRDNet [29] is long due to the time-consuming optimization process.

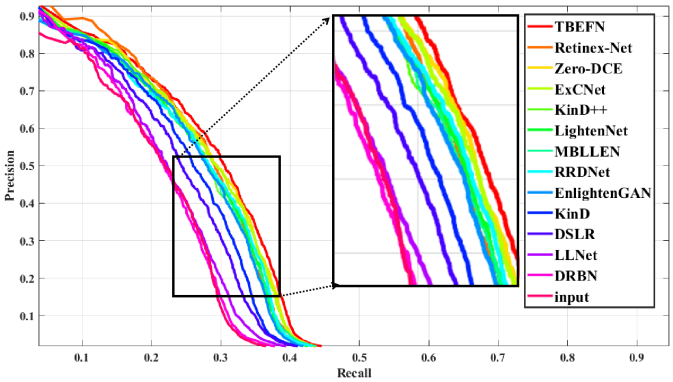


Fig. 10. The P-R curves of face detection in the dark.

4.5 Application-Based Evaluation

We investigate the performance of low-light image enhancement methods on face detection in the dark. Following the setting presented in Guo et al. [28], we use the DARK FACE dataset [80] that is composed of images with faces taken in the dark. Since the bounding boxes of the test set are not publicly available, we perform the evaluation on 500 images randomly sampled from the training and validation sets. The Dual Shot Face Detector (DSFD) [93] trained on WIDER FACE dataset [94] is used as the face detector. We feed the results of different LLIE methods to the DSFD [93] and depict the precision-recall (P-R) curves under 0.5 IoU threshold in Figure 10. We compare the average precision (AP) under different IoU thresholds using the evaluation tool[3] provided in DARK FACE dataset [80] in Table 10.

As shown in Figure 10, all the deep learning-based solutions improve the performance of face detection in the dark, suggesting the effectiveness of deep learning-based LLIE solutions for face detection in the dark. As shown in Table 10, the AP scores of best performers under different IoU thresholds range from 0.268 to 0.013 and the AP scores of input under different IoU thresholds are very low. The results suggest that there is still room for improvement. It is noteworthy that Retinex-Net [4], Zero-DCE [28], and TBEFN [20] achieve relatively robust performance on face detection in the dark. We show the visual results of different methods in Figure 11. Although Retinex-Net [4] performs better than other methods on the AP score, its visual result contains obvious artifacts and unnatural textures. In general, Zero-DCE [28] obtains a good balance between the AP score and the perceptual quality for face detection in the dark. Note that the results of face detection in the dark are related to not only the enhanced results but also the face detector including the detector model and the training data of the detector. Here, we only take the pre-trained DSFD [93] as an example to validate the low-light image enhancement performance of different methods to some extent.

4.6 Discussion

From the experimental results, we obtain several interesting observations and insights:

1) The performance of different methods significantly varies based on the test datasets and evaluation metrics.

3. https://github.com/Ir1d/DARKFACE_eval_tools

TABLE 9
Quantitative comparisons of computational complexity in terms of runtime (in second), number of trainable parameters (#Parameters) (in M), and FLOPs (in G). The best result is in red whereas the second and third best results are in blue and purple under each case, respectively. '-' indicates the result is not available.

Learning	Method	RunTime↓	#Parameters ↓	FLOPs↓	Platform
SL	LLNet [1]	36.270	17.908	4124.177	Theano
	LightenNet [5]	-	0.030	30.540	MATLAB
	Retinex-Net [4]	0.120	0.555	587.470	TensorFlow
	MBLLEN [3]	13.995	0.450	301.120	TensorFlow
	KinD [11]	0.148	8.160	574.954	TensorFlow
	KinD++ [61]	1.068	8.275	12238.026	TensorFlow
	TBEFN [20]	0.050	0.486	108.532	TensorFlow
	DSLR [21]	0.074	14.931	96.683	PyTorch
UL	EnlightenGAN [26]	0.008	8.637	273.240	PyTorch
SSL	DRBN [33]	0.878	0.577	196.359	PyTorch
ZSL	ExCNet [27]	23.280	8.274	-	PyTorch
	Zero-DCE [28]	0.003	0.079	84.990	PyTorch
	RRDNet [29]	167.260	0.128	-	PyTorch

(a) input (b) LightenNet [5] (c) Retinex-Net [4] (d) MBLLEN [3]
(e) KinD++ [61] (f) TBEFN [20] (g) DSLR [21] (h) EnlightenGAN [26]
(i) DRBN [33] (j) ExCNet [27] (k) Zero-DCE [28] (l) RRDNet [29]

Fig. 11. Visual results of different methods on a low-light image sampled from DARK FACE dataset. Better see with zoom in for the bounding boxes of faces.

In terms of the full-reference IQA metrics on commonly used test datasets, MBLLEN [3], KinD++ [61], and DSLR [21] are generally better than other compared methods. For real-world low-light images taken by mobile phones, supervised learning-based Retinex-Net [4] and KinD++ [61] obtain better scores measured in the non-reference IQA metrics. For real-world low-light videos taken by mobile phones, TBEFN [20] preserves the temporal coherence better. When coming to the computational efficiency, LightenNet [5] and Zero-DCE [28] are outstanding. From the aspect of face detection in the dark, TBEFN [20], Retinex-Net [4], and Zero-DCE [28] rank the first three. No method always wins. Overall, Retinex-Net [4], [20], Zero-DCE [28], and DSLR [21] are better choice in most cases. For raw data, EEMEFN [16] obtains relatively better qualitative and quantitative performance than SID [85]. However, from the visual results, EEMEFN [16] and [85] cannot recover the color well when compared with the corresponding ground truth.

2) LLIV-Phone dataset fails most methods. The generalization capability of existing methods needs further improvements. It is worth noting that it is inadequate to use only the average luminance variance to evaluate the performance of different methods for low-light video enhancement. More effective and comprehensive assessment metrics would guide the development of low-light video enhancement towards the right track.

3) Regarding learning strategies, supervised learning achieves better performance in most cases, but requires high computational resources and paired training data. In comparison, zero-shot learning is more appealing in practical

TABLE 10
Quantitative comparisons of AP under different IoU thresholds of face detection in the dark. The best result is in red whereas the second and third best results are in blue and purple under each case, respectively.

Learning	Method	IoU thresholds		
		0.5	0.6	0.7
	input	0.195	0.061	0.007
SL	LLNet [1]	0.208	0.063	0.006
	LightenNet [5]	0.249	0.085	0.010
	Retinex-Net [4]	0.261	0.101	0.013
	MBLLEN [3]	0.249	0.092	0.010
	KinD [11]	0.235	0.081	0.010
	KinD++ [61]	0.251	0.090	0.011
	TBEFN [20]	0.268	0.099	0.011
	DSLR [21]	0.223	0.067	0.007
UL	EnlightenGAN [26]	0.246	0.088	0.011
SSL	DRBN [33]	0.199	0.061	0.007
ZSL	ExCNet [27]	0.256	0.092	0.010
	Zero-DCE [28]	0.259	0.092	0.011
	RRDNet [29]	0.248	0.083	0.010

applications because it does not require paired or unpaired training data. Consequently, zero-shot learning-based methods enjoy better generalization capability. However, their quantitative performance is inferior to other methods.

4) There is a gap between visual results and quantitative IQA scores. In other words, a good visual appearance does not always yield a good IQA score. The relationships between human perception and IQA scores are worth more investigation. Pursuing better visual perception or quantitative scores depends on specific applications. For instance, to show the results to observers, more attention should be paid to visual perception. In contrast, accuracy is more important when LLIE methods are applied to face detection in the dark. Thus, more comprehensive comparisons should be performed when comparing different methods.

5) Deep learning-based LLIE methods are beneficial to face detection in the dark. Such results further support the significance of enhancing low-light images and videos. However, in comparison to the high accuracy of face detection in normal-light images, the accuracy of face detection in the dark is extremely low, despite using LLIE methods.

6) In comparison to RGB format-based LLIE methods, raw format-based LLIE methods usually recover details better, obtain more vivid color, and reduce noises and artifacts more effectively. This is because raw data contain more information such as wider color gamut and higher dynamic range. However, raw format-based LLIE methods are limited to specific sensors and formats such as the Bayer pattern of the Sony camera and the APS-C X-Trans pattern of the Fuji camera. In contrast, RGB format-based LLIE methods are more convenient and versatile since RGB images are commonly found as the final imagery form produced by mobile devices. However, RGB format-based LLIE methods cannot cope well with cases that exhibit low light and excessive noise.

5 Open Issues

In this section, we summarize the open issues in low-light image and video enhancement as follows.

Generalization Capability. Although existing methods can produce some visually pleasing results, they have limited generalization capability. For example, a method trained on MIT-Adobe FiveK dataset [79] cannot effectively enhance the low-light images of LOL dataset [4]. Albeit synthetic data are used to augment the diversity of training data, the models trained on the combination of real and synthetic data cannot solve this issue well. Improving the generalization capability of LLIE methods is an unsolved open issue.

Removing Unknown Noises. Observing the results of existing methods on the low-light images captured by different types of phones' cameras, we can find that these methods cannot remove the noises well and even amplify the noises, especially when the types of noises are unknown. Despite some methods add Gaussian and/or Poisson noises in their training data, the noise types are different from real noises, thus the performance of these methods is unsatisfactory in real scenarios. Removing unknown noises is still unsolved.

Removing Unknown Artifacts. One may enhance a low-light image downloaded from the Internet. The image may have gone through a serial of degradations such as JPEG compression or editing. Thus, the image may contain unknown artifacts. Suppressing unknown artifacts still challenges existing low-light image and video enhancement methods.

Correcting Uneven Illumination. Images taken in real scenes usually exhibit uneven illumination. For example, an image captured at night has both dark regions and normal-light or over-exposed regions such as the regions of light sources. Existing methods tend to brighten both the dark regions and the light source regions, affecting the visual quality of the enhanced result. It is expected to enhance dark regions but suppress over-exposed regions. However, this open issue is not solved well in existing LLIE methods.

Distinguishing Semantic Regions. Existing methods tend to enhance a low-light image without considering the semantic information of its different regions. For example, the black hair of a man in a low-light image is enhanced to be off-white as the black hair is treated as the low-light regions. An ideal enhancement method is expected to only enhance the low-light regions induced by external environments. How to distinguish semantic regions is an open issue.

Using Neighbouring Frames. Despite some methods that have been proposed to enhance low-light videos, they commonly process a video frame-by-frame. How to make full use of the neighboring frames to improve the enhancement performance and speed up the processing speed is an unsolved open issue. For example, the well-lit regions of neighboring frames are used to enhance the current frame. For another example, the estimated parameters for processing neighboring frames can be reused to enhance the current frame for reducing the time of parameter estimation.

6 Future Research Directions

Low-light enhancement is a challenging research topic. As can be observed from the experiments presented in Section 4 and the unsolved open issues in Section 5, there is still room for improvement. We suggest potential future research directions as follows.

Effective Learning Strategies. As aforementioned, current LLIE models mainly adopt supervised learning that requires massive paired training data and may overfit on a specific

dataset. Although some researchers attempted to introduce unsupervised learning into LLIE, the inherent relationships between LLIE and these learning strategies are not clear and their effectiveness in LLIE needs further improvements. Zero-shot learning has shown robust performance for real scenes while not requiring paired training data. The unique advantage suggests zero-shot learning as a potential research direction, especially on the formulation of zero-reference losses, deep priors, and optimization strategies.

Specialized Network Structures. A network structure can significantly affect enhancement performance. As previously analyzed, most LLIE deep models employ U-Net or U-Net-like structures. Though they have achieved promising performance in some cases, the investigation if such an encoder-decoder network structure is most suitable for the LLIE task is still lacking. Some network structures require a high memory footprint and long inference time due to their large parameter space. Such network structures are unacceptable for practical applications. Thus, it is worthwhile to investigate a more effective network structure for LLIE, considering the characteristics of low-light images such as non-uniform illumination, small pixel values, noise suppression, and color constancy. One can also design more efficient network structures via taking into account the local similarity of low-light images or considering more efficient operations such as depthwise separable convolution layer [95] and self-calibrated convolution [96]. Neural architecture search (NAS) technique [97], [98] may be considered to obtain more effective and efficient LLIE network structures. Adapting the transformer architecture [99], [100] into LLIE may be a potential and interesting research direction.

Loss Functions. Loss functions constrain the relationships between an input image and ground truth and drive the optimization of deep networks. In LLIE, the commonly used loss functions are borrowed from related vision tasks. Thus, designing loss functions that are more well-suited for LLIE is desired. Recent studies have shown the possibility of using deep neural networks to approximate human visual perception of image quality [101], [102]. These ideas and fundamental theories could be used to guide the designs of loss functions for low-light enhancement networks.

Realistic Training Data. Although there are several training datasets for LLIE, their authenticity, scales, and diversities fall behind real low-light conditions. Thus, as shown in Section 4, current LLIE deep models cannot achieve satisfactory performance when encountering low-light images captured in real-world scenes. More efforts are needed to study the collection of large-scale and diverse real-world paired LLIE training datasets or to generate more realistic synthetic data.

Standard Test Data. Currently, there is no well-accepted LLIE evaluation benchmark. Researchers prefer selecting their test data that may bias to their proposed methods. Despite some researchers leave some paired data as test data, the division of training and test partitions are mostly ad-hoc across the literature. Consequently, conducting a fair comparison among different methods is often laborious if not impossible. Besides, some test data are either easy to be handled or not originally collected for low-light enhancement. It is desired to have a standard low-light image and video test dataset, which includes a large number of test samples with the corresponding ground truths, covering diverse scenes and challenging illumination conditions.

Task-Specific Evaluation Metrics. The commonly adopted evaluation metrics in LLIE can reflect the image quality to some extent. However, how to measure how good a result is enhanced by an LLIE method still challenges current IQA metrics, especially for non-reference measurements. The current IQA metrics either focus on human visual perceptual such as subjective quality or emphasize machine perceptual such as the effects on high-level visual tasks. Therefore, more works are expected in this research direction to make efforts on designing more accurate and task-specific evaluation metrics for LLIE.

Robust Generalization Capability. Observing the experimental results on real-world test data, most methods fail due to their limited generalization capability. The poor generalization is caused by several factors such as synthetic training data, small-scaled training data, ineffective network structures, or unrealistic assumptions. It is important to explore ways to improve the generalization.

Extension to Low-Light Video Enhancement. Unlike the rapid development of video enhancement in other low-level vision tasks such as video deblurring [103], video denoising [104], and video super-resolution [105], low-light video enhancement receives less attention. A direct application of existing LLIE methods to videos often leads to unsatisfactory results and flickering artifacts. More efforts are needed to remove visual flickering effectively, exploit the temporal information between neighboring frames, and speed up the enhancement speed.

Integrating Semantic Information. Semantic information is crucial for low-light enhancement. It guides the networks to distinguish different regions in the process of enhancement. A network without access to semantic priors can easily deviate the original color of a region, e.g., turning black hair to gray color after enhancement. Therefore, integrating semantic priors into LLIE models is a promising research direction. Similar work has been done on image super-resolution [106], [107] and face restoration [108].

ACKNOWLEDGMENTS

This study is supported under the RIE2020 Industry Alignment Fund Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). It is also partially supported by the NTU SUG and NAP grant. Chunle Guo is sponsored by CAAI-Huawei MindSpore Open Fund.

REFERENCES

[1] K. G. Lore, A. Akintayo, and S. Sarkar, "LLNet: A deep autoencoder approach to natural low-light image enhancement," *PR*, vol. 61, pp. 650–662, 2017.

[2] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *CVPR*, 2018, pp. 3291–3300.

[3] F. Lv, F. Lu, J. Wu, and C. Lim, "MBLLEN: Low-light image/video enhancement using cnns," in *BMVC*, 2018.

[4] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," in *BMVC*, 2018.

[5] C. Li, J. Guo, F. Porikli, and Y. Pang, "LightenNet: A convolutional neural network for weakly illuminated image enhancement," *PRL*, vol. 104, pp. 15–22, 2018.

[6] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *TIP*, vol. 27, no. 4, pp. 2049–2062, 2018.

[7] R. Wang, Q. Zhang, C.-W. Fu, X. Shen, W.-S. Zheng, and J. Jia, "Underexposed photo enhancement usign deep illumination estimation," in *CVPR*, 2019, pp. 6849–6857.

[8] C. Chen, Q. Chen, M. N. Do, and V. Koltun, "Seeing motion in the dark," in *ICCV*, 2019, pp. 3185–3194.

[9] H. Jiang and Y. Zheng, "Learning to see moving object in the dark," in *ICCV*, 2019, pp. 7324–7333.

[10] Y. Wang, Y. Cao, Z. Zha, J. Zhang, Z. Xiong, W. Zhang, and F. Wu, "Progressive retinex: Mutually reinforced illumination-noise perception network for low-light image enhancement," in *ACMMM*, 2019, pp. 2015–2023.

[11] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *ACMMM*, 2019, pp. 1632–1640.

[12] W. Ren, S. Liu, L. Ma, Q. Xu, X. Xu, X. Cao, J. Du, and M.-H. Yang, "Low-light image enhancement via a deep hybrid network," *TIP*, vol. 28, no. 9, pp. 4364–4375, 2019.

[13] K. Xu, X. Yang, B. Yin, and R. W. H. Lau, "Learning to restore low-light images via decomposition-and-enhancement," in *CVPR*, 2020, pp. 2281–2290.

[14] M. Fan, W. Wang, W. Yang, and J. Liu, "Integrating semantic segmentation and retinex model for low light image enhancement," in *ACMMM*, 2020, pp. 2317–2325.

[15] F. Lv, B. Liu, and F. Lu, "Fast enhancement for non-uniform illumination images using light-weight cnns," in *ACMMM*, 2020, pp. 1450–1458.

[16] M. Zhu, P. Pan, W. Chen, and Y. Yang, "EEMEFN: Low-light image enhancement via edge-enhanced multi-exposure fusion network," in *AAAI*, 2020, pp. 13 106–13 113.

[17] 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 *ECCV*, 2020, pp. 103–119.

[18] J. Li, J. Li, F. Fang, F. Li, and G. Zhang, "Luminance-aware pyramid network for low-light image enhancement," *TMM*, 2020.

[19] L. Wang, Z. Liu, W. Siu, and D. P. K. Lun, "Lightening network for low-light image enhancement," *TIP*, vol. 29, pp. 7984–7996, 2020.

[20] K. Lu and L. Zhang, "TBEFN: A two-branch exposure-fusion network for low-light image enhancement," *TMM*, 2020.

[21] S. Lim and W. Kim, "DSLR: Deep stacked laplacian restorer for low-light image enhancement," *TMM*, 2020.

[22] F. Zhang, Y. Li, S. You, and Y. Fu, "Learning temporal consistency for low light video enhancement from single images," in *CVPR*, 2021.

[23] J. Li, X. Feng, and Z. Hua, "Low-light image enhancement via progressive-recursive network," *TCSVT*, 2021.

[24] W. Yang, W. Wang, H. Huang, S. Wang, and J. Liu, "Sparse gradient regularized deep retinex network for robust low-light image enhancement," *TIP*, vol. 30, pp. 2072–2086, 2021.

[25] R. Yu, W. Liu, Y. Zhang, Z. Qu, D. Zhao, and B. Zhang, "DeepExposure: Learning to expose photos with asynchronously reinforced adversarial learning," in *NeurIPS*, 2018, pp. 2149–2159.

[26] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "EnlightenGAN: Deep light enhancement without paired supervision," *TIP*, vol. 30, pp. 2340–2349, 2021.

[27] L. Zhang, L. Zhang, X. Liu, Y. Shen, S. Zhang, and S. Zhao, "Zero-shot restoration of back-lit images using deep internal learning," in *ACMMM*, 2019, pp. 1623–1631.

[28] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *CVPR*, 2020, pp. 1780–1789.

[29] A. Zhu, L. Zhang, Y. Shen, Y. Ma, S. Zhao, and Y. Zhou, "Zero-shot restoration of underexposed images via robust retinex decomposition," in *ICME*, 2020, pp. 1–6.

[30] C. Li, C. Guo, and C. C. Loy, "Learning to enhance low-light image via zero-reference deep curve estimation," *TPAMI*, 2021.

[31] Z. Zhao, B. Xiong, L. Wang, Q. Ou, L. Yu, and F. Kuang, "Retinexdip: A unified deep framework for low-light image enhancement," *TCSVT*, 2021.

[32] 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 *CVPR*, 2021.

[33] 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 *CVPR*, 2020, pp. 3063–3072.

[34] W. Yang, S. Wang, Y. F. nd Yue Wang, and J. Liu, "Band representation-based semi-supervised low-light image enhancement: Bridging the gap between signal fidelity and perceptual quality," *TIP*, vol. 30, pp. 3461–3473, 2021.

[35] H. Ibrahim and N. S. P. Kong, "Brightness preserving dynamic histogram equalization for image contrast enhancement," *TCE*, vol. 53, no. 4, pp. 1752–1758, 2007.

[36] M. Abdullah-AI-Wadud, M. H. Kabir, M. A. A. Dewan, and O. Chae, "A dynamic histogram equalization for image contrast enhancement," *TCE*, vol. 53, no. 2, pp. 593–600, 2007.

[37] S. Wang, J. Zheng, H. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *TIP*, vol. 22, no. 9, pp. 3538–3548, 2013.

[38] X. Fu, Y. Liao, D. Zeng, Y. Huang, X. Zhang, and X. Ding, "A probabilistic method for image enhancement with simultaneous illumination and reflectance estimation," *TIP*, vol. 24, no. 12, pp. 4965–4977, 2015.

[39] X. Guo, Y. Li, and H. Ling, "LIME: Low-light image enhancement via illumination map estimation," *TIP*, vol. 26, no. 2, pp. 982–993, 2016.

[40] S. Park, S. Yu, B. Moon, S. Ko, and J. Paik, "Low-light image enhancement using variational optimization-based retinex model," *TCE*, vol. 63, no. 2, pp. 178–184, 2017.

[41] M. Li, J. Liu, W. Yang, X. Sun, and Z. Guo, "Structure-revealing low-light image enhancement via robust retinex model," *TIP*, vol. 27, no. 6, pp. 2828–2841, 2018.

[42] Z. Gu, F. Li, F. Fang, and G. Zhang, "A novel retinex-based fractional-order variational model for images with severely low light," *TIP*, vol. 29, pp. 3239–3253, 2019.

[43] X. Ren, W. Yang, W.-H. Cheng, and J. Liu, "LR3M: Robust low-light enhancement via low-rank regularized retinex model," *TIP*, vol. 29, pp. 5862–5876, 2020.

[44] S. Hao, X. Han, Y. Guo, X. Xu, and M. Wang, "Low-light image enhancement with semi-decoupled decomposition," *TMM*, vol. 22, no. 12, pp. 3025–3038, 2020.

[45] Y. Hu, H. He, C. Xu, B. Wang, and S. Lin, "Exposure: A white-box photo post-processing framework," *ACM Graph.*, vol. 37, no. 2, pp. 1–17, 2018.

[46] M. Gharbi, J. Chen, J. T. Barron, S. W. Hasinoff, and F. Durand, "Deep bilateral learning for real-time image enhancement," *ACM Graph.*, vol. 36, no. 4, pp. 1–12, 2017.

[47] Y. Chen, Y. Wang, M. Kao, and Y. Chuang, "Deep photo enhancer: Unpaired learning for image enhancement form photographs wigh gans," in *CVPR*, 2018, pp. 6306–6314.

[48] Y. Deng, C. C. Loy, and X. Tang, "Aesthetic-driven image enhancement by adversarial learning," in *ACMMM*, 2018, pp. 870–878.

[49] Z. Yan, H. Zhang, B. Wang, S. Paris, and Y. Yu, "Automatic photo adjustment using deep neural networks," *ACM Graph.*, vol. 35, no. 2, pp. 1–15, 2016.

[50] Q. Chen, J. Xu, and V. Koltun, "Fast image processing with fully-convolutional networks," in *CVPR*, 2017, pp. 2497–2506.

[51] Z. Ni, W. Yang, S. Wang, L. Ma, and S. Kwong, "Towards unsupervised deep image enhancement with generative adversarial network," *TIP*, vol. 29, pp. 9140–9151, 2020.

[52] H. Zeng, J. Cai, L. Li, Z. Cao, and L. Zhang, "Learning image-adaptive 3D lookup tables for high performance photo enhancement in real-time," *TPAMI*, 2020.

[53] C. Li, C. Guo, Q. Ai, S. Zhou, and C. C. Loy, "Flexible piecewise curves estimation for photo enhancement," *arXiv preprint arXiv:2010.13412*, 2020.

[54] W. Wang, X. Wu, X. Yuan, and Z. Gao, "An experiment-based review of low-light image enhancement methods," *IEEE Access*, vol. 8, pp. 87 884–87 917, 2020.

[55] J. Liu, D. Xu, W. Yang, M. Fan, and H. Huang, "Benchmarking low-light image enhancement and beyond," *IJCV*, 2021.

[56] V. Jain and S. Seung, "Natural image denoising with convolutional networks," in *NeurIPS*, 2008, pp. 1–8.

[57] K. Xu, X. Yang, B. Yin, and R. W. H. Lau, "Learning to restore low-light images via decomposition-and-enhancement," in *CVPR*, 2020, pp. 2281–2290.

[58] E. H. Land, "An alternative technique for the computation of the designator in the retinex theory of color vision," *National Academy of Sciences*, vol. 83, no. 10, pp. 3078–3080, 1986.

[59] D. J. Jobson, Z. ur Rahman, and G. A. Woodell, "Properties and performance of a center/surround retinex," *TIP*, vol. 6, no. 3, pp. 451–462, 1997.

[60] R. Wang, Q. Zhang, C.-W. Fu, X. Shen, W.-S. Zheng, and J. Jia, "Underexposed photo enhancement using deep illumination estimation," in *CVPR*, 2019, pp. 6849–6857.

[61] X. Guo, Y. Zhang, J. Ma, W. Liu, and J. Zhang, "Beyond brightening low-light images," *IJCV*, 2020.

[62] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[63] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *IJCV*, vol. 127, no. 8, pp. 1106–1125, 2019.

[64] K. He, J. Sun, and X. Tang, "Guided image filtering," *TPAMI*, vol. 35, no. 6, pp. 1397–1409, 2013.

[65] P. Whittle, "The psychophysics of contrast brightness," *A. L. Gilchrist (Ed.), Brightness, lightness, and transparenc (1994)*,, pp. 35–110, 1993.

[66] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *CVPR*, 2018.

[67] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *TIP*, vol. 3, no. 1, pp. 47–56, 2017.

[68] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *CVPR*, 2017, pp. 4681–4690.

[69] K. Simoayan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[70] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[71] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *CVPR*, 2017.

[72] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *ECCVW*, 2018.

[73] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *TPAMI*, vol. 38, no. 2, pp. 295–307, 2015.

[74] Q. Xu, C. Zhang, and L. Zhang, "Denoising convolutional neural network," in *ICIA*, 2015, pp. 1184–1187.

[75] X. Fu, J. Huang, D. Zeng, Y. Huang, X. Ding, and J. Paisley, "Removing rain from single images via a deep detail network," in *CVPR*, 2017, pp. 3855–3863.

[76] W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan, "Deep joint rain detection and removal from a single image," in *CVPR*, 2017, pp. 1357–1366.

[77] X. Fu, Q. Qi, Z. Zha, X. Ding, F. Wu, and J. Paisley, "Successive graph convolutional network for image deraining," *IJCV*, vol. 129, no. 5, pp. 1691–1711, 2021.

[78] J. Sun, W. Cao, Z. Xu, and J. Ponce, "Learning a convolutional neural network for non-uniform motion blur removal," in *CVPR*, 2015, pp. 769–777.

[79] V. Bychkovsky, S. Paris, E. Chan, and F. Durand, "Learning photographic global tonal adjustment with a database of input/output image pairs," in *CVPR*, 2011, pp. 97–104.

[80] Y. Yuan, W. Yang, W. Ren, J. Liu, W. JScheirer, and W. Zhangyang, "UG+ Track 2: A collective benchmark effort for evaluating and advancing image understanding in poor visibility environments," *arXiv arXiv:1904.04474*, 2019.

[81] Y. P. Loh and C. S. Chan, "Getting to know low-light images with the exclusively dark dataset," *CVIU*, vol. 178, pp. 30–42, 2019.

[82] L. Chulwoo, L. Chul, L. Young-Yoon, and K. Chang-su, "Power-constrained contrast enhancement for emissive displays based on histogram equalization," *TIP*, vol. 21, no. 1, pp. 80–93, 2012.

[83] C. Lee, C. Lee, and C.-S. Kim, "Contrast enhancement based on layered difference representation of 2d histograms," *TIP*, vol. 22, no. 12, pp. 5372–5384, 2013.

[84] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving video database with scalable annotation tooling," *arXiv preprint arXiv:1805.04687*, 2018.

[85] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *CVPR*, 2018, pp. 3291–3300.

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

[87] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595.

[88] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *SPL*, vol. 20, no. 3, pp. 209–212, 2013.

[89] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *CVPR*, 2018, pp. 6228–6237.

[90] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a non-reference quality metric for single-image super-resolution," *CVIU*, vol. 158, pp. 1–16, 2017.

[91] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *ICCV*, 2020, pp. 3677–3686.

[92] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *CVPR*, 2017, pp. 6638–6646.

[93] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, and F. Huang, "DSFD: Dual shot face detector," in *CVPR*, 2019, pp. 5060–5069.

[94] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider Face: A face detection benchmark," in *CVPR*, 2016, pp. 5525–5533.

[95] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision application," *arXiv preprint arXiv:1704.04861*, 2017.

[96] J. Liu, Q. Hou, M. M. Cheng, C. Wang, and J. Feng, "Improving convolutional networks with self-calibrated convolutions," in *CVPR*, 2020, pp. 10 096–10 105.

[97] C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L. Li, L. F. Fei, A. Yuille, J. Huang, and K. Murphy, "Progressive neural architecture search," in *ECCV*, 2018, pp. 19–34.

[98] C. Liu, L. C. Chen, F. Schroff, H. Adam, W. Hua, A. Yuille, and L. F. Fei, "Auto-Deeplab: Hierarchical neural architecture search for semantic image segmentation," in *CVPR*, 2019, pp. 82–92.

[99] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. D. M. Minderer, G. Heigold, S. Gelly, J. Uszekoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[100] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," *arXiv preprint arXiv:2012.00364*, 2020.

[101] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone phtography," in *CVPR*, 2020, pp. 3677–3686.

[102] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *TIP*, vol. 27, no. 8, pp. 3998–4011, 2018.

[103] T. H. Kim, K. M. Lee, B. Scholkopf, and M. Hirsch, "Online video deblurring via dynamic temporal blending network," in *ICCV*, 2017, pp. 4038–4047.

[104] T. Ehret, A. Davy, J.-M. Morel, G. Facciolo, and P. Arias, "Model-blind video denoising via frame-to-frame training," in *CVPR*, 2019, pp. 11 369–11 378.

[105] K. C. K. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "BasicVSR: The search for essential components in video super-resolution and beyond," in *CVPR*, 2021.

[106] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *CVPR*, 2018, pp. 606–615.

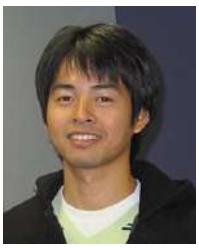
[107] K. C. K. Chan, X. Wang, X. Xu, J. Gu, and C. C. Loy, "GLEAN: Generative latent bank for large-factor image super-resolution," in *CVPR*, 2021.

[108] X. Li, C. Chen, S. Zhou, X. Lin, W. Zuo, and L. Zhang, "Blind face restoration via deep multi-scale component dictionaries," in *ECCV*, 2020, pp. 399–415.

Chongyi Li is a Research Assistant Professor with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He received the Ph.D. degree from Tianjin University, China in 2018. From 2016 to 2017, he was a joint-training Ph.D. Student with Australian National University, Australia. Prior to joining NTU, he was a postdoctoral fellow with City University of Hong Kong and Nanyang Technological University from 2018 to 2021. His current research focuses on image processing, computer vision, and deep learning, particularly in the domains of image restoration and enhancement. He serves as an associate editor of the Journal of Signal, Image and Video Processing and a lead guest editor of the IEEE Journal of Oceanic Engineering.

Chunle Guo received his PhD degree from Tianjin University in China under the supervision of Prof. Jichang Guo. He conducted the Ph.D. research as a Visiting Student with the School of Electronic Engineering and Computer Science, Queen Mary University of London (QMUL), UK. He continued his research as a Research Associate with the Department of Computer Science, City University of Hong Kong (CityU), from 2018 to 2019. Now he is a postdoc research fellow working with Prof. Ming-Ming Cheng at Nankai University. His research interests lie in image processing, computer vision, and deep learning.

Linhao Han is currently a master student at the College of Computer Science, Nankai University, under the supervision of Prof. Ming-Ming Cheng. His research interests include deep learning and computer vision.



Jun Jiang received the PhD degree in Color Science from Rochester Institute of Technology in 2013. He is a Senior Researcher in SenseBrain focusing on algorithm development to improve image quality on smartphone cameras. His research interest includes computational photography, low-level computer vision, and deep learning.

Ming-Ming Cheng (Senior Member, IEEE) received the Ph.D. degree from Tsinghua University in 2012. Then he did two years research fellowship with Prof. Philip Torr at Oxford. He is currently a Professor at Nankai University and leading the Media Computing Laboratory. His research interests include computer graphics, computer vision, and image processing. He received research awards, including the ACM China Rising Star Award, the IBM Global SUR Award, and the CCF-Intel Young Faculty Researcher Program. He is on the Editorial Board Member of IEEE Transactions on Image Processing (TIP).

Jinwei Gu (Senior Member, IEEE) is the R&D Executive Director of SenseTime USA. His current research focuses on low-level computer vision, computational photography, smart visual sensing and perception, and robotics. He obtained his Ph.D. degree in 2010 from Columbia University, and his B.S and M.S. from Tsinghua University, in 2002 and 2005 respectively. Before joining SenseTime, he was a senior research scientist in NVIDIA Research from 2015 to 2018. Prior to that, he was an assistant professor in Rochester Institute of Technology from 2010 to 2013, and a senior researcher in the media lab of Futurewei Technologies from 2013 to 2015. He is an associate editor for IEEE Transactions on Computational Imaging and an IEEE senior member since 2018.

Chen Change Loy (Senior Member, IEEE) is an Associate Professor with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He is also an Adjunct Associate Professor at The Chinese University of Hong Kong. He received his Ph.D. (2010) in Computer Science from the Queen Mary University of London. Prior to joining NTU, he served as a Research Assistant Professor at the MMLab of The Chinese University of Hong Kong, from 2013 to 2018. He was a postdoctoral researcher at Queen Mary University of London and Vision Semantics Limited, from 2010 to 2013. He serves as an Associate Editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence and International Journal of Computer Vision. He also serves/served as an Area Chair of major conferences such as ICCV, CVPR, ECCV and AAAI. His research interests include image/video restoration and enhancement, generative tasks, and representation learning.

Low-Light Image and Video Enhancement Using Deep Learning: A Survey (Supplementary Material)

Chongyi Li, Chunle Guo, Linghao Han, Jun Jiang, Ming-Ming Cheng, *Senior Member, IEEE*, Jinwei Gu, *Senior Member, IEEE*, and Chen Change Loy, *Senior Member, IEEE*

Project Page: https://www.mmlab-ntu.com/project/lliv_survey/index.html.

- In this supplementary material, we provide more visual comparisons of the results enhanced by different methods on a variety of input scenes sampled from different testing benchmark datasets. Besides, we follow Jiang and Zheng [1] to plot the luminance curves for comparing the temporal coherence of enhanced videos. The smoother the curve is, the better the method for temporal coherence is. We also provide the average luminance variance value (the smaller, the better).
- We also upload the video results of different methods to YouTube at https://www.youtube.com/watch?v=Elo9TkrG5Oo&t=6s.
- One can test the performance of different methods with any inputs on our online platform at http://mc.nankai.edu.cn/ll/.
- We collect low-light image and video enhancement methods, datasets, and evaluation metrics and periodically update the content in at https://github.com/Li-Chongyi/Lighting-the-Darkness-in-the-Deep-Learning-Era-Open.
- We release our proposed dataset at https://drive.google.com/file/d/1QS4FgT5aTQNYy-eHZ_A89rLoZgx_iysR/view.

In what follows, we present the visual results of different methods. Specifically,

Figures 1, 2, and 3 show the results enhanced by different deep learning-based low-light image enhancement methods on the low-light images sampled from LOL-test dataset [2].

Figures 4, 5, and 6 show the results enhanced by different deep learning-based low-light image enhancement methods on the low-light images sampled from MIT-Adobe FiveK-test dataset [3].

Figures 7, 8, and 9 show the results enhanced by different deep learning-based low-light image enhancement methods on the low-light images sampled from our proposed LLIV-Phone-imgT dataset.

Figures 10 and 11 show the results enhanced by different deep learning-based low-light image enhancement methods on the raw low-light images sampled from SID-test dataset [4].

Figures 12, 13, and 14 show the results enhanced by different deep learning-based low-light image enhancement methods on the low-light images sampled from DARK FACE dataset [5] and their face detection resutls.

Figures 15, 16, 17, 18, 19, 20, 21, 22, 23, and 24 show the luminance curves of the videos enhanced by different methods. These ten low-light videos were taken by different mobile phones' cameras and sampled from our proposed LLIV-Phone-vidT dataset.

(a) input (b) LLNet [6] (c) LightenNet [7]

(d) Retinex-Net [2] (e) MBLLEN [8] (f) KinD [9]

(g) KinD++ [10] (h) TBEFN [11] (i) DSLR [12]

(j) EnlightenGAN [13] (k) DRBN [14] (l) ExCNet [15]

(m) Zero-DCE [16] (n) RRDNet [17] (o) GT

Fig. 1: Visual results of different methods on a low-light image sampled from LOL-test dataset [2].

C. Li and C. C. Loy are with the S-Lab, Nanyang Technological University (NTU), Singapore (e-mail: chongyi.li@ntu.edu.sg and ccloy@ntu.edu.sg).
C. Guo, L. Han, and M.-M. Cheng are with the College of Computer Science, Nankai University, Tianjin, China (e-mail: guochunle@nankai.edu.cn, lhhan@mail.nankai.edu.cn, and cmm@nankai.edu.cn).
J. Jiang and J. Gu are with the SenseTime (e-mail: jiangjun@sensebrain.site and gujinwei@sensebrain.site).
C. Li and C. Guo contribute equally.
C. C. Loy is the corresponding author.

(a) input (b) LLNet [6] (c) LightenNet [7]

(d) Retinex-Net [2] (e) MBLLEN [8] (f) KinD [9]

(g) KinD++ [10] (h) TBEFN [11] (i) DSLR [12]

(j) EnlightenGAN [13] (k) DRBN [14] (l) ExCNet [15]

(m) Zero-DCE [16] (n) RRDNet [17] (o) GT

Fig. 2: Visual results of different methods on a low-light image sampled from LOL-test dataset [2].

(a) input (b) LLNet [6] (c) LightenNet [7]

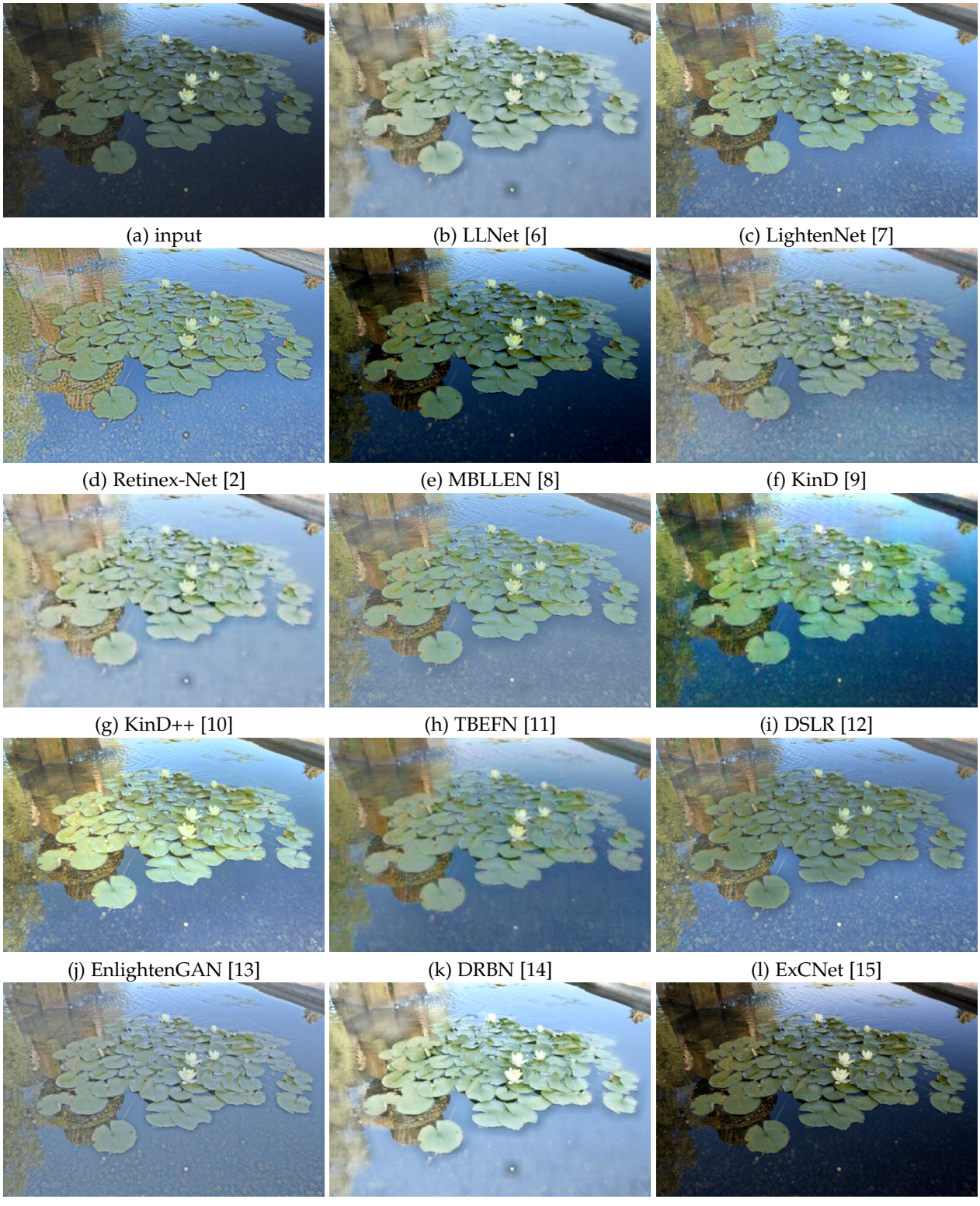
(d) Retinex-Net [2] (e) MBLLEN [8] (f) KinD [9]

(g) KinD++ [10] (h) TBEFN [11] (i) DSLR [12]

(j) EnlightenGAN [13] (k) DRBN [14] (l) ExCNet [15]

(m) Zero-DCE [16] (n) RRDNet [17] (o) GT

Fig. 3: Visual results of different methods on a low-light image sampled from LOL-test dataset [2].



(a) input (b) LLNet [6] (c) LightenNet [7]

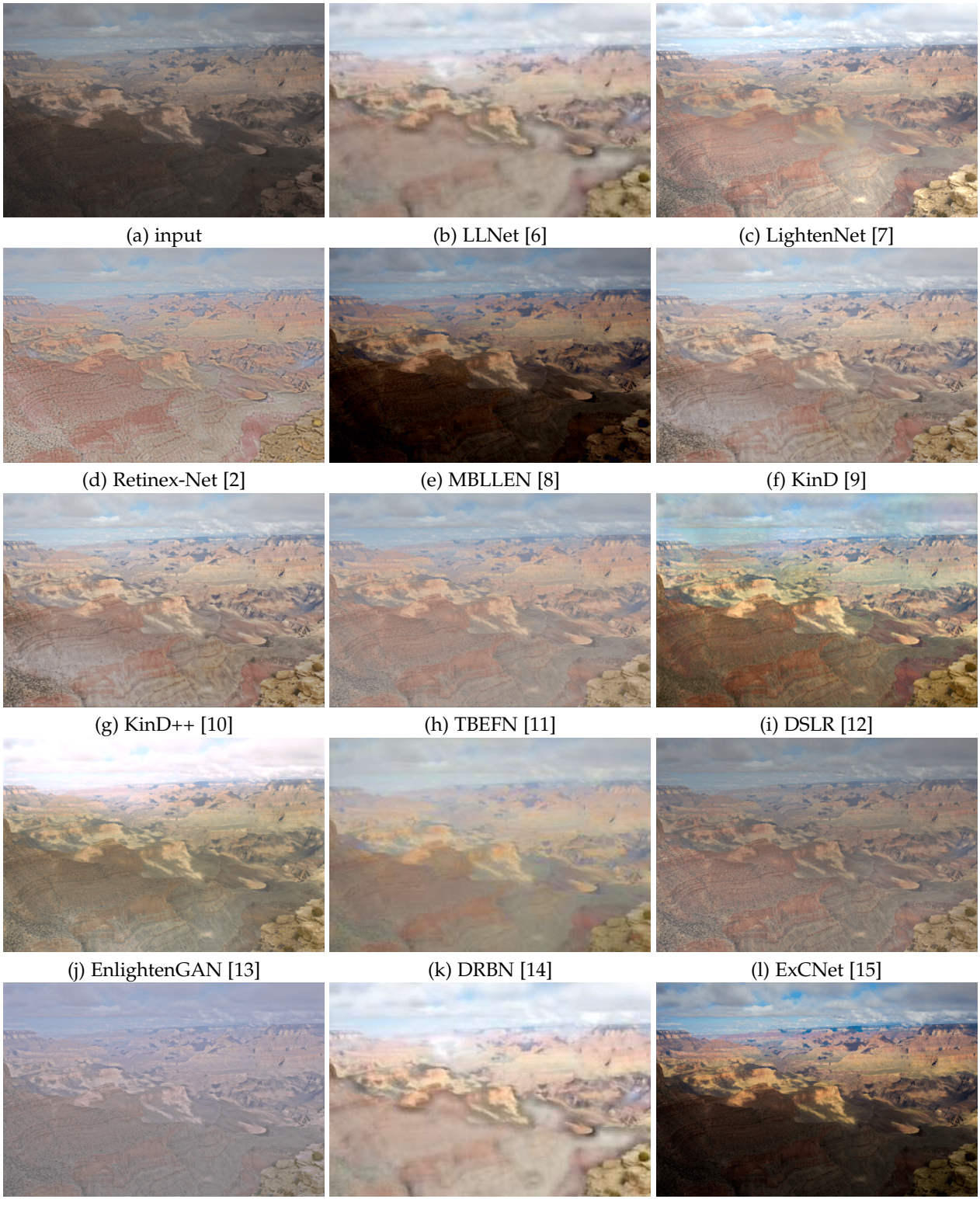
(d) Retinex-Net [2] (e) MBLLEN [8] (f) KinD [9]

(g) KinD++ [10] (h) TBEFN [11] (i) DSLR [12]

(j) EnlightenGAN [13] (k) DRBN [14] (l) ExCNet [15]

(m) Zero-DCE [16] (n) RRDNet [17] (o) GT

Fig. 4: Visual results of different methods on a low-light image sampled from MIT-Adobe FiveK-test dataset [3].



(a) input (b) LLNet [6] (c) LightenNet [7]

(d) Retinex-Net [2] (e) MBLLEN [8] (f) KinD [9]

(g) KinD++ [10] (h) TBEFN [11] (i) DSLR [12]

(j) EnlightenGAN [13] (k) DRBN [14] (l) ExCNet [15]

(m) Zero-DCE [16] (n) RRDNet [17] (o) GT

Fig. 5: Visual results of different methods on a low-light image sampled from MIT-Adobe FiveK-test dataset [3].

(a) input (b) LLNet [6] (c) LightenNet [7]

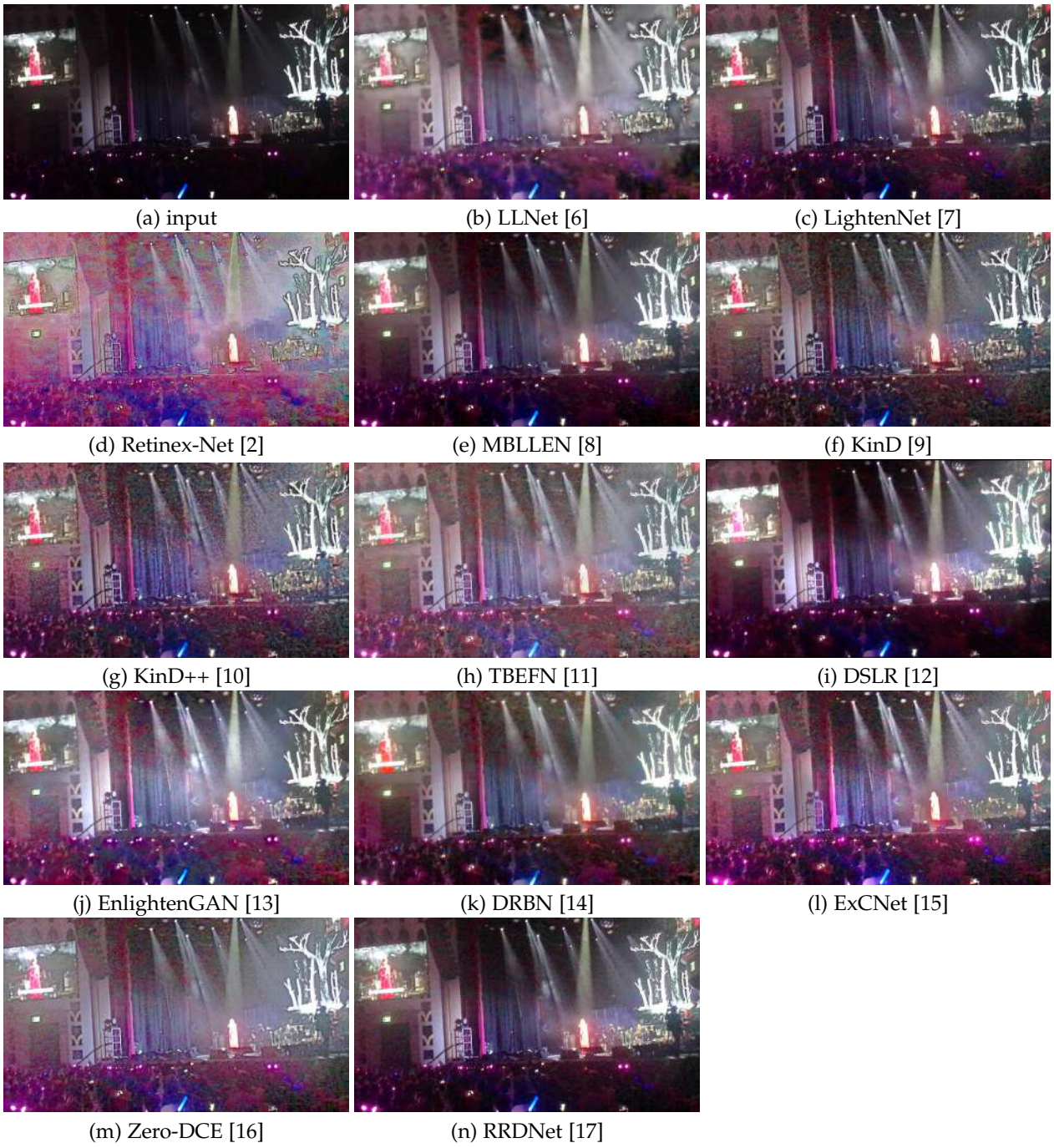
(d) Retinex-Net [2] (e) MBLLEN [8] (f) KinD [9]

(g) KinD++ [10] (h) TBEFN [11] (i) DSLR [12]

(j) EnlightenGAN [13] (k) DRBN [14] (l) ExCNet [15]

(m) Zero-DCE [16] (n) RRDNet [17] (o) GT

Fig. 6: Visual results of different methods on a low-light image sampled from MIT-Adobe FiveK-test dataset [3].



(a) input (b) LLNet [6] (c) LightenNet [7]

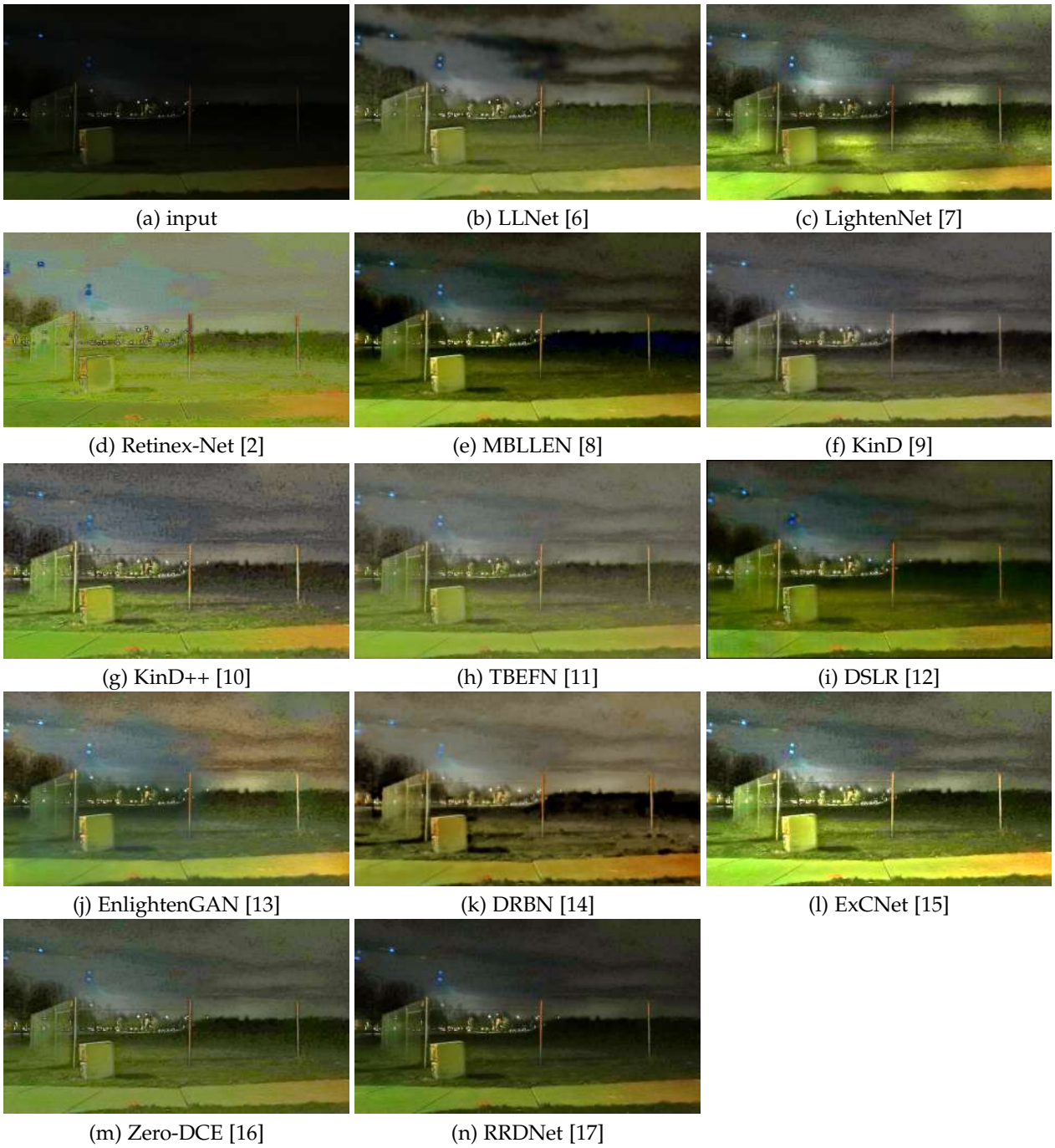
(d) Retinex-Net [2] (e) MBLLEN [8] (f) KinD [9]

(g) KinD++ [10] (h) TBEFN [11] (i) DSLR [12]

(j) EnlightenGAN [13] (k) DRBN [14] (l) ExCNet [15]

(m) Zero-DCE [16] (n) RRDNet [17]

Fig. 7: Visual results of different methods on a low-light image sampled from LLIV-Phone-imgT dataset.



(a) input (b) LLNet [6] (c) LightenNet [7]

(d) Retinex-Net [2] (e) MBLLEN [8] (f) KinD [9]

(g) KinD++ [10] (h) TBEFN [11] (i) DSLR [12]

(j) EnlightenGAN [13] (k) DRBN [14] (l) ExCNet [15]

(m) Zero-DCE [16] (n) RRDNet [17]

Fig. 8: Visual results of different methods on a low-light image sampled from LLIV-Phone-imgT dataset.

(a) input (b) LLNet [6] (c) LightenNet [7]

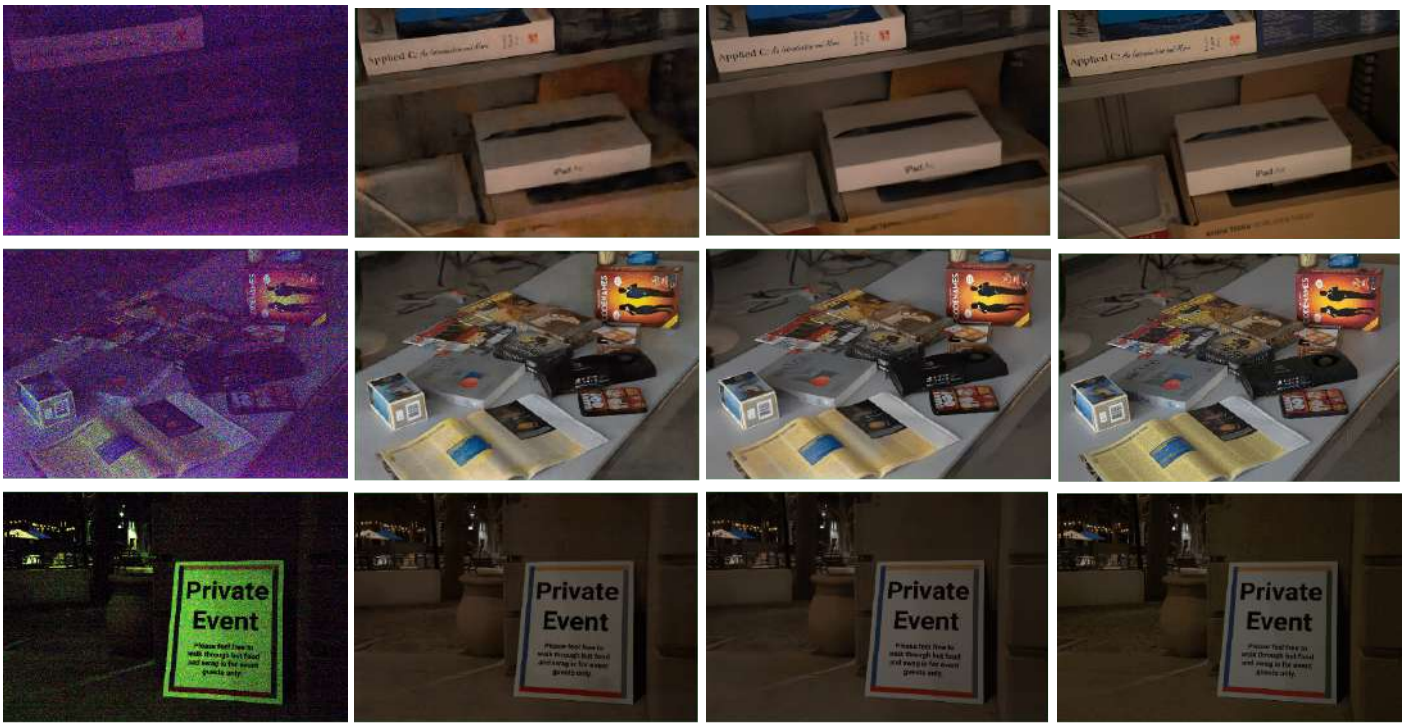
(d) Retinex-Net [2] (e) MBLLEN [8] (f) KinD [9]

(g) KinD++ [10] (h) TBEFN [11] (i) DSLR [12]

(j) EnlightenGAN [13] (k) DRBN [14] (l) ExCNet [15]

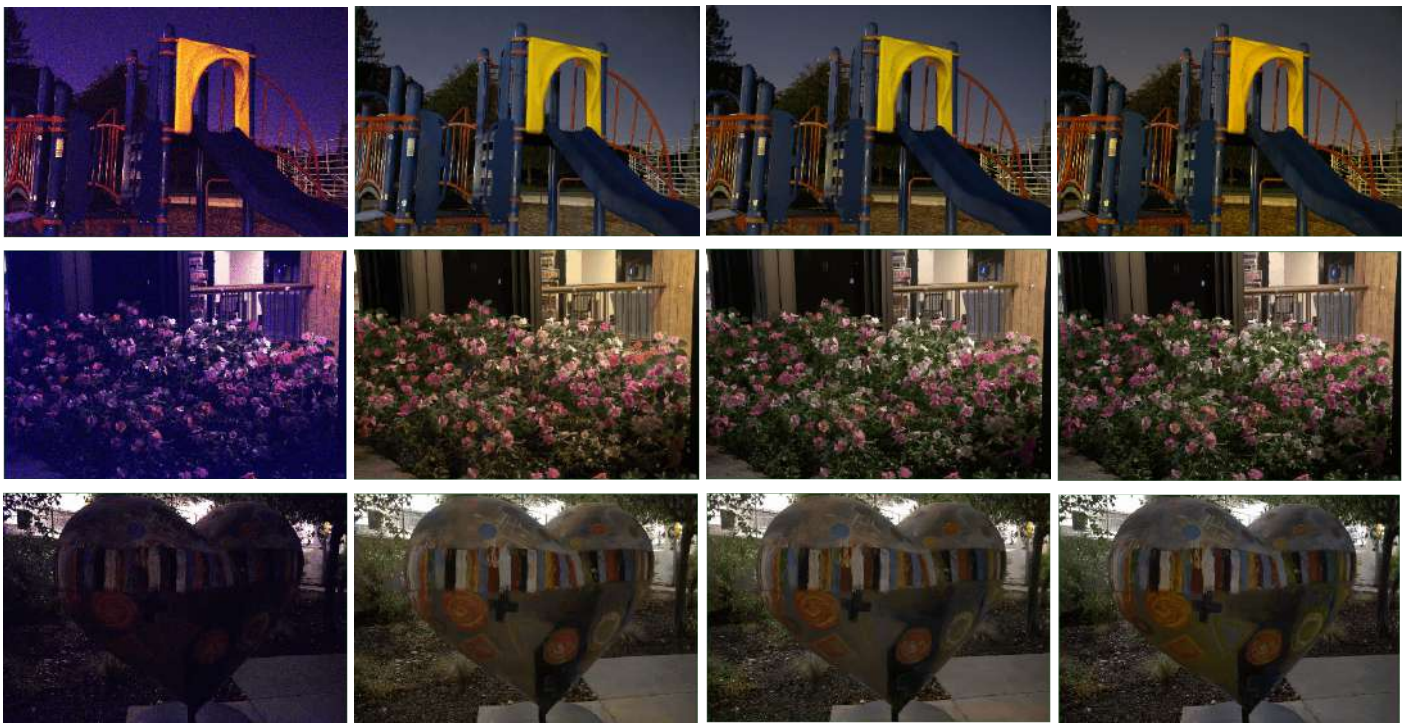
(m) Zero-DCE [16] (n) RRDNet [17]

Fig. 9: Visual results of different methods on a low-light image sampled from LLIV-Phone-imgT dataset.



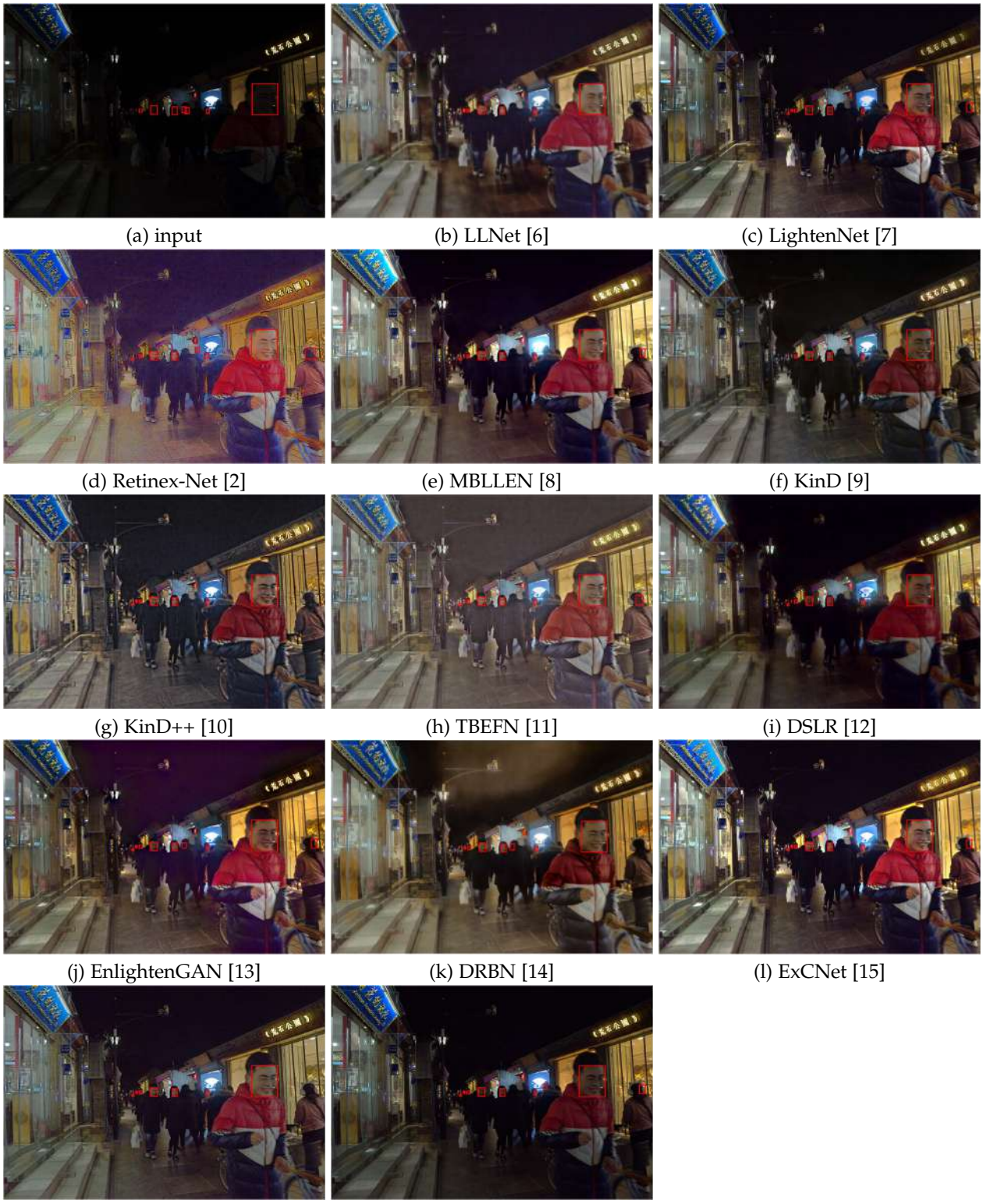
(a) inputs (b) SID [18] (c) EEMEFN [19] (d) GT

Fig. 10: Visual results of different methods on raw low-light images of Bayer pattern sampled from SID-test-Bayer test dataset. The inputs are amplified for visualization.



(a) inputs (b) SID [18] (c) EEMEFN [19] (d) GT

Fig. 11: Visual results of different methods on raw low-light images of APS-C X-Trans pattern sampled from SID-test-X-Trans test dataset. The inputs are amplified for visualization.



(a) input (b) LLNet [6] (c) LightenNet [7]

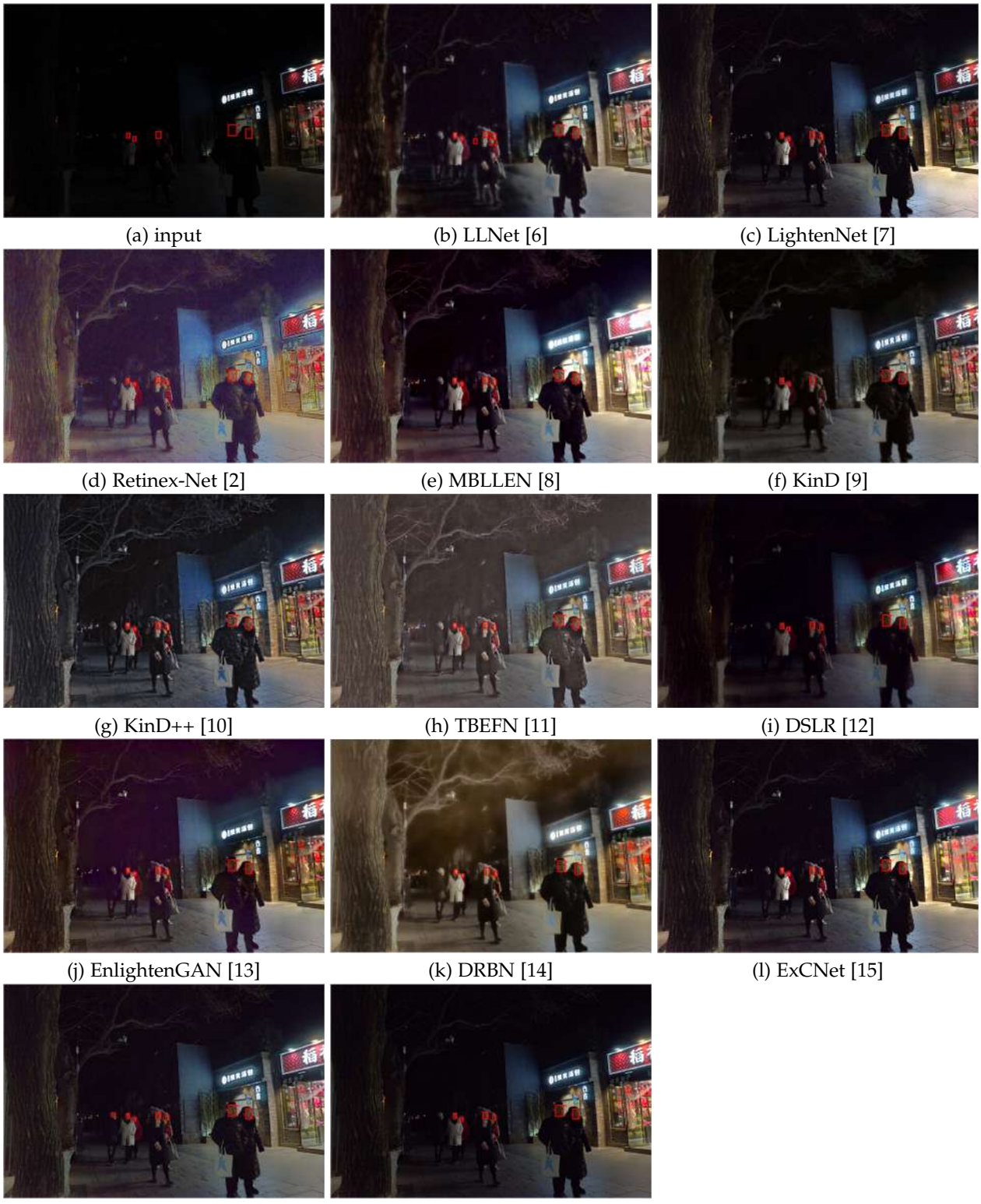
(d) Retinex-Net [2] (e) MBLLEN [8] (f) KinD [9]

(g) KinD++ [10] (h) TBEFN [11] (i) DSLR [12]

(j) EnlightenGAN [13] (k) DRBN [14] (l) ExCNet [15]

(m) Zero-DCE [16] (n) RRDNet [17]

Fig. 12: Visual results of different methods on a low-light image sampled from DARK FACE dataset [5]. Better see with zoom in for the bounding boxes of faces.



(a) input (b) LLNet [6] (c) LightenNet [7]

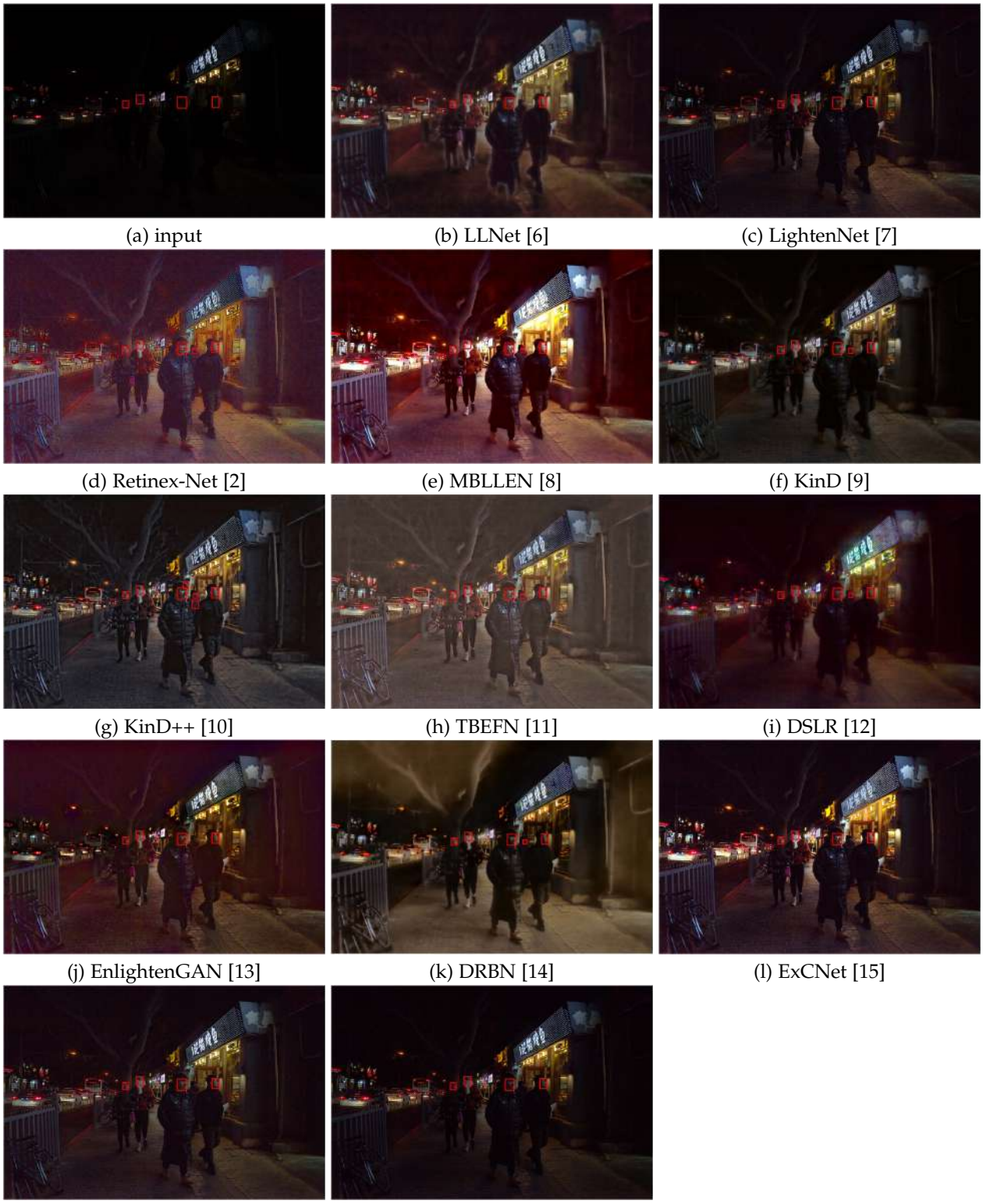
(d) Retinex-Net [2] (e) MBLLEN [8] (f) KinD [9]

(g) KinD++ [10] (h) TBEFN [11] (i) DSLR [12]

(j) EnlightenGAN [13] (k) DRBN [14] (l) ExCNet [15]

(m) Zero-DCE [16] (n) RRDNet [17]

Fig. 13: Visual results of different methods on a low-light image sampled from DARK FACE dataset [5]. Better see with zoom in for the bounding boxes of faces.



(a) input (b) LLNet [6] (c) LightenNet [7]

(d) Retinex-Net [2] (e) MBLLEN [8] (f) KinD [9]

(g) KinD++ [10] (h) TBEFN [11] (i) DSLR [12]

(j) EnlightenGAN [13] (k) DRBN [14] (l) ExCNet [15]

(m) Zero-DCE [16] (n) RRDNet [17]

Fig. 14: Visual results of different methods on a low-light image sampled from DARK FACE dataset [5]. Better see with zoom in for the bounding boxes of faces.

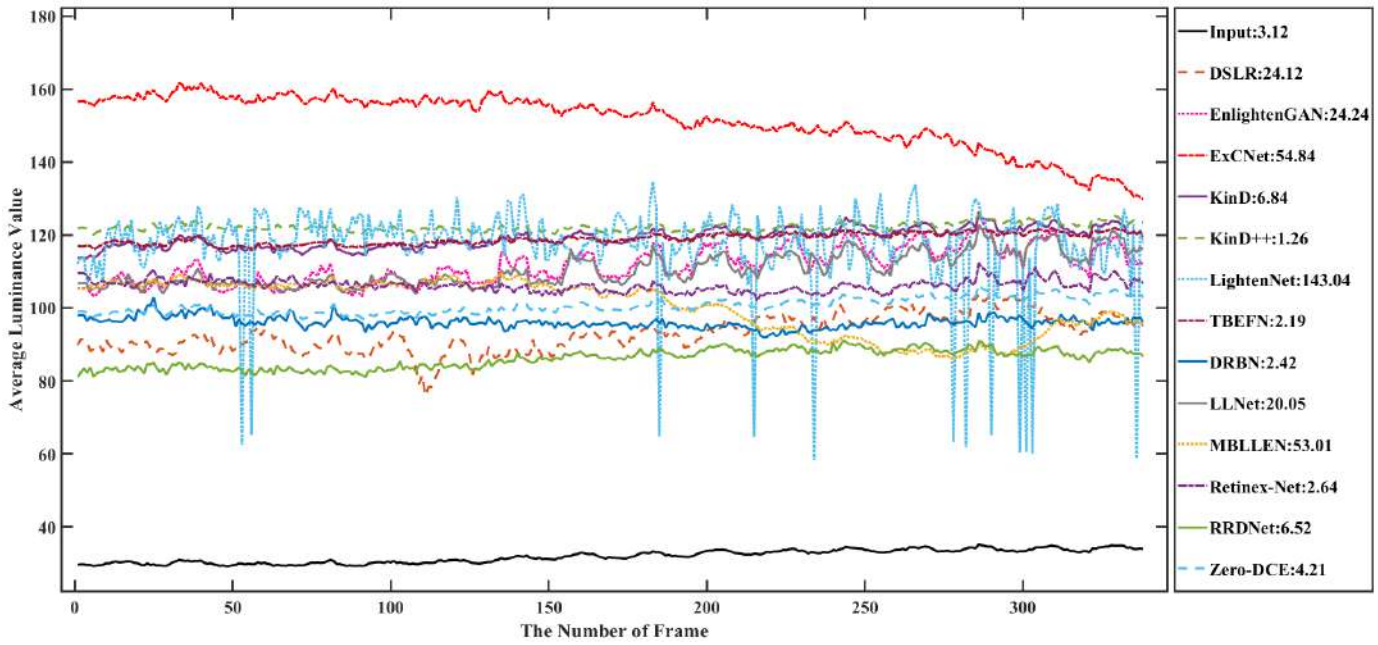


Fig. 15: Luminance curves of the video enhanced by different methods. The curve is plotted by computing average luminance value of each bounding box in consecutive frames. The smoother the curve is, the better the method for video enhancement (i.e., temporal coherence) is. The numbers in the legend represent the average luminance variance values (the smaller, the better). The low-light video was taken by a Huawei Mate 20 Pro phone's camera.

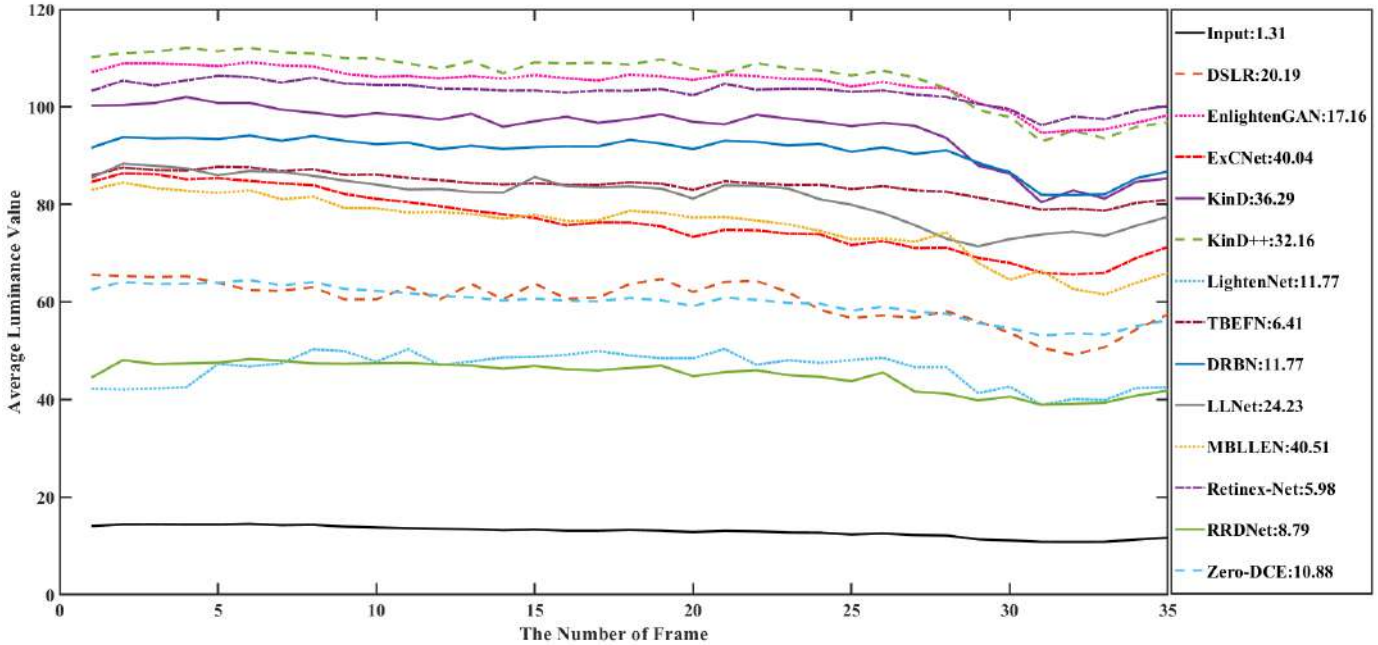


Fig. 16: Luminance curves of the video enhanced by different methods. The curve is plotted by computing average luminance value of each bounding box in consecutive frames. The smoother the curve is, the better the method for video enhancement (i.e., temporal coherence) is. The numbers in the legend represent the average luminance variance values (the smaller, the better). The low-light video was taken by an iPhone 6s phone's camera.

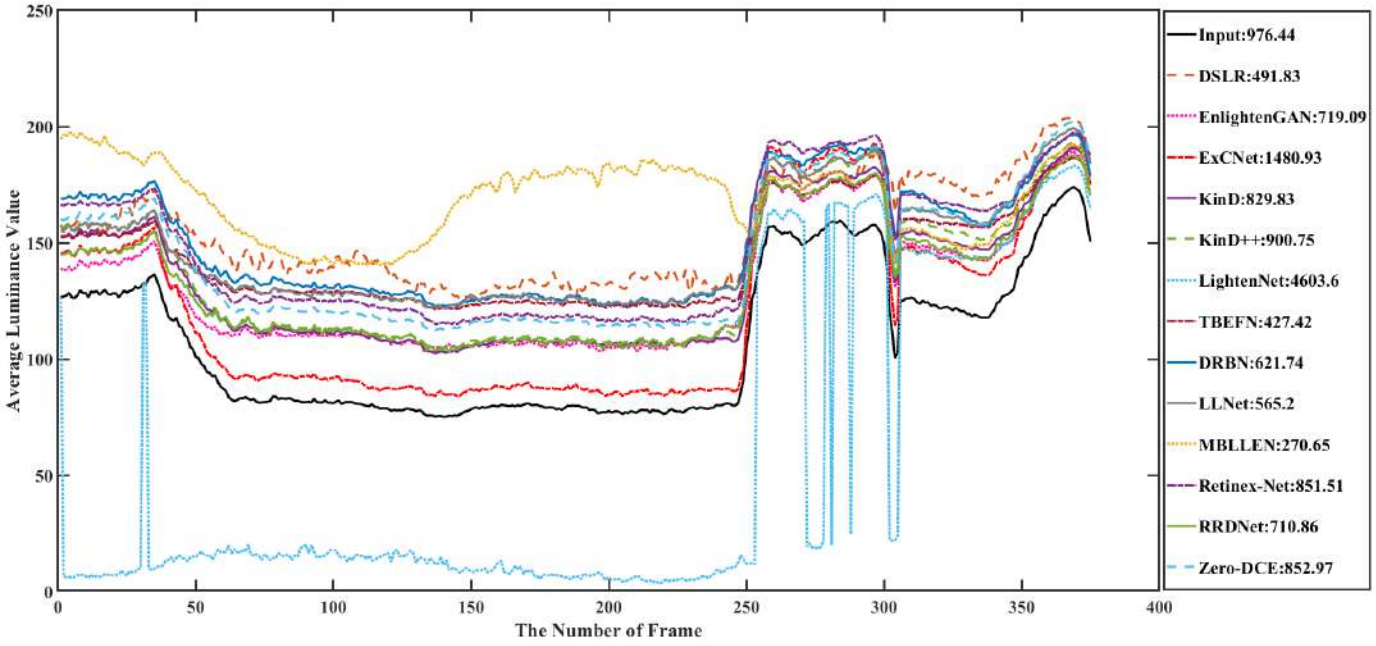


Fig. 17: Luminance curves of the video enhanced by different methods. The curve is plotted by computing average luminance value of each bounding box in consecutive frames. The smoother the curve is, the better the method for video enhancement (i.e., temporal coherence) is. The numbers in the legend represent the average luminance variance values (the smaller, the better). The low-light video was taken by an iPhone 7 phone's camera.

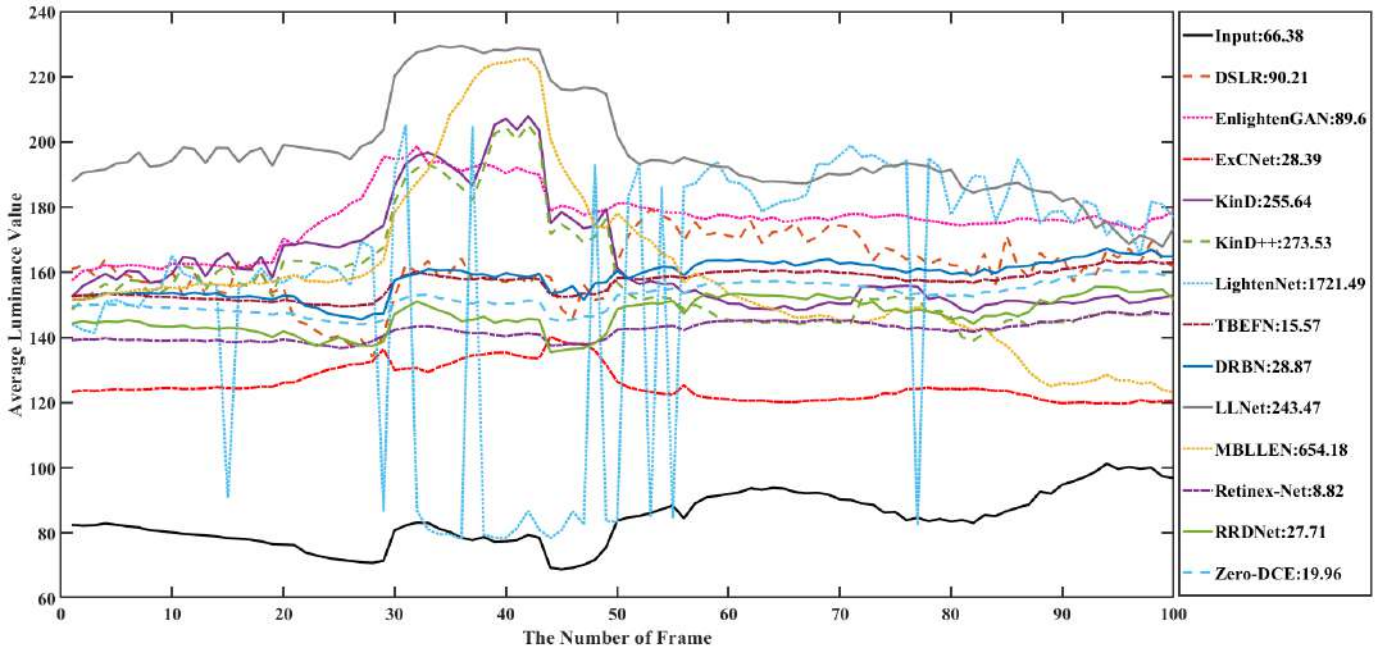


Fig. 18: Luminance curves of the video enhanced by different methods. The curve is plotted by computing average luminance value of each bounding box in consecutive frames. The smoother the curve is, the better the method for video enhancement (i.e., temporal coherence) is. The numbers in the legend represent the average luminance variance values (the smaller, the better). The low-light video was taken by an iPhone7 Plus phone's camera.

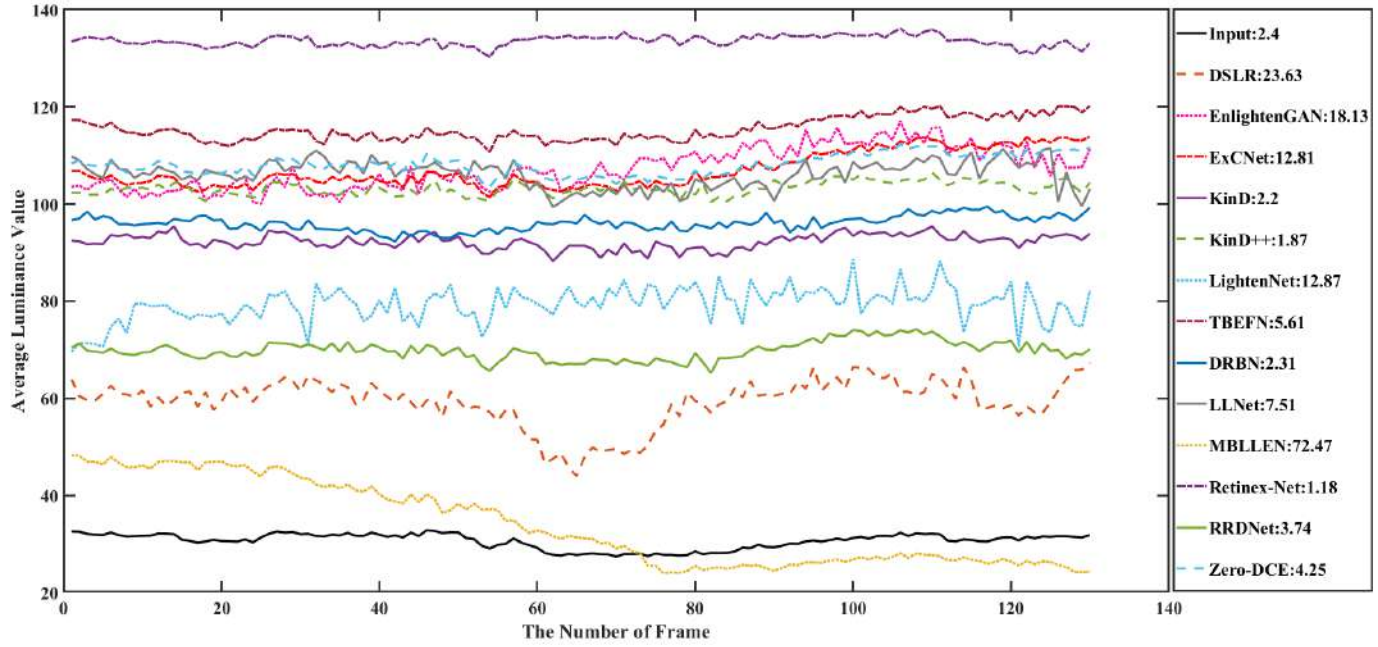


Fig. 19: Luminance curves of the video enhanced by different methods. The curve is plotted by computing average luminance value of each bounding box in consecutive frames. The smoother the curve is, the better the method for video enhancement (i.e., temporal coherence) is. The numbers in the legend represent the average luminance variance values (the smaller, the better). The low-light video was taken by an iPhone8 Plus phone's camera.

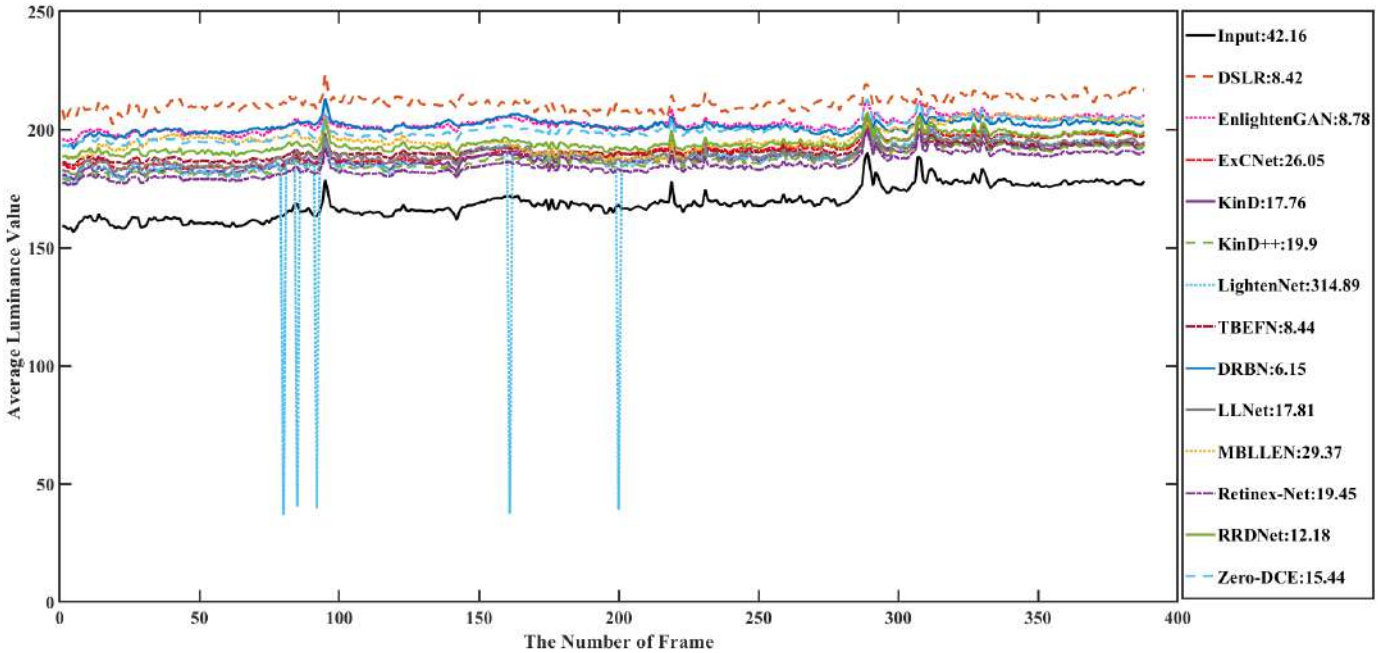


Fig. 20: Luminance curves of the video enhanced by different methods. The curve is plotted by computing average luminance value of each bounding box in consecutive frames. The smoother the curve is, the better the method for video enhancement (i.e., temporal coherence) is. The numbers in the legend represent the average luminance variance values (the smaller, the better). The low-light video was taken by an iPhone 11 phone's camera.

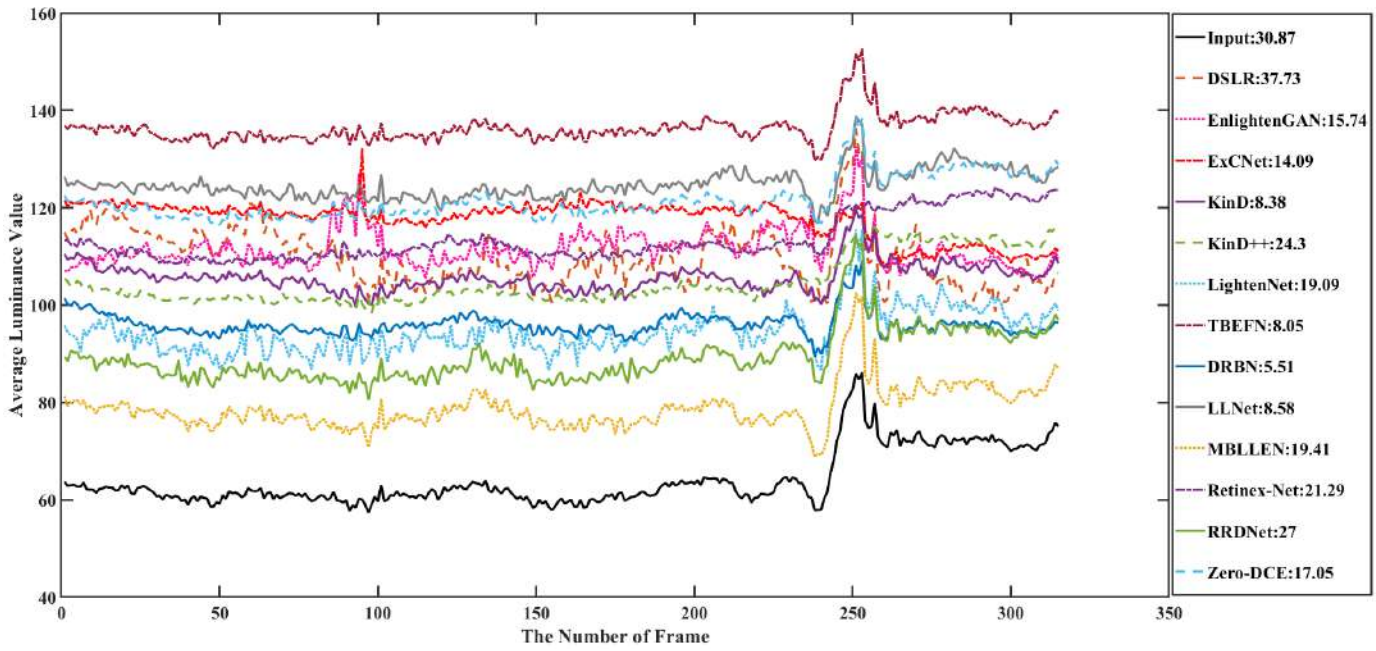


Fig. 21: Luminance curves of the video enhanced by different methods. The curve is plotted by computing average luminance value of each bounding box in consecutive frames. The smoother the curve is, the better the method for video enhancement (i.e., temporal coherence) is. The numbers in the legend represent the average luminance variance values (the smaller, the better). The low-light video was taken by an iPhone11 Pro phone's camera.

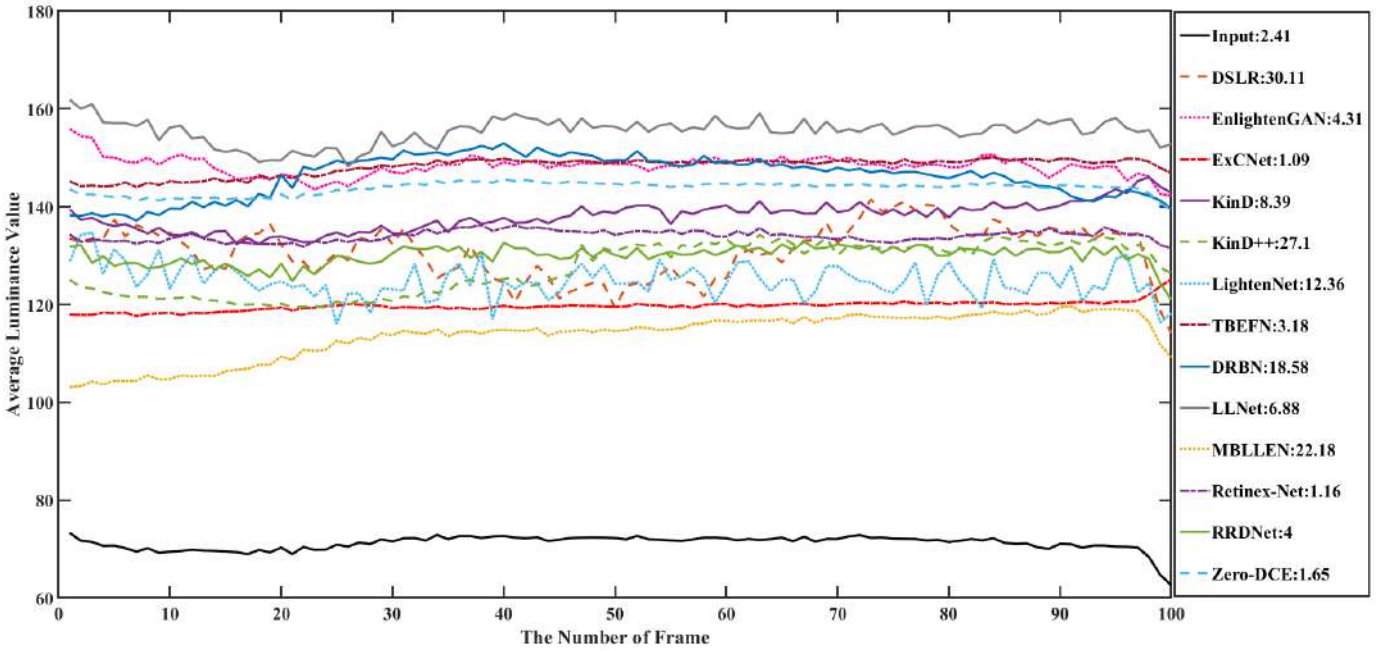


Fig. 22: Luminance curves of the video enhanced by different methods. The curve is plotted by computing average luminance value of each bounding box in consecutive frames. The smoother the curve is, the better the method for video enhancement (i.e., temporal coherence) is. The numbers in the legend represent the average luminance variance values (the smaller, the better). The low-light video was taken by an iPhone SE phone's camera.

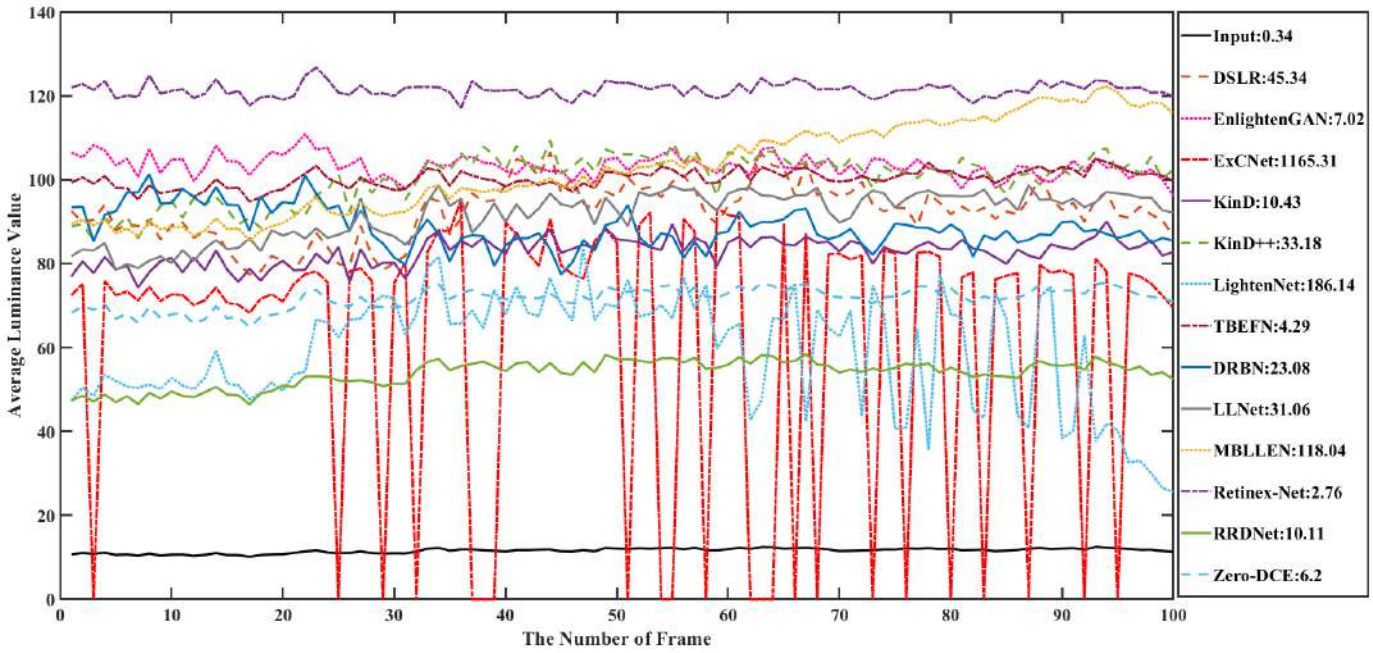


Fig. 23: Luminance curves of the video enhanced by different methods. The curve is plotted by computing average luminance value of each bounding box in consecutive frames. The smoother the curve is, the better the method for video enhancement (i.e., temporal coherence) is. The numbers in the legend represent the average luminance variance values (the smaller, the better). The low-light video was taken by an iPhone XS phone's camera.

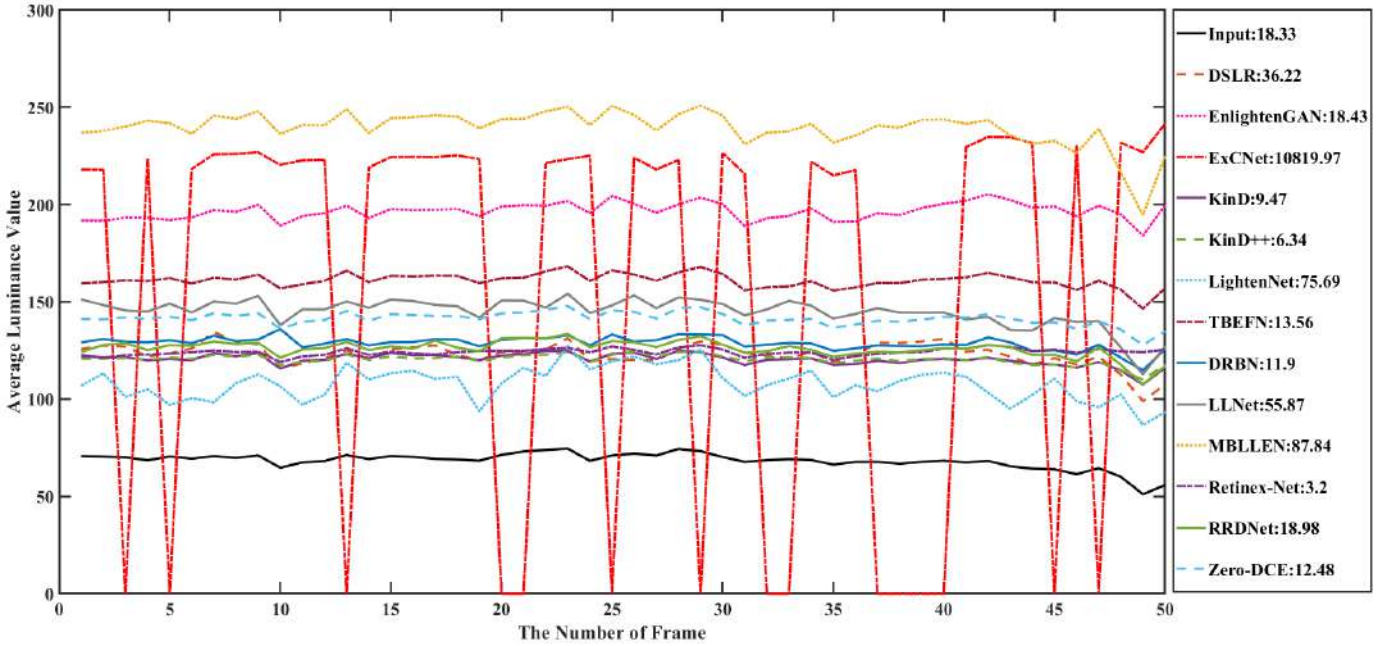


Fig. 24: Luminance curves of the video enhanced by different methods. The curve is plotted by computing average luminance value of each bounding box in consecutive frames. The smoother the curve is, the better the method for video enhancement (i.e., temporal coherence) is. The numbers in the legend represent the average luminance variance values (the smaller, the better). The low-light video was taken by an OnePlus 5T phone's camera.

REFERENCES

[1] H. Jiang and Y. Zheng, "Learning to see moving object in the dark," in *ICCV*, 2019, pp. 7324–7333.
[2] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," in *BMVC*, 2018.
[3] V. Bychkovsky, S. Paris, E. Chan, and F. Durand, "Learning photographic global tonal adjustment with a database of input/output image pairs," in *CVPR*, 2011, pp. 97–104.
[4] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *CVPR*, 2018, pp. 3291–3300.
[5] Y. Yuan, W. Yang, W. Ren, J. Liu, W. JScheirer, and W. Zhangyang, "UG+ Track 2: A collective benchmark effort for evaluating and advancing image understanding in poor visibility environments," *arXiv arXiv:1904.04474*, 2019.
[6] K. G. Lore, A. Akintayo, and S. Sarkar, "LLNet: A deep autoencoder approach to natural low-light image enhancement," *PR*, vol. 61, pp. 650–662, 2017.
[7] C. Li, J. Guo, F. Porikli, and Y. Pang, "LightenNet: A convolutional neural network for weakly illuminated image enhancement," *PRL*, vol. 104, pp. 15–22, 2018.
[8] F. Lv, F. Lu, J. Wu, and C. Lim, "MBLLEN: Low-light image/video enhancement using cnns," in *BMVC*, 2018.
[9] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *ACMMM*, 2019, pp. 1632–1640.
[10] X. Guo, Y. Zhang, J. Ma, W. Liu, and J. Zhang, "Beyond brightening low-light images," *IJCV*, 2020.
[11] K. Lu and L. Zhang, "TBEFN: A two-branch exposure-fusion network for low-light image enhancement," *TMM*, 2020.
[12] S. Lim and W. Kim, "DSLR: Deep stacked laplacian restorer for low-light image enhancement," *TMM*, 2020.
[13] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "EnlightenGAN: Deep light enhancement without paired supervision," *TIP*, vol. 30, pp. 2340–2349, 2021.
[14] 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 *CVPR*, 2020, pp. 3063–3072.
[15] L. Zhang, L. Zhang, X. Liu, Y. Shen, S. Zhang, and S. Zhao, "Zero-shot restoration of back-lit images using deep internal learning," in *ACMMM*, 2019, pp. 1623–1631.
[16] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *CVPR*, 2020, pp. 1780–1789.
[17] A. Zhu, L. Zhang, Y. Shen, Y. Ma, S. Zhao, and Y. Zhou, "Zero-shot restoration of underexposed images via robust retinex decomposition," in *ICME*, 2020, pp. 1–6.
[18] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *CVPR*, 2018, pp. 3291–3300.
[19] M. Zhu, P. Pan, W. Chen, and Y. Yang, "EEMEFN: Low-light image enhancement via edge-enhanced multi-exposure fusion network," in *AAAI*, 2020, pp. 13 106–13 113.