International Journal of Computer Vision (2021) 129:1153–1184
https://doi.org/10.1007/s11263-020-01418-8

Benchmarking Low-Light Image Enhancement and Beyond

Jiaying Liu[1] · Dejia Xu[1] · Wenhan Yang[1] · Minhao Fan[1] · Haofeng Huang[1]

Received: 16 March 2020 / Accepted: 4 December 2020 / Published online: 11 January 2021

Abstract

In this paper, we present a systematic review and evaluation of existing single-image low-light enhancement algorithms. Besides the commonly used low-level vision oriented evaluations, we additionally consider measuring machine vision performance in the low-light condition via face detection task to explore the potential of joint optimization of high-level and low-level vision enhancement. To this end, we first propose a large-scale low-light image dataset serving both low/high-level vision with diversified scenes and contents as well as complex degradation in real scenarios, called Vision Enhancement in the LOw-Light condition (VE-LOL). Beyond paired low/normal-light images without annotations, we additionally include the analysis resource related to human, i.e. face images in the low-light condition with annotated face bounding boxes. Then, efforts are made on benchmarking from the perspective of both human and machine visions. A rich variety of criteria is used for the low-level vision evaluation, including full-reference, no-reference, and semantic similarity metrics. We also measure the effects of the low-light enhancement on face detection in the low-light condition. State-of-the-art face detection methods are used in the evaluation. Furthermore, with the rich material of VE-LOL, we explore the novel problem of joint low-light enhancement and face detection. We develop an enhanced face detector to apply low-light enhancement and face detection jointly. The features extracted by the enhancement module are fed to the successive layer with the same resolution of the detection module. Thus, these features are intertwined together to unitedly learn useful information across two phases, i.e. enhancement and detection. Experiments on VE-LOL provide a comparison of state-of-the-art low-light enhancement algorithms, point out their limitations, and suggest promising future directions. Our dataset has supported the Track "Face Detection in Low Light Conditions" of CVPR UG2+ Challenge (2019–2020) (http://cvpr2020.ug2challenge.org/).

Keywords Low-light enhancement · Benchmark · Dataset · Face detection

1 Introduction

Low-light image capturing conditions lead to several kinds of degradations presented in images, including low visibility, color cast, and intensive noise, *etc*. To handle this degradation, low-light image enhancement methods are proposed to enhance the visibility and visual quality of the input image. The earliest methods directly amplify the illumination uniformly. Later methods adjust the global illumination property of an image, e.g. histogram equalization (HE) to make dark images visible by stretching the dynamic range of an image (Pizer et al. 1990; Abdullah-Al-Wadud et al. 2007). However, these methods might fail to well adjust the visual quality from all aspects, e.g. noise suppression. The Retinex-based methods (Jobson et al. 1997a, b; Fu et al. 2016; Guo et al. 2017; Fu et al. 2016; Li et al. 2018; Ren et al. 2018) are a very important branch. It is first proposed as a model of human visual perception (Land and McCann 1971) designed

Communicated by Dengxin Dai.

✉ Jiaying Liu
liujiaying@pku.edu.cn

✉ Wenhan Yang
yangwenhan@pku.edu.cn

Dejia Xu
dejia@pku.edu.cn

Minhao Fan
fanminhao@pku.edu.cn

Haofeng Huang
huang6013@pku.edu.cn

[1] Wangxuan Institute of Computer Technology, Peking University, Beijing, China

Table 1 Comparison between existing low-light enhancement datasets and VE-LOL

Properties	Number (training/testing)	Synthetic/real	Source
Paired dataset			
Phos (Vonikakis et al. 2013)	225	Real	Camera
LLNet (Lore et al. 2017)[1]	211,250/211,250 (patches)	Synthetic	Testing images
MSR-Net (Shen et al. 2017)[1]	8000/2000	Synthetic	UCID/BSD/Google
SID (Chen et al. 2018)	5094 (RAW images)	Real	Camera
SICE (Cai et al. 2018)	589 (sequences)	Real	Camera
SMOID (Jiang and Zheng 2019)	202 (RAW videos)	Real	Camera
DRV (Chen et al. 2019)	179 (RAW videos)	Real	Camera
LOL (Wei et al. 2018)	1485/15	Synthetic + Real	RAISE/Camera
DeepUPE (Wei et al. 2018)[1]	2750/250	Synthetic + Real	Flicker/Camera
VE-LOL-L	2100/400	Synthetic + Real	RAISE/Camera

Properties	Number (training/testing)	Synthetic/real	Annotations
Unpaired dataset			
LIME (Guo et al. 2017)	10	Real	No
DICM (Lee et al. 2013a)	69	Real	No
MEF (Ma et al. 2015)	17	Real	No
NPE (Wang et al. 2013a)	85	Real	No
VV[2]	24	Real	No
ExDARK (Loh and Chan 2019)	7363	Real	Object bounding boxes/categories
Nighttime driving (Dai and Gool 2018)	50 (35,000 unlabeled)	Real	Densely Annotated Semantic Labels
Dark Zurich (Sakaridis et al. 2019)	151 (5336 unlabeled)	Real	Densely annotated semantic labels
VE-LOL-H	6940/4000	Real	Face bounding boxes

[1] The datasets are still not publicly available

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

to compute visual appearance. Later on, based on the theoretical basis, the variational versions of Retinex model follow the paradigm of layer decomposition and are applied to low-light image enhancement (Jobson et al. 1997b). The methods decompose images into two components: reflectance and illumination. Then, the enhanced results are obtained by further processing and combining these two parts. With properly enforced priors and regularizations, the superiority of this branch is witnessed in noise suppression and high-frequency detail preservation. Recently, data-driven image processing applications have emerged with promising performance. Learning-based low-light image enhancement methods (Lore et al. 2017; Shen et al. 2017; Wei et al. 2018) have been studied. With the knowledge of large-scale data, these methods present the general effectiveness in handling low-light images in diversified conditions, to achieve an overall good visual quality.

Due to the rapid development of single-image low-light enhancement methods, researchers are paying more attention to this domain. However, few efforts have been made on a systematic review and comprehensive benchmark work based on a large-scale dataset for both low-level and high-level vision tasks, which can provide the community a retrospect of previous methods and a prospect of future work. However, in the first step, it is non-trivial to collect a comprehensive large-scale low-light image dataset, as it is hard to capture real low-light images (paired/unpaired) with certain kinds of contents, e.g. given objects, in a controllable environment (given the degradation condition). This limits our capacity to train, evaluate, and compare the strengths and limitations of different approaches based on large-scale benchmarks, especially the recent data-driven approaches, e.g. deep-learning methods. To the best of our knowledge, all the currently available low-light image datasets listed in Table 1 have limitations in various aspects.

First, previous datasets usually serve only one of human and machine vision purposes. Previous datasets (Vonikakis et al. 2018; Guo et al. 2017; Wang et al. 2013a; Lee et al. 2013a; Ma et al. 2015) consisting of only low-light images (without annotations) mainly focus on the subjective evaluation and user study from the perspective of visual quality. With the rise of deep learning-based methods (Lore et al. 2017; Shen et al. 2017; Wei et al. 2018), some new datasets provide paired low/normal-light images for training and evaluating low-light enhancement methods from the perspective of low-level signal fidelity. These two kinds of datasets are limited to the evaluation of low-level visual quality. Recently, Exclusively Dark dataset (Loh and Chan 2019) is proposed including low-light images with image classes and local object bounding box annotations to evaluate high-level vision tasks, i.e. image recognition and object detection. In Dai and Gool (2018), Sakaridis et al. (2019), two datasets including unlabeled nighttime images, unlabeled twilight images with correspondences to their daytime versions, and nighttime images with pixel-level dense annotations are proposed to serve evaluation of semantic segmentation at night.

Second, most of the previous datasets are not close to real scenario and diverse at the same time. Except for LOL, previous paired datasets include either synthetic or real low-light images, therefore suffer from the limitation: for synthetic images, the simulated degradation might deviate from the real one; for real images, the diversity of captured contents are usually limited, due to the resource cost of on-site shooting.

Third, previous low-light datasets do not pay enough attentions to analytics related to human. However, in real applications, the videos that include human and the related behaviors, such as surveillance videos, should be given a priority as they are high valuable and include the most critical information. The absence of such a dataset captured in low-light conditions leads to a lack of exploration and development of the related analytic approaches in the given degraded conditions. For example, there are also few considerations on how to construct and optimize a joint pipeline of image enhancement and detection methods.

Finally, the scale and diversity of most existing datasets are limited. Especially for unpaired datasets, before ExDARK, the images of previous datasets are fewer than 100, which limits the potential to provide a comprehensive and systematic evaluation.

In order to address these issues, a large-scale benchmark dataset, **V**isual **E**nhancement in **LO**w-**L**ight conditions (**VE-LOL**) with analysis resources related to human, is developed as the research material for exploring and evaluating vision enhancement approaches for both human perception and machine analytics. Based on the dataset, we make efforts in providing a detailed survey on low-light enhancement methods and a quantitative benchmarking of these methods from both perspectives of human and machine visions. Beyond that, with the wealth of the proposed dataset, we also explore the novel problem of joint optimization of low-light enhancement and face detection.

The paper has the following contributions:

- We introduce the newly proposed VE-LOL single-image low-light enhancement benchmark. It includes two subsets in the low-light condition: one including paired low and normal-light images (synthesized and captured) for the evaluation of visual quality enhancement, the other including captured low-light images with bounding boxes annotating the human faces for evaluation of face detection from the perspective of high-level computer vision. To the best of our knowledge, this is the first large-scale dataset captured in the low-light condition for both kinds of evaluation purposes. The superiorities of our VE-LOL are illustrated in detail in Sect. 3.
- We provide a detailed survey on previous datasets and methods focusing on single-image low-light enhancement. Our survey provides a holistic view of most of the existing methods. We believe it can provide a useful starting point to understand the main development of the field, the limitations of existing methods, and the possible future directions.
- Based on VE-LOL, we go a step further to conduct extensive and systematic experiments to quantitatively compare state-of-the-art single-image low-light enhancement methods with various evaluation criteria, including no-reference, full-reference, and high-level feature similarity metrics as well as the task-driven metric, *i.e.* face detection accuracy. Our evaluation and analysis demonstrate the performance and limitations of state-of-the-art algorithms, and bring in rich insights.
- Beyond the rich dataset and comprehensive benchmark analysis, we also explore to build a powerful face detector jointly optimized with a learnable low-light enhancement. A half cyclic constraint is introduced for image modeling and regularizing the training of the low-light image enhancer. The features at different levels of the low-light enhancement and face detection modules are correlated across the two phases, and learn to benefit each other mutually. Our preliminary attempt improves the face detection accuracy in the low-light condition and provides useful insights on the combination of low and high-level vision tasks to the community.

The rest of this paper is organized as follows. Section 2 provides a brief but systematic review of previous low-light enhancement datasets and approaches. Section 3 presents the proposed dataset and related analysis. In Sect. 4, the evaluation of representative state-of-the-art methods with diverse kinds of metrics is illustrated. Section 5 shows our exploration of joint low-light enhancement and face detection, with the wealth of VE-LOL. Finally, concluding remarks are provided in Sect. 6.

2 Literature Review of Low-Light Enhancement

2.1 Existing Datasets and Evaluations

Paired Datasets As data-driven methods become popular, some datasets are proposed for training and evaluations in low-light image enhancement. In (2013), Vonikakis et al. built a dataset including 225 images captured in 15 scenes. In every scene, 15 images are included: 9 images captured under various strengths of uniform illumination, while 6 images under different degrees of non-uniform illumination. In (2017), Lore et al. synthesized 422,500 patches based on 169 standard images[1] with the random Gamma transformation and Gaussian noise. In (2017), Shen et al. synthesized 10,000 low and normal-light image pairs with nature images from UCID (Schaefer and Stich 2004), BSD (Martin et al. 2001) and web images collected using Google search engine, and used 8000 and 2000 images for training and testing, respectively. In (2018), Chen et al. introduced the dataset See-in-the-Dark (SID) including 5094 short-exposure low-light raw images and corresponding long-exposure reference raw images. Cai et al. (2018) built the SICE dataset, including under/over-contrast and normal-contrast encoded image pairs, in which the reference normal-contrast images are generated from 589 image sequences and 4413 high-resolution images of different exposures by Multi-Exposure image Fusion (MEF) or High Dynamic Range (HDR) algorithm. LOw-Light (LOL) dataset (Wei et al. 2018) includes 500 captured paired images (485 pairs for training and another 15 ones for evaluation) and 1000 synthetic ones for training. Deep Underexposed Photo Enhancement (DeepUPE) (Wang et al. 2019) proposes a new dataset of 3000 underexposed photos (2750/250 for training and testing respectively) covering diverse lighting conditions. 85% images are captured in the resolution of 6000, 4000 with Canon EOS 5D Mark III and Sony ILCE-7 while around 15% more images are collected from Flickr. Dark Raw Video (DRV) (Jiang and Zheng 2019) includes 202 static videos, captured in indoor and outdoor scenes in different lighting conditions. The lighting range of the captured videos is in 0.5 to 5 lux range. The long and short exposed frames are well-aligned. See-Moving-Objects-in-the-Dark (SMOID) (Chen et al. 2019) includes 179 street view video pairs including moving vehicles and pedestrians at different exposure levels. Well-exposed videos are generated with a demosaicing procedure to obtain ground truth videos. These works pay attention to enhancing low-light images to satisfy the need for human visual perception.

Unpaired Datasets Several public datasets provide lots of under-exposed images used for subjective evaluations. VV[1] includes 24 images, part of which are normally exposed and other parts of severely under/over-exposed and provide the most challenging cases for low-light enhancement. LIME (Guo et al. 2017) contains 10 low-light images. NPE (Wang et al. 2013a) contains 85 low-light images downloaded from the Internet captured in 8 outdoor nature scenes. DICM (Lee et al. 2013a) contains 69 captured images from commercial digital cameras. MEF (Ma et al. 2015) contains 17 high-quality image sequences including natural scenarios, indoor and outdoor views and man-made architectures. (Loh and Chan 2019) developed the Exclusively Dark dataset including 7363 low-light images from very low-light environments to twilight (i.e 10 different conditions) annotated with 12 object classes using both image level categories and local object bounding boxes. (Hwang et al. 2015) introduced the KAIST Dataset, which contains various night-time traffic sequences for pedestrian detection. These works are designed to meet the need for machine visions, namely improving the performance of high-level vision tasks.

Other Attempts and Evaluation Criteria There are also important attempts on datasets in the related image processing tasks for multiple purposes, e.g. dehazing (Li et al. 2019), unconstrained conditions (Nada et al. 2018), *etc.* Previous metrics for low-light enhancement, including full-reference and no-reference metrics, are summarized in Table 2. In our work, we build a comprehensive dataset capturing both paired and unpaired images in low-light conditions to meet the need for both visual quality enhancement and face detection accuracy, and make an effort on benchmarking the existing methods with diverse metrics based on the dataset.

2.2 Low-Light Enhancement Approach

Based on the working mechanism and processed data type, we categorize the single-image low-light enhancement into seven categories: histogram equalization, dehazing, statistical model, Retinex model, deep-learning, compound degradation, and RAW. We will discuss the existing methods of these approaches in detail in the subsequent sections. A summary of previous works is given in Tables 3 and 4. A timeline is provided in Fig. 1.

Histogram Equalization Histogram equalization (HE) makes dark images visible by stretching the dynamic range of an image (Pizer et al. 1990) via manipulating the corresponding histogram. However, HE applies the adjustment globally, leads to undesirable local illumination and amplifying buried intensive noise. Later methods apply several kinds of constraints, e.g. mean intensity preservation (Ibrahim and Pik Kong 2007), noise robustness, white and black stretching (Arici et al. 2009), and a new distortion model (Lee et al. 2014a), to achieve the improved overall visual quality. To better adjust the histograms in a more finely-grained way,

[1] http://decsai.ugr.es/cvg/dbimagenes/

Table 2 Summary of previous evaluation metrics for low-light enhancement

Metrics	Reference	Measurement	Work
LOE (Wang et al. 2013a, 2019)	No	Lightness distortion	Guo et al. (2017), Ying et al. (2017a, c), Wang et al. (2019), Zhang et al. (2019), Tao et al. (2017), Lv et al. (2018)
Color difference (Wang and Zhang 2010)	No	Color distortion	Ying et al. (2017c)
NIQE (Mittal et al. 2013)	No	Natural preservation	Shen et al. (2017), Wang et al. (2019), Zhang et al. (2019), Jiang et al. (2019)
Discrete entropy (Ye et al. 2007)	No	Color richness and sharpness	Shen et al. (2017)
Angular error (Hordley and Finlayson 2004)	Full	Color Distortion	Shen et al. (2017)
VIF (Sheikh and Bovik 2006)	Full	Information fidelity	Ying et al. 2017a; Lv et al. 2018
PSNR	Full	Signal fidelity	Chen et al. (2018), Lore et al. (2017), Wang et al. (2019), Wang et al. (2019), Zhang et al. (2019), Cai et al. (2018), Lv et al. (2018), Ren et al. (2019)
SSIM (Wang et al. 2004)	Full	Signal and structure fidelity	Shen et al. (2017), Chen et al. (2018), Lore et al. (2017), Wang et al. (2019), Wang et al. (2019), Zhang et al. (2019), Lv et al. (2018), Ren et al. (2019)
FSIM (Zhang et al. 2011)	Full	Signal and structure fidelity	Cai et al. (2018)
TMQI (Yeganeh and Wang 2013)	Full	Signal and structure fidelity	Tao et al. (2017), Lv et al. (2018)
Average brightness (Chen et al. 2006)	Full	Brightness fidelity	Lv et al. (2018)

in Lee et al. (2013a), Nakai et al. (2013), the histogram equalization is applied to the pixel's difference adaptively. In some methods, side information, e.g. depth information (Lee et al. 2014b), is introduced to guide the pixel value transformation adaptively. In Ying et al. (2017b), Wu et al. (2017), the imaging and visual perception models are injected to guide the low-light image enhancement, e.g. camera response model (Ying et al. 2017b) to find the best exposure ratio and visual importance (Wu et al. 2017) to control the contrast gain. In general, with more side information and constraints, HE-based methods improve local adaptivity of the enhancement process. However, most methods are not flexible enough for visual property adjustment in local regions and lead to undesirable local appearances, e.g. under/over-exposure and amplified noise.

Dehazing Some methods (Li et al. 2015; Zhang et al. 2012; Dong et al. 2011) regard the inverted low-light images as haze images, and enhance visibility by applying dehazing. The dehazing result is inverted as the enhancement result. These methods also consider noise suppression. In Zhang et al. (2012), a joint-bilateral filter is applied after enhancement. In Li et al. (2015), adaptive BM3D denoising operations (Dabov et al. 2007) are conducted to separate the base and enhancement layers, and then adjust these two layers adaptively. These methods obtain reasonable results. However, a convincing physical explanation on their basic model is missing, and applying denoising as post-processing may lead to blurred details.

Statistical Model-Based Methods A wide range of methods depicts desirable properties of images with statistical models. They are carefully designed with expert domain knowledge. Some are based on the physical and statistical measures, such as interpixel relationship (Celik and Tjahjadi 2011), local contrast measure (Pierre et al. 2016), perceptual quality measure (Zhang et al. 2016). Some are designed based on mathematical processes, such as nonlinear diffusion filter (Liang et al. 2016), generalized Gaussian mixture model (Li et al. 2018). There are also methods built based on imaging or visual perception guided models (Ying et al. 2017c, a), such as camera response model and just noticeable difference (Chang and Jung 2016; Su and Jung 2017). These methods achieve good results in their focused aspects. However, they are not adaptive enough when coming across the cases out of assumed input ranges, such as the input images with intensive noise.

Retinex Theory Based Methods Retinex model is proposed as a human visual perception model (Land and McCann 1971) to compute visual appearance. The successive variational versions of Retinex model follow the layer decomposition paradigm, which is generally adopted in low-light image enhancement (Jobson et al. 1997b). The methods decompose images into two components: reflectance and illumination. Then, the enhanced results are obtained by further processing

Table 3 An overview of low-light enhancement methods (Part 1): including histogram equalization (HE), dehaze, Retinex, statistical model-based methods

Category	Methods	Variables/models	Main idea	Publication
HE	CLAHE	Contrast; local histogram; partitioned regions	The method divides the input into regions and performs the adaptive histogram equalization locally with contrast limitation, which reduces noise by partially suppressing local histogram equalization	Pizer et al. (1990) [CVBC]
	BPDHE	Smoothed histogram; Histogram partition	The mean intensity of the output image is kept to be almost the same to that of the input to prevent visual deterioration	Ibrahim and Pik Kong (2007) [TCE]
	WAHE	Contrast adjustment; Noise robustness; White/black stretching; Mean-brightness preservation	A general framework based on histogram equalization is presented to integrate contrast adjustment, noise robustness, white/black stretching and mean-brightness preservation	Arici et al. (2009) [TIP]
	BCCE	Brightness compensation distortion; Backlight -scaled image contrast	The work formulates an objective function consisting of contrast enhancement and a newly proposed distortion model to adjust the backlight-scaled image contrast	Lee et al. (2014a) [TCSVT]
	LDR	2D histogram; Tree -like gray-level differences	The image contrast is enhanced by amplifying the gray-level differences between adjacent pixels	Lee et al. (2013a) [TIP]
	DHECI	Intensity histogram; Saturation histogram	A differential gray-levels histogram equalization is designed for color images with two differential gray-level histograms, i.e. intensity gray-levels histogram and saturation gray-levels histogram	Nakai et al. (2013) [SITIS]
	DGACE	Depth; 2D histograms; Adaptive space-variant transform function	A novel contrast enhancement method utilizes 2D histograms to transform pixel values adaptively based on the depth information	Lee et al. (2014b) [ICIP]
	EFF	Weighting matrix; Camera response model; Best exposure ratio	The weighting matrix and camera response model are introduced to synthesize multi-exposure images with the best exposure ratio	Ying et al. (2017b) [CAIP]
	CAHE	Visual importance; Dark-pass filtered gradients	The method adaptively controls the contrast gain based on the potential visual importance of intensities and pixels	Wu et al. (2017) [ICIP]
Dehaze	Dehazing	Inverted video	The method first inverts an input video and then applies a dehazing approach on the inverted video	Dong et al. (2011) [ICME]
	ENR	Inverted image; Filter weighting	After enhancement, the joint bilateral filter is introduced to suppress noise	Li et al. (2015) [ICPR]
	SepDehaze	Base layer; Detail layer; Superpixel	The input image is decomposed into base layer and detail layer and then enhance them adaptively	Zhang et al. (2012) [ICIP]
	SSR	Single-scale Retinex; Chromaticity coordinates; Color restoration function	It defines a practical implementation of Retinex center and surround Retinex, and treats the reflectance as the final enhanced result	Jobson et al. (1997b) [TIP]
	MSR	Multi-scale Retinex; Chromaticity coordinates; Color restoration function	It creates the enhanced results by fusing different single-scale Retinex outputs	Jobson et al. (1997a) [TIP]

Table 3 continued

Category	Methods	Variables/models	Main idea	Publication
	AMSR	Single-scale Retinex; Stretched results; Weighting matrix	The weight of each single-scale retinex is adaptively computed based on the input image	Lee et al. (2013b) [SITIS]
Retinex	NPE	Lightness-order-error; a Bright-pass filter; Bi-log transform	A lightness-order-error measure accesses naturalness preservation and a bi-log transform is derived to make a balance between details and naturalness	Wang et al. (2013a) [TIP]
	VBR	Prior distribution; Parameter distribution	We utilize the Gibbs distributions as prior distributions for the reflectance and the illumination, and the gamma distributions for the model parameters to construct a hierarchical Bayesian model	Wang et al. (2014) [TIP]
	RIA	Reflectance; Illumination	A novel model without the logarithmic transform is built to well preserve edges	Fu et al. (2014) [ICASSP]
	SRIE	Revised total variation	A weighted variational model is proposed the original domain (instead of logarithmic domain) for better prior modeling	Fu et al. (2016) [CVPR]
	LIME	Structure prior	The initial illumination map is refined with an imposed structure prior	Guo et al. (2017) [TIP]
	MF	Luminance-improved and contrast-enhanced input; Weights	Two inputs are derived to represent luminance-improved and contrast-enhanced versions of the illumination using the sigmoid function and adaptive histogram equalization. Then, weights are inferred to produce an adjusted illumination in a multi-scale fashion	Fu et al. (2016) [SP]
	JIEP	Shape prior; Texture prior; Illumination prior	A novel model is proposed to preserve the structure information by shape prior, estimate the reflectance with fine details by texture prior, and capture the luminous source by illumination prior	Cai et al. (2017) [ICCV]
	RPCE	Luminance just-noticeable difference; Illumination weakening factor	We first estimate the illumination component by adaptive smoothing and get luminance just-noticeable difference using luminance adaptation. Then, an illumination weakening factor is calculated for detail enhancement	Xu and Jung (2017) [ICASSP]
	Robust	Local derivatives; Structure map; Texture map	Exponential filters are designed for the local derivatives modeling. Different exponents are chosen to extract structure and texture maps	Li et al. (2018) [TIP]
	JED	Illumination map; Reflectance map; Gradient constraint	Retinex decomposition is applied in a sequence, where a piece-wise smoothed illumination and a noise-suppressed reflectance are sequentially estimated	Ren et al. (2018) [ICASSP]
	STAR	Exponential filters; Local derivatives; Structure map; Texture Map	Exponential filters are designed for the local derivatives modeling. Different exponents are chosen to extract structure and texture maps	Xu et al. (2019) [arXiv]

Table 3 continued

Category	Methods	Variables/models	Main idea	Publication
Statistical model	CVC	2-D interpixel relationship histogram;	The work enhances the contrast of an input image using interpixel contextual information, a 2-D histogram to depict a mutual relationship between each pixel and its neighboring pixels	Celik and Tjahjadi (2011) [TIP]
	HPPCE	Local contrast measure; Discrete total variation	A variational model is introduced to adjust the average local contrast measure, preserve the hue and model the lateral inhibition. The control of the level of contrast can be tuned	Pierre et al. (2016) [ICIP]
	PDPF	Multi-exposed results	The video frame is adjusted via tentative tone mapping curves. Guided by some visual perception quality measures, the best exposed regions are integrated in a temporally consistent manner	Zhang et al. (2016) [TVCG]
	PCE	Gradient map; Just noticeable difference	The textural coefficient is inferred by gray level difference and just noticeable difference, filtered by a Gaussian kernel. Then, the optimal contrast tone mapping is obtained	Chang and Jung (2016) [VCIP]
	NDF	Nonlinear diffusion filtering; Texture suppression;	The illumination is estimated by a iterative nonlinear diffusion. Surround suppression is embedded in the conductance function to enhance the diffusive strength in textural areas of the image	Liang et al. (2016) [TIP]
	PLM	Environmental light; Light-scattering attenuation	The initial environmental light is estimated via a Gaussian surrounding function. Then, the environmental light and light-scattering attenuation rate are iteratively refined with the information loss constraint	Yu and Zhu (2019) [TCSVT]
	BIMEF	Weighting matrix; Camera response model; Best exposure ratio	The weighting matrix is extracted via illumination estimation. Then, camera response model is introduced to synthesize multi-exposure images and the best exposure ratio is found for each region	Ying et al. (2017a) [arXiv]
	CRM	Camera response model; Exposure ratio map	The method uses the inferred camera response model to adjust the pixel intensity to the desired exposure based on the estimated exposure ratio map	Ying et al. (2017c) [ICCVW]
	TSNS	Noise level function; Just noticeable difference	The method first performs noise aware contrast enhancement using noise level function and then utilizes a just noticeable difference model to suppress noise	Su and Jung (2017) [ICASSP]
	3GGMM	Generalized Gaussian mixture model	A three-component generalized Gaussian mixture model is used to fit the histogram of the illuminance image, and probabilistically characterize under-, normal-, and overexposures	Li et al. (2018) [ICIP]

Table 4 An overview of low-light enhancement methods (Part 2): including deep learning-based, compound degradation-targeted, RAW oriented methods

Methods	Category	Variables/Models	Main Idea	Publication
Deep learning	LLNet*	Stacked sparse denoising autoencoder	This work uses a class of deep neural networks, stacked sparse denoising autoencoder, to enhance natural low-light images	Lore et al. (2017) [PR]
	LLCNN	Multi-scale feature maps	A special module is designed to utilize multiscale feature maps and SSIM loss is used to train the model	Tao et al. (2017) [VCIP]
	MBLLEN	Feature extraction module; Enhancement module; Fusion module	The method extracts rich features up to different levels via feature extraction module, enhances the multi-level features respectively via enhancement module and obtains the final output by multi-branch fusion via fusion module	Lv et al. (2018) [BMVC]
	HybridNet	Content features; Edge features	The proposed network consists of two streams to simultaneously learn the global content and the salient structures of the clear image. A novel spatially variant RNN as an edge stream to compensate for edge details.	Ren et al. (2019) [TIP]
	EnlightenGAN	Attention map; Self-regularization	The method regularizes the unpaired training with the information extracted from the input low-light images, a global/local discriminator structure, a self-regularized perceptual loss, and attention mechanism	Jiang et al. (2019) [arXiv]
	LowLightGAN	Task-driven training set	This paper proposes a low-light enhancement method using an advanced generative adversarial network with spectral normalization and advanced loss functions as well as a task-driven training set	Kim et al. (2019) [ICIP]
	EEMEFN	Multi-exposure fusion mode; Edge enhancement module	The multi-exposure fusion model first produces several images with different illumination and then fuses these image together into a high-quality one. After that, The edge enhancement module further estimate and refine an edge map to generate the final enhanced image	Zhu et al. (2020) [AAAI]
	Zero-DCE	Light-enhancement curves; Iterations	The method trains a lightweight deep network to estimate pixel-wise and high-order curves for dynamic range adjustment of the input low-light image	Guo et al. (2020) [CVPR]
	DRBL	Coarse-to-fine band representations; Recompose band representations; Signal fidelity; Perceptual visual quality	The first stage of the network extracts a series of coarse-to-fine band representations. In the second stage, the model learns to recompose the band representation towards fitting perceptual visual quality of high-quality images	Yang et al. (2020) [CVPR]
	D&E	Frequency-based decomposition and enhancement	A novel network first learns to recover image objects in the low-frequency layer and then enhances high-frequency details	Xu et al. (2020) [CVPR]
Deep learning + Retinex	MSR-Net	Multi-scale logarithmic transform; Difference of convolution; Color restoration function	The relationship between multi-scale Retinex and feed-forward CNN is built and the surround functions in Retinex theory are formulated as convolutional layers	Shen et al. (2017) [arXiv]
	SICE	Low frequency luminance; High frequency detail	This work builds a dataset of low-contrast and good-contrast image pairs, making the discriminative learning of SICE enhancers possible	Cai et al. (2018) [TIP]

Table 4 continued

Methods	Category	Variables/Models	Main Idea	Publication
	RetinexNet	Illumination/Reflectance layers; Structure-aware smoothness; Multi-scale illumination adjustment	The network is trained with the assumed consistency between the reflectance of paired low/normal-light images, and the smoothness of illumination. Subsequent lightness enhancement is conducted on illumination and there is a denoising operation on reflectance	Wei et al. (2018) [BMVC]
	KinD	Illumination/Reflectance layers	This work builds a network to decomposes images into two components and adjust them adaptively	Zhang et al. (2019) [ACM MM]
	DeepUPE	Local/Global features Low-res illumination	A network for enhancing underexposed photos by estimating an image-to-illumination mapping is built and reconstruction, smoothness and color losses are applied	Wang et al. (2019) [CVPR]
	ProgRetinex	Pointwise convolutional neural networks; statistical regularities of ambient light and image noise	This paper proposes a progressive Retinex framework, where illumination and noise of low-light image are perceived in a mutually reinforced manner	Wang et al. (2019) [ACM MM]
	SemanticRetinex	Illumination layer; Reflectance layer; Semantic prior	Semantic segmentation, reflectance, and illumination are estimated from the input underexposed image. Reflectance is enhanced with the help of semantic information, and the reconstructed reflectance helps adjust the illumination	Fan et al. (2020) [ACM MM]
Compound degradation	RCE	Denoised results; Reliable weight; Intensity histogram; Complete matrix	The method first applies denoising and compute the reliability weight to categorize each pixel into noise-free or noisy pixels. Then , selective histogram equalization is applied. Finally, missing values of the noisy pixels are filled via a low-rank matrix completion	Lim et al. (2015) [ICIP]
	GLO	Graph Laplacian operator; Edge weight filtering	The enhancement method based on a graph Laplacian matrix enhances the high frequency details without amplifying additive noise. Then, a graph-based low-pass filtering approach is used to denoise edge weights in the graph	Liu et al. (2015) [ICASSP]
	SE	Superpixels; Noise-texture level; Base/Detail layer; Haze-related variables	The low-light image is segmented into superpixels and the noise/texture level of each superpixel is estimated. Based on the noise/texture level, a smooth base layer and detail layer are extracted and adaptively combined to get a noise-free and detail-preserved image. Finally, an adaptive enhancement parameter is to adjust the dark channel prior to enlarge contrast	Li et al. (2015) [TIP]
	PNM-WTV	Weighted total variation	A low light image denoising is built based on Poisson noise model and weighted total variation regularization	Yang et al. (2018) [ICIP]

Table 4 continued

Methods	Category	Variables/Models	Main Idea	Publication
	LED2Net	Illumination map	The illumination map is taken as the component for three tasks, i.e. atmospheric light estimation, transmission map estimation, and low-light enhancement. The model is trained simultaneously based on the retinex theory	Kim and Kwon (2019) [arXiv]
RAW+ Deep learning	SID	Amplification ratio	The work introduces a dataset of raw short-exposed low-light images and the corresponding long-exposure reference images. Furthermore, am end-to-end trainable pipeline for processing low-light images is designed	Chen et al. (2018) [CVPR]
	SDM	Filtered results; Siamese network	A new dataset including raw low-light videos is constructed, where the high-resolution raw data is captured at video rate and several under-exposed versions are captured at the same time. Based on the dataset, a siamese network is built and trained on static raw videos, and generalizes to handle videos of dynamic scenes in the testing phase	Chen et al. (2019) [ICCV]
	SMOID	U-Net	A novel optical system is developed to capture paired bright and dark videos. A fully convolutional network with 3D and 2D miscellaneous operations is built to learn the enhancement mapping	Jiang and Zheng (2019) [ICCV]
Application	Nighttime driving	Synthetic twilight set; Synthetic nighttime set; Pseudo/Refined labels	A curriculum framework is proposed to adapt semantic segmentation models from day to nigh progressively	Sakaridis et al. (2019) [ICCV] Sakaridis et al. (2020) [arXiv]
	Dark YOLO	Glue layers; Generative model	The pre-trained enhancement and detection models are merged with newly proposed glue layers and a generative model	Sasagawa and Nagahara (2020) [ECCV]
	Defog	High/low-frequency component; gray-scale/color images; consistency loss	A two-step method employs separate operations on the high/low-frequency component of the gray-scale and color images, respectively, with the consistency loss between the two-step outputs	Jiang and Zheng (2019) [ICCV]

∗ LLNet is first published by arXiv on Nov. 2015

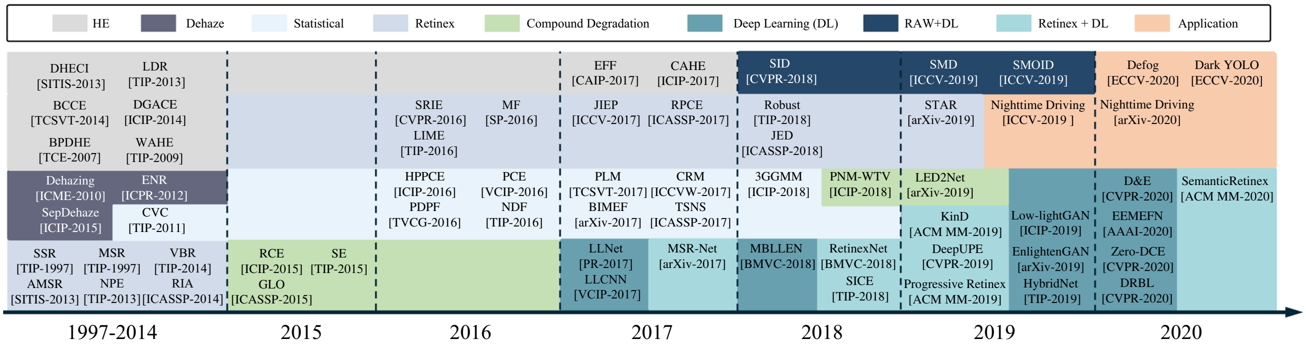


Fig. 1 Milestones of single-image low-light enhancement methods: histogram equalization, dehazing, statistical model, Retinex model, deep-learning (DL), RAW+DL, Retinex+DL, compound degradation, and related applications

and combining these two parts. To simultaneously suppress the noises and preserve high-frequency details, a series of methods built on Retinex theory (Land 1977) with diversified priors and constraints. Single-scale Retinex (Jobson et al. 1997b) defines a practical implementation of Retinex center and surrounding Retinex, and treats the reflectance as the final enhanced result. Multi-scale Retinex (Jobson et al. 1997a) creates enhanced results by fusing different single-scale Retinex outputs. Successive methods (Lee et al. 2013b; Wang et al. 2013b, 2014) increase the adaptivity of enhancement operations on the decomposed layers. In Lee et al. (2013b), the weight of each single-scale Retinex is adaptively computed based on the input image. Wang et al. (2013b) construct a bright-pass filter for Retinex decomposition, and try to preserve the naturalness while enhancing details in low-light images. In Wang et al. (2014), prior distributions of the reflectance and the illumination, as well as the parameters of the enhancement process, are jointly modeled with a hierarchical Bayesian model. Some methods explore the proper domain to apply the reconstruction prior. In Fu et al. (2014), a novel model without the logarithmic transform is built to well preserve edges. There are also methods focusing on exploiting more effective priors (Fu et al. 2016; Guo et al. 2017; Fu et al. 2016; Cai et al. 2017; Xu and Jung 2017; Xu et al. 2019) to regularize the enhancement of illumination and reflectance layers. Fu et al. (2016) propose an improved version by fusing different merits into a single one based on multiple derivatives of the estimated illumination. Guo et al. (2017) proposed to refine an initial illumination map with a structure aware prior. In Fu et al. (2016), a weighted variational model is proposed to impose better prior representation in the regularization terms. These methods consider less on the constraints on the reflectance, and the latent intensive noises in the low-light regions are usually amplified. Li et al. (2018) proposed to extend the traditional Retinex model to a robust one with an explicit noise term, and made the first attempt to estimate a noise map out of that model via an alternating direction minimization algorithm. Ren et al. (2018) also aimed to enhance low-light images based on that robust Retinex model, and developed a sequential algorithm to estimate a piecewise smoothed illumination and a noise-suppressed reflectance. These methods show impressive results in stretching the contrast of the image and removing noise in some cases. However, as the methods and the related priors are hand-crafted, they have poor adaptability and usually generate unpromising results when being applied to the large-scale testing data.

Deep-Learning Based Methods The era of deep-learning (DL) low-light enhancement starts in year 2017. After that, due to its distinguished performance and flexibility, this branch gradually becomes the mainstream. Lore et al. (2017) used a deep auto-encoder named Low-Light Net (LLNet) to perform contrast enhancement and denoising. In Shen et al. (2017), Tao et al. (2017), and Lv et al. (2018), the multi-scale features are injected into the multi-branch architecture to form better low-light enhancement results. In some of these works (Lore et al. 2017; Cai et al. 2018; Wang et al. 2019), efforts are put into creating paired low/normal-light datasets for network training. Diversified losses are enforced to regularizing the enhancement model training, such as, MSE (Lore et al. 2017), SSIM loss (Cai et al. 2018), and compound loss (Wang et al. 2019). In Shen et al. (2017), Wei et al. (2018); Wang et al. (2019), Retinex structure is fused into the design of effective deep networks, to absorb the advantages of both Retinex-based methods, *i.e.* good signal structure, and deep learning-based methods, *i.e.* the general useful priors extracted from the large-scale dataset. In Ren et al. (2019), layer decomposition and separative processing are introduced for better structure and detail modeling. In Jiang et al. (2019), Kim et al. (2019), the adversarial learning is introduced to capture the visual properties beyond the traditional metrics. Especially for EnlightenGAN (Jiang et al. 2019), unpaired learning is introduced to train a light enhancement model, which is the potential to get rid of paired dataset construction and address the domain shift problem between the training data and the practical applications. In general, with the powerful priors extracted from the large-scale data, deep learning methods achieve the general superiority in performance. Some traditional ideas are injected to guide the design of the deep networks, such as Retinex model and the layer separation.

Compound Degradation and RAW Enhancement Some works consider addressing the problem of low-light enhancement as well as its accompanying issues, such as denoising (Lim et al. 2015; Liu et al. 2015; Li et al. 2015; Yang et al. 2018) and dehazing (Kim and Kwon 2019). Some methods address the issue with a sequential architecture (Lim et al. 2015; Liu et al. 2015) while others achieve joint processing with a unified model (Li et al. 2015; Yang et al. 2018). In general, these methods can achieve good results in their assumed conditions, while a comprehensive model to capture all degradation and handle the corresponding degradation is still absent. Besides, there are works (Chen et al. 2018, 2019; Jiang and Zheng 2019) considering the application scenario to obtain enhanced images from raw images (un-processing images). The datasets of raw short-exposure low-light images and the corresponding raw long-exposure reference images are introduced and novel end-to-end trainable pipelines for processing low-light images/videos are designed. The attempt is meaningful while this direction expects more attention.

Related Applications There are also recent works focusing on the related applications in low-light conditions or at nighttime. In Sasagawa and Nagahara (2020), Loh and Chan (2019), the object detection problem in the low-light condition is explored. Loh et al. (2019) offered a large-scale collection consisting of 7363 low-light images with 12 object classes, annotated with both image-level classes and object-level bounding boxes. Yukihiro et al. (2020) proposed to merge the pre-trained enhancement model (from RAW image to RGB image) and pretrained detection model (from RGB image to bounding boxes) using newly proposed glue layers and a generative model, which can save the effort to create an new dataset (from RAW image to bounding boxes). In Dai and Gool (2018), Sakaridis et al. (2019, 2020), two datasets including unlabeled nighttime images, unlabeled twilight images with correspondences to their daytime versions, and nighttime images with pixel-level dense annotations are proposed to serve evaluation of semantic segmentation at night. A curriculum framework is proposed to adapt semantic segmentation models from day to nigh progressively. The cross-time-of-day correspondences are utilized to guide the label inference in the nighttime domains. In (2020), Yan et al. proposed a two-step method that employs separate operations on the high/low-frequency component of the gray-scale and color images, respectively, with the consistent loss between the two-step outputs.

Summary and Prospect We can obtain several interesting observations from the literature review:

- Retinex-based methods are the most widely adopted prior while in recent years since 2017, deep learning methods become the mainstream, which demonstrates the effectiveness of Retinex signal structure and data-driven priors extracted from the large-scale data.
- Statistical model-based methods are also a large group. However, the different methods within the same group also vary from each other. Their designs accompany much expert domain knowledge, which is not flexible and general to incorporate other widely used priors.
- Deep-learning methods are augmented with some traditional priors, such as Retinex structure and layer-specific priors, to achieve better enhancement performance.
- Adversarial learning is utilized to capture the visual properties beyond the traditional metrics to provide more visually pleasing results.
- Unsupervised or semi-supervised (unpaired) learning that benefits to getting rid of laborious paired dataset construction and address the domain shift problem between the training data and the practical applications are expected in the future works.

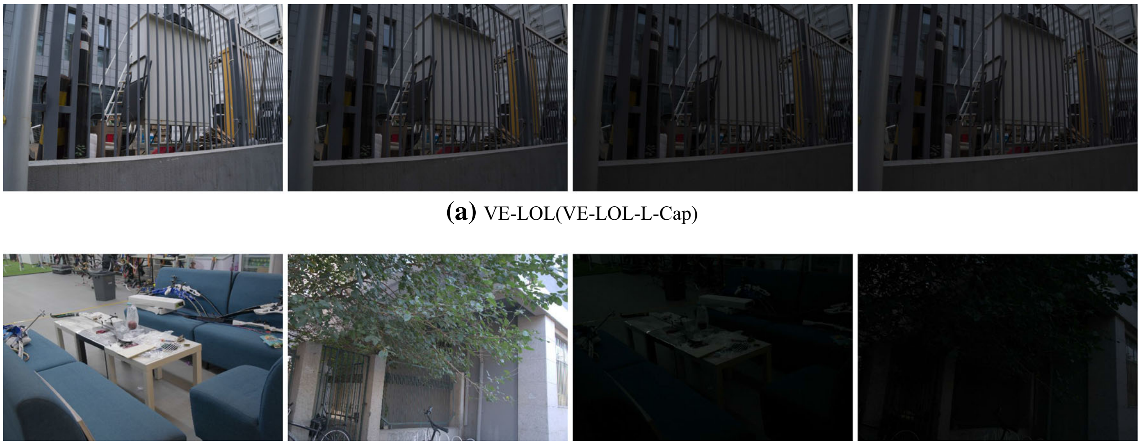
Despite the prosperity of low-light enhancement, there is a lack of extensive and systematic analysis of existing state-of-the-art low-light enhancement methods with comprehensive evaluation criteria. Therefore, in the following sections, we propose a novel dataset serving the purpose and apply extensive comparisons to show performance from the perspectives of human and machine visions.

3 A Large-Scale New Dataset: VE-LOL

We propose the Vision Enhancement in LOw Light conditions (VE-LOL) dataset, a novel large-scale dataset including both paired images, and unpaired images with annotations. It provides a wealth of materials to fairly evaluate and compare the performance of single-image low-light enhancement methods. A wide range of evaluation metrics, including no-reference, full-reference, and high-level feature metrics as well as the task-driven metric, *i.e.* face detection accuracy, are utilized in the evaluations.

The advantages of our proposed dataset can be summarized as follows.

- *Comprehensive Consideration*: VE-LOL supports evaluation for both low-level (with the subset VE-LOL-L) and high-level vision (with the subset VE-LOL-H).



(a) VE-LOL(VE-LOL-L-Cap)

(b) LOL

Fig. 2 Compared with LOL (bottom panel) consisting of paired low and normal-light images with a single under-exposure level, our proposed VE-LOL-L additionally includes paired low and normal-light images with different under-exposure levels at the same scene (top panel)

- *Reality*: VE-LOL contains real-captured paired images under both low-light and normal-light conditions, as well as the low-light face images with the corresponding annotations.
- *Diversity*: VE-LOL-L includes synthesized images with diversified backgrounds and a variety of objects.
- *Human-Relevant*: VE-LOL-H includes analysis resources related to human, *i.e.* annotated human face bounding boxes, which enables to evaluate existing methods from the perspective of machine vision and to develop joint enhancement and detection method.
- *Large Scale*: VE-LOL-H contains 10,940 images, whose scale is comparable to WIDER-FACE, the largest dataset captured in normal-light conditions and includes 32,203 images. Therefore, VE-LOL-H is so far the largest low-light detection dataset for high-level vision tasks. Beyond solely enabling testing evaluation as UFDD (Nada et al. 2018) does, VE-LOL supports fully supervised training, which might promote new directions or facilitate new methods in the related fields.

3.1 Dataset Overview

VE-LOL consists of two subsets: a paired one VE-LOL Low-Level Vision (VE-LOL-L) for training and evaluating low-level vision enhancement, and an unpaired one VE-LOL High-Level Vision (VE-LOL-H) for training low-light enhancement models and evaluating the effect of low-light enhancement models on high-level vision tasks, e.g. face detection.

VE-LOL-L includes 2500 paired images. Among them, 1000 pairs (VE-LOL-L-Syn) are synthesized from RAW images in RAISE dataset (Dang-Nguyen et al. 2015). The synthesis process follows a similar way to that of LOL dataset (Wei et al. 2018) considering both the low-light degradation process and natural image statistics. Differently, we additionally consider noise modeling at the RAW image level following (Brooks et al. 2019). The parameters of the noise model are estimated based on the Darmstadt Noise Dataset, which is captured by using four different cameras: Sony A7R, Olympus E-M10, Sony RX100 IV, and Huawei Nexus 6P. Therefore, the noise model in theory can adapt to a wide range of cameras. Hence, in our work, we directly use the default setting in Brooks et al. (2019) to synthesize our low-light noisy data. For this collection, we mainly hope to capture more diversified scenes and contents as well as more abundant illumination variation. The other 1500 images are real image pairs (VE-LOL-L-Cap). 500 of them are also captured in the same way as LOL dataset (Wei et al. 2018) while the other 1000 pairs are captured with different under-exposure levels. That is, for a given captured normal-light image, we also capture its low-light versions with different low-light exposure levels. The difference between our multiple under-exposed image pairs in VE-LOL and the pairs in LOL dataset (Wei et al. 2018) is visualized in Fig. 2. The multiple under-exposure levels make the contained degradation more diverse and provide more abundant resources to evaluate the effectiveness and robustness of the enhancement models. Note that, the images in the captured collection include real visual degradation in low-light conditions. Therefore, the whole VE-LOL-L collection (VE-LOL-L-Syn and VE-LOL-L-Cap) includes diversified scenes and contents, abundant illumination variations, and real low-light visual degradation (including intensive noise), which provides the desirable resources for evaluating in low-level visual quality.

Table 5 Summary of our VE-LOL dataset

Subset	#Image	Real/Synthetic	Paired	Annotations
VE-LOL-L-Syn	1000	Synthetic	Yes	No
VE-LOL-L-Cap	1500	Real	Yes	No
VE-LOL-H	10,940	Real	No	Yes

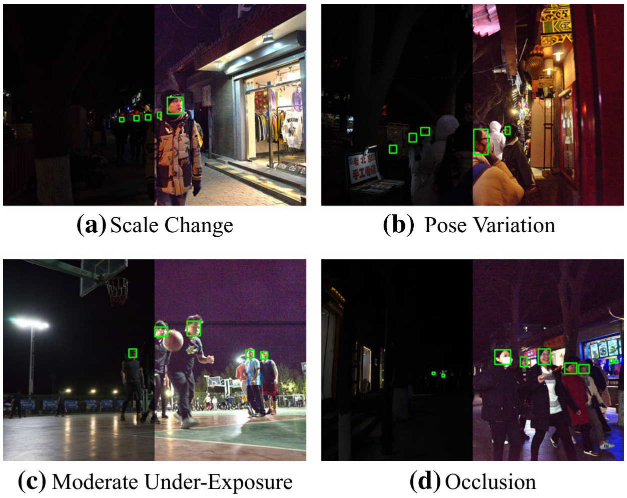

(a) Scale Change **(b)** Pose Variation **(c)** Moderate Under-Exposure **(d)** Occlusion

Fig. 3 Our proposed VE-LOL-H dataset for face detection has a high degree of variability in scale, pose, appearance, occlusion, and illumination. The left half of each image is the original one while the right half is enhanced by LIME for better visualization

Inspired by Anaya and Barbu (2018), we use a three-step shooting strategy to process the images in the captured collection of VE-LOL-L. For one scene, we first shoot two normal-light images N_1 and N_2. Then, we change the exposure time and ISO to capture a series of low-light images. Finally, we set the exposure time and ISO back to shoot another two normal-light images N_3 and N_4. Following (Anaya and Barbu 2018), the average of N_i (i=1,2,3,4) is treated as the ground-truth $G = \frac{1}{4}\sum_{i=1}^{4} N_i$. Then, we check whether there is object or camera movement. Specifically, the misalignment for these normal-light images is measured by $M = \frac{1}{4}\sum_{i=1}^{4} \mathrm{MSE}(N_i, G)$. If $M > 0.1$, we abandon the corresponding pair.

VE-LOL-H is composed of 10,940 images (6940 for training and validation, and 4000 for testing) taken in under-exposure conditions where human faces are manually annotated with bounding boxes. The training and evaluation sets include 53,619 annotated faces and the testing set includes 37,711 annotated faces. Table 5 presents a summary of our VE-LOL-L and VE-LOL-H dataset. Because previous works, e.g. LOL (Wei et al. 2018), have made great efforts in building datasets similar to VE-LOL-L, in the next part, we only focus on illustrating VE-LOL-H in detail, which is largely beyond considerations of previous works.

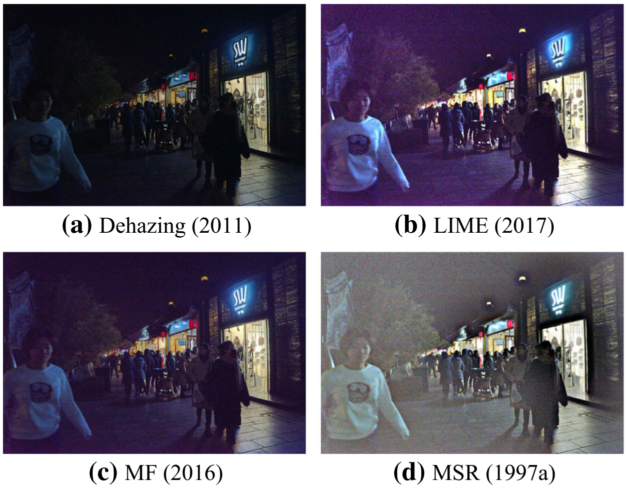

(a) Dehazing (2011) **(b)** LIME (2017) **(c)** MF (2016) **(d)** MSR (1997a)

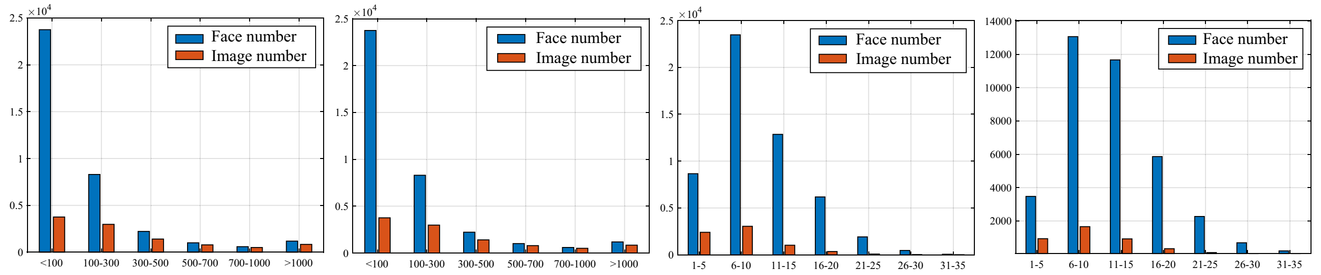
Fig. 4 Example images after low-light enhancement

3.2 VE-LOL-H for face detection in Low-Light Condition

Overview Beyond VE-LOL-L used for training and evaluating low-light enhancement methods from the perspective of low-level vision, we make endeavors to build a dataset captured in the low-light condition with high-level annotations, *i.e.* human face bounding boxes. Images in VE-LOL-H are captured in the under-exposed condition. Besides human faces, VE-LOL-H contains diversified objects, as shown in Fig. 3. Bounding boxes in images denote where faces are. They are manually selected using LabelImg Toolbox.[2] The bounding boxes provide resources to train and perform related evaluation experiments. Table 6 shows a comparison of VE-LOL-H to previous datasets, including both detection datasets in degraded conditions and face detection datasets.

Collection and Annotation This collection consists of images recorded from Sony α6000 and Sony α7 E-mount cameras with different capturing parameters on several busy streets around Beijing, where faces with various scales, poses, and appearances are captured. The resolution of these images is 1080 × 720 (down-sampled from 6K × 4K for maximum convenience). This collection includes 21,422 captured images in total. After filtering out those without sufficient information (lacking faces, too dark to see anything, *etc.*),

[2] https://github.com/tzutalin/labelImg



(a) FR in Train **(b)** FR in Test **(c)** FN in Train **(d)** FN in Test

Fig. 5 Face distribution in VE-LOL-H collections. Image number denotes the number of images belonging to a certain category. Face number denotes the summation number of faces belonging to a certain category. FR and FN denote face resolution (pixel^2) and number, respectively

Table 6 Comparison of VE-LOL-H to previous datasets

Dataset	#Image	#Object (Face)	#Train/Test	Conditions
ExDark (Loh and Chan 2019)	7363	23,710	4800/2563	Low light
UFDD (Nada et al. 2018)	6424	10,895	0/6424	Complex
MALF (Yang et al. 2015)	5250	11,931	250/5000	Normal
WIDER FACE (Yang et al. 2016)	32,303	393,703	12,921/16,152	Normal
VE-LOL-H	10,940	91,330	6940/4000	Low light

we select 10,940 images for human annotation. The bounding boxes are labeled for all the recognizable faces in our collection. We make the bounding tightly fit the forehead, chin, and cheek. If a face is occluded, we only label the exposed regions with skins. If the most of a face is occluded, we just ignore it. For this collection, we observe commonly seen degradations including poor image quality, under-exposure, and intensive noise in the results generated by enhancement methods, as shown in Fig. 4.

Data Distribution Each annotated image contains up to 34 human faces. The resolutions of faces in these images range from 1 × 2 to 335 × 296. The specific distributions of the ranges of the resolution and the number of faces are analyzed in Fig. 5. It is observed that, the resolution of most faces in our dataset is below 300 pixel^2 and the number of faces mostly falls into the range [1, 20]. Our face resolution is smaller than the object resolutions of the commonly used object detection datasets, e.g. MNIST (28 × 28) and CIFAR-10 (32 × 32), which poses new challenges jointly with low-light conditions in the community.

4 Algorithm Benchmarking

Based on the rich resources provided by VE-LOL, we evaluate representative state-of-the-art methods with diverse kinds of metrics.

4.1 Evaluation Protocols

From Full-Reference to No-Reference As denoted in Li et al. (2019), the full-reference signal and structure fidelity-driven PSNR and SSIM metrics are not enough to evaluate the visual quality of a series of image processing tasks, e.g. dehazing and low-light enhancement, because of their misalignment to human visual perception. Thus, based on our reviews of previous metrics in Table 2, we additionally select two full-reference metrics, VIF and angular error, to measure the information fidelity and color distortion of the enhanced results. Besides, we adopt several no-reference IQA metrics (i.e. LOE (Wang et al. 2013a), NIQE (Mittal et al. 2013), BRISQUE (Mittal et al. 2012), ENIQA (Chen et al. 2018), IL-NIQE (Zhang et al. 2015), HOSA (Xu et al. 2016), SSEQ (Liu et al. 2014), and BLIINDS-II (Saad et al. 2011)), to measure the lightness distortion, spatial domain statistics, and naturalness preservation.

From Low-Level to High-Level Feature Similarity Besides measuring the enhancement quality from the perspective of low-level signal structures, we also hope to measure whether the low-light enhancement methods well preserve the high-level semantics. Therefore, we use the perceptual metric (Johnson et al. 2016) to measure the similarity of the enhanced results and ground truth from the semantic view. Here, we use the first and fourth layers of VGG features for metric calculation, denotes as Perceptual_1 and Perceptual_4.

Task Driven Metric Additionally, we hope to measure the effect of low-light enhancement methods on the final per-

Table 7 The code sources of compared methods

Methods	Project Page
Multi-Scale Retinex (MSR), In Inverse Dehazing (Dehazing), Brightness Preserving Dynamic Histogram Equalization (BPDHE), Naturalness Preserved Enhancement (NPE), Multiple image Fusion (MF), Simultaneous Reflectance and Illumination Estimation (SRIE), Bio-Inspired Multi-Exposure Fusion (BIMEF)	https://eithub.com/baidut/BIMEF
Contextual and Variational Contrast enhancement (CVC), DHECI, Histogram Equalization (HE), Layered Difference Representation (LDR), Weighted Approximated Histogram Equalization (WAHE), Adaptive MultiScale Retinex (AMSR)	https://github.com/baidut/OpenCE
LLNet	https://github.com/kglore/llnetcolor
RetinexNet	https://github.com/weichen582/RetinexNet
Joint Enhancement and Denoising (JED)	https://github.com/tonghelen/JED-Method
Robust Retinex Model (Robust)	https://github.com/martinli0822/Low-light-image-enhancement
Single Image Contrast Enhancer (SICE)	https://github.com/csjcai/SICE
Kindling the Darkness (KinD)	https://github.com/zhangyhuaee/KinD
Deep Underexposed Photo Enhancement (Deep-UPE)	https://github.com/wangruixing/DeepUPE
Low-light IMage Enhancement (LIME)	https://sites.google.com/view/xjguo/lime
DSFD	https://github.com/yxlijun/DSFD.pytorch
PyramidBox	https://github.com/yxlijun/Pyramidbox.pytorch
SRN	https://github.com/ChiCheng123/SRN
SSH	https://github.com/dechunwang/SSH-pytorch
Faster-RCNN	https://github.com/hdjsjyl/face-faster-rcnn.pytorch

formance of high-level vision tasks. With the wealth of VE-LOL-H, we cascade low-light enhancement methods as a preprocessing of face detection, and use the face detection performance to measure the effectiveness of low-light enhancement from the machine vision view.

4.2 Baseline Enhancement and Detection Methods

We test state-of-the-art algorithms for light/contrast enhancement: Multi-Scale Retinex (MSR) (Jobson et al. 1997a), Inverse Dehazing (Dehazing) (Dong et al. 2011), Brightness Preserving Dynamic Histogram Equalization (BPDHE) (Ibrahim and Pik Kong 2007), Naturalness Preserved Enhancement (NPE) (Wang et al. 2013b), Low-light IMage Enhancement (LIME) (Guo et al. 2017), Multiple image Fusion (MF) (Fu et al. 2016), Simultaneous Reflectance and Illumination Estimation (SRIE) (Fu et al. 2016), Bio-Inspired Multi-Exposure Fusion (BIMEF) (Ying et al. 2017a), Joint Enhancement and Denoising (JED) (Ren et al. 2018), LLNet (Lore et al. 2017), RetinexNet (Wei et al. 2018), Contextual and Variational Contrast enhancement (CVC) (Celik and Tjahjadi 2011), DHECI (Nakai et al. 2013), Layered Difference Representation (LDR) (Lee et al. 2013a), Robust Retinex Model (Robust) (Li et al. 2018), Single Image Contrast Enhancer (SICE) (Cai et al. 2018), Weighted Approximated Histogram Equalization (WAHE) (Arici et al. 2009), Kindling the Darkness (KinD) (Zhang et al. 2019), Deep Underexposed Photo Enhancement (DeepUPE) (Wang et al. 2019). CVC, DHECI, LDR, WAHE, and BPDHE are the histogram equalization-based method. Inverse Dehazing conducts the dehazing operation in the inverse domain to enhance low-light images. Robust, MSR, NPE, LIME, MF, SRIE, and JED are Retinex model-based methods. BIMEF is the multiple hypothesis fusion-based method. Meanwhile, SICE, LLNet, RetinexNet, KinD, and DeepUPE are deep learning-based methods. For the task-driven metrics, we adopt the face detection results of Dual Shot Face Detectorf (DSFD) (Li et al. 2019), PyramidBox (Tang et al. 2018), Single Shot Scale-Invariant Face Detector (S^3FD) (Zhang et al. 2017), Single Stage Headless Face Detector (SSH) (Najibi et al. 2017), Selective Refinement Network (SRN) (Chi et al. 2018), and Faster RCNN (Jiang and Learned-Miller 2017). The code sources of compared methods are provided in Table 7. We do not retrain learning-based low-light enhancement methods and only adopt their pretrained models. We believe that, for the low-light enhancement task, the training dataset represents author's belief about what the results look like. Therefore, the datasets used in fact belong to parts of contributions of a work. For example, in RetinexNet, LLNet, SICE, DeepUPE, and KinD, the datasets are all listed as their contributions and belong to a part of their methods. With this in mind, we do not retrain learning-based methods and compare different methods with the authors provided pertained models.

4.3 Results on VE-LOL-L

Objective Evaluation We conduct the objective evaluation on the testing set of VE-LOL-L, including 100 synthetic images and 100 real-world paired images, respectively. The objective evaluation results are presented in Table 8. From the results, we can obtain several interesting observations:

- The results of different methods show different superiorities using different metrics.
- In general, deep learning-methods, KinD, SICE, and LLNet, achieve better performance in full-reference metrics, especially obtaining higher PSNR and SSIM results.
- For non-reference image quality assessment methods, deep-learning and Retinex-based methods, e.g. KinD, LLNet, JED, MSR, and SRIE, obtain superior results, which demonstrates the effectiveness of both the powers of data-driven learning and Retinex vision theory.
- For perceptual quality, deep-learning based methods, *i.e.* KinD and SICE win in semantic similarity metrics, *i.e.* Perceptual_1 and Perceptual_4 (Johnson et al. 2016), which shows that, driven by big data, enhancement methods are better at restoring perceptual properties of images.
- In general, KinD, LLNet, SICE, and JED are the best four methods as they enter the top three under most of the evaluation metrics.

Subjective Evaluation We also compare the subjective quality of different methods in Figs. 6, 7 (real) and 8 (synthesized). It is observed that, the visual results of different methods show different superiorities. For example, in Fig. 6, NPE, DHECI, MF, and LIME, generate visually good results. However, in Fig. 7, KinD achieves a much superior result than other methods. For real images, BPDHE, BIMEF, JED, SRIE, Robust, and DeepUPE's results are under-exposed. The results of DHECI, NPE, MF and LIME have rich saturation. The results of WAHE, SICE, LDR and CVC have a dull color distribution. The images are blurry and details are missing in the result of LLNet when zooming-in the results. RetinexNet generates similar results to the ground truths from the view of the overall signal distribution. However, the visual quality is not good. Except for Robust, KinD, DeepUPE, LLNet, and JED, other methods suffer from amplified intensive noise. For Fig. 8, most methods achieve more visually pleasing results. However, BPDHE, WAHE, LDR, CVC and DeepUPE still include under-exposed regions, especially for the left region of the bridge. More visual results will be presented in the supplementary material.

4.4 Running Time Evaluation

Table 9 reports the per-image running time of each method, averaged over the images (1080 × 720) in VE-LOL-H, on a machine with Intel(R) Xeon(TM) E5-1620 v3 3.50 GHz CPU and 16G RAM. All methods are implemented in MATLAB with CPU, except SICE by Caffe, Retinex-Net, KinD, DeepUPE by Tensorflow and LLNet by Theano with NVIDIA GeForce GTX 1080 Ti. It is observed that, most methods can finish processing an image within 2 seconds. BIMEF achieves the shortest running time. With the help of GPU, RetinexNet ranks fourth among all methods. It is worth mentioning that, all methods are still far away from the need for real-time processing (30 frames per second).

4.5 Results on VE-LOL-H

Detection Results with Low-Light Inputs Fig. 9a depicts the precision-recall curves of the baseline face detection methods for the VE-LOL-H collection, without enhancement. The baseline methods are trained on WIDER FACE (Yang et al. 2016), a large dataset captured in the well-exposed condition with large scale variations in diversified attributes and conditions. The results demonstrate that state-of-the-art methods cannot achieve desirable detection accuracies on VE-LOL-H. Some examples are illustrated in Fig. 10. The evidences may imply that though covering variations in poses, appearances, and scales, previous face datasets are not with sufficient training sources for images captured in the under-exposed condition, e.g. images in VE-LOL-H. The poor performance of state-of-the-art face detection methods then calls a novel dataset having the diversified distributions of face images in the under-exposed condition.

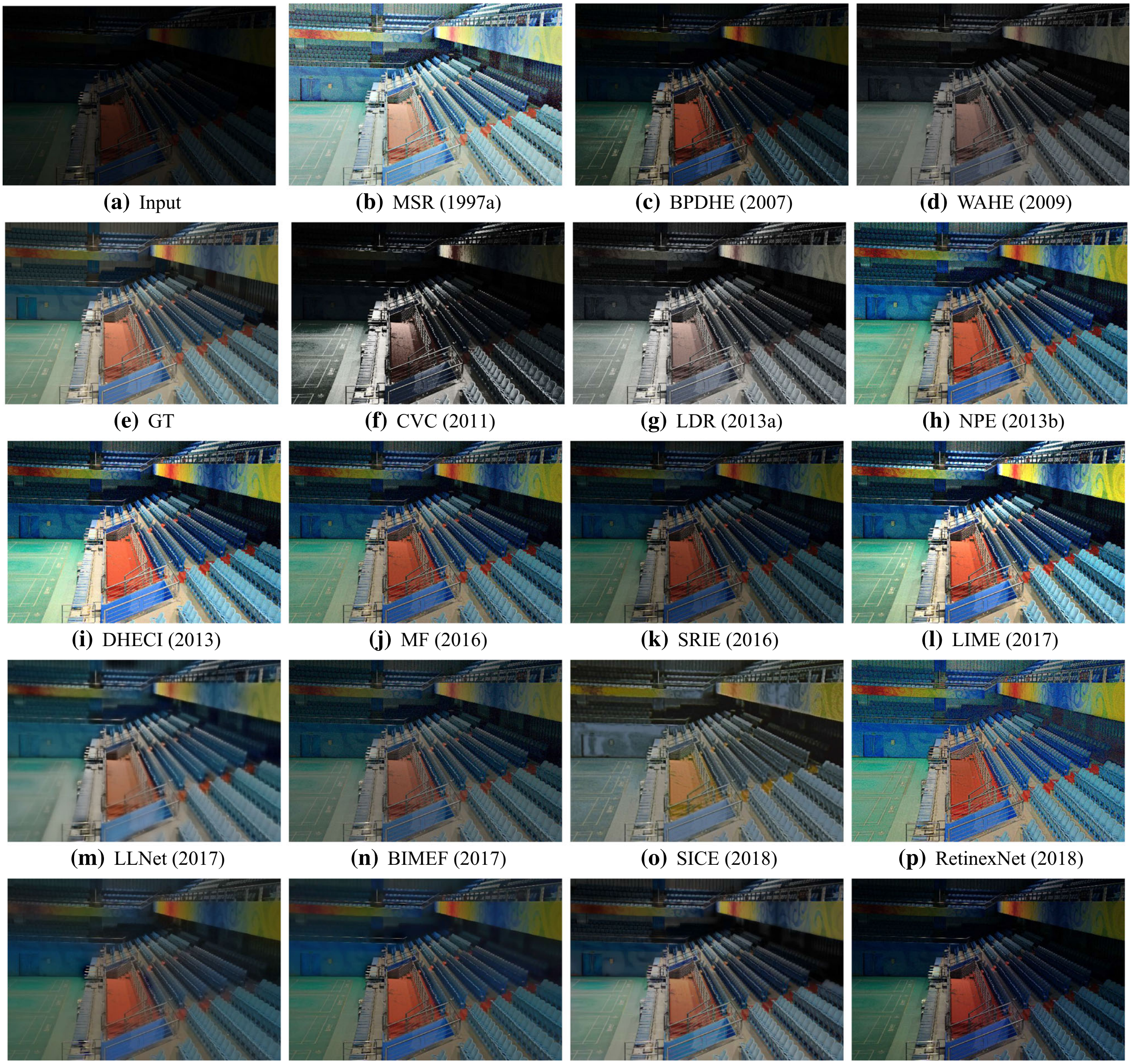
We further analyze the performance of face detectors on subsets of VE-LOL-H with different levels of difficulties. We split the testset based on two criteria: face scale and light condition. All faces in VE-LOL-H are divided into three levels based on the average size of the faces in an image: small ($<$100 pixel^2), medium (100 $\sim$ 300 pixel^2), and large ($>$300 pixel^2). Considering the facial illumination, all faces are also divided into three levels based on the average pixel value of the faces in an image: low illumination ($<$5), medium illumination (5 10), and high illumination ($>$10). The results are presented in Figs. 11 and 12. Clearly, the performance degrades when faces are small and in low illumination. DSFD achieves the best performance, with average precision rates greater than other detectors in all cases. The results suggest that the performance of current state-of-the-art face detectors will also degrade when the scales and light conditions change to some extreme conditions.

Detection Results with Enhanced Inputs Ideally, image restoration and enhancement algorithms should help object

Table 8 Average full-reference, no-reference and perceptual evaluations results of low-light enhancement results of different methods

Metrics	↑ or ↓	input	MSR	Dehazing	NPE	LIME	MF	SRIE	BIMEF	BPDHE	LLNET
PSNR	↑	10.24	11.95	15.38	15.38	14.07	16.26	13.66	15.95	12.75	*17.57*
SSIM	↑	0.2941	0.5493	0.5471	0.5670	0.5274	0.5998	0.5469	0.6386	0.4651	**0.7388**
VIF	↑	0.2937	*0.4525*	0.3772	0.4502	***0.4821***	0.4378	0.4351	0.4377	0.3802	0.3347
Angular_Error	↓	25.32	17.90	20.33	19.73	19.79	18.58	19.83	16.07	25.69	13.20
LOE	↓	0.00	1245.56	200.62	445.46	889.51	186.53	*140.11*	142.41	***18.11***	452.50
NIQE	↓	24.62	28.38	30.70	29.66	30.96	30.63	27.70	27.83	27.57	***18.97***
BRISQUE	↓	21.39	38.26	42.69	43.40	46.87	44.54	33.56	34.74	41.10	***20.58***
ENIQA	↓	0.1999	0.2405	0.1703	0.2287	0.1748	0.1962	0.1951	0.1708	**0.0600**	0.2116
ILNIQE	↓	52.88	37.72	46.07	47.79	50.85	47.75	47.06	45.98	44.48	*32.76*
HOSA	↓	37.18	54.70	44.95	47.12	47.01	47.58	38.31	42.80	40.09	38.18
SSEQ	↓	18.69	35.00	34.16	34.34	36.53	34.40	27.42	29.14	30.23	30.79
BLIINDS-II	↓	30.66	31.05	35.99	32.87	32.70	33.91	31.22	31.38	33.25	***12.70***
Perceptual_1	↓	20,522	25,595	15,392	16,877	28,213	13,695	13,213	11,781	18,514	11,138
Perceptual_4	↓	3481.91	4123.95	3904.14	3618.12	4744.79	3307.97	3111.39	*2971.76*	3899.04	3161.43
Metrics	↑ or ↓	JED	RetinexNet	CVC	DHECI	LDR	Robust	SICE	WAHE	KinD	DeepUPE
PSNR	↑	16.73	14.68	13.01	14.24	15.11	15.78	**18.06**	15.07	***18.42***	13.19
SSIM	↑	0.6817	0.5252	0.4469	0.5312	0.6114	0.6378	*0.7094*	0.6309	***0.7658***	0.4902
VIF	↑	0.3744	0.3482	0.3501	0.4299	**0.4681**	0.3750	0.3747	0.4377	0.4381	0.4222
Angular_Error	↓	*13.02*	21.32	28.83	19.58	19.33	16.06	**12.42**	17.08	***11.67***	22.70
LOE	↓	405.38	808.58	243.59	**15.60**	231.21	466.72	439.61	200.02	363.29	262.05
NIQE	↓	*23.07*	31.52	25.11	30.58	30.36	24.89	24.36	27.75	**21.38**	27.68
BRISQUE	↓	*28.51*	55.43	34.08	50.23	40.48	41.99	30.06	39.49	**23.30**	29.70
ENIQA	↓	0.1293	0.4049	*0.0659*	0.0699	0.0941	0.1837	0.1470	***0.0549***	0.1118	0.1906
ILNIQE	↓	35.53	47.27	36.08	51.63	36.42	46.32	33.85	36.13	**29.01**	48.99
HOSA	↓	36.53	55.47	37.99	47.11	44.02	43.22	***30.57***	43.18	**32.98**	*34.88*
SSEQ	↓	***18.39***	38.88	23.41	37.29	24.39	26.58	26.36	**22.29**	*23.19*	25.45
BLIINDS-II	↓	21.44	43.53	35.29	32.54	31.07	26.33	*20.96*	31.02	**20.62**	29.97
Perceptual_1	↓	*11,028*	20,333	26,335	25,581	14,901	13,211	**9871**	13,333	***9735***	14,108
Perceptual_4	↓	2998.40	4340.71	4752.17	4409.88	3289.93	3201.35	***2837.81***	3180.37	***2433.50***	3183.71

The value in bold italic, bold and italic to denote the first, second and third best results, respectively.



(a) Input **(b)** MSR (1997a) **(c)** BPDHE (2007) **(d)** WAHE (2009)

(e) GT **(f)** CVC (2011) **(g)** LDR (2013a) **(h)** NPE (2013b)

(i) DHECI (2013) **(j)** MF (2016) **(k)** SRIE (2016) **(l)** LIME (2017)

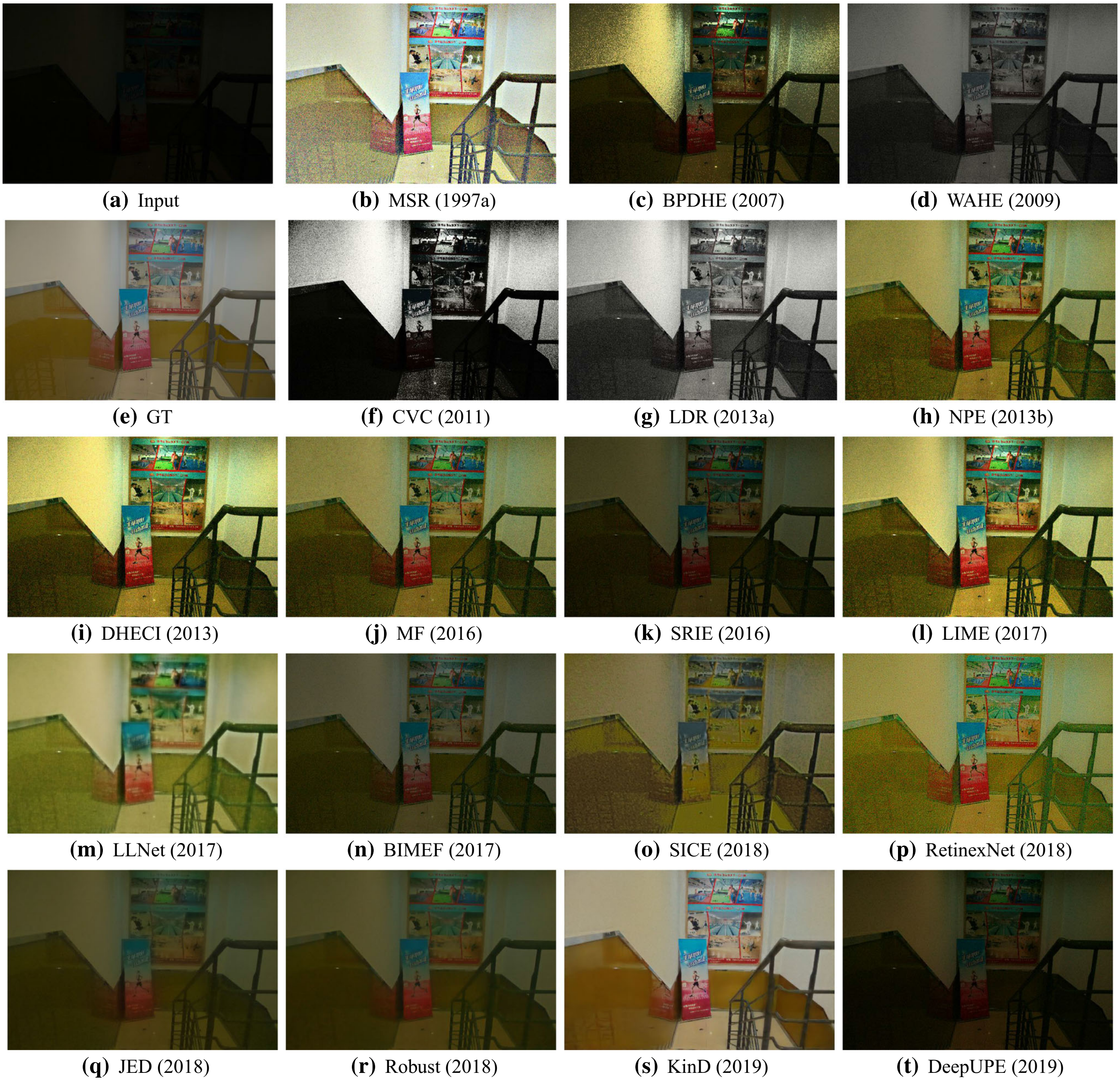
(m) LLNet (2017) **(n)** BIMEF (2017) **(o)** SICE (2018) **(p)** RetinexNet (2018)

(q) JED (2018) **(r)** Robust (2018) **(s)** KinD (2019) **(t)** DeepUPE (2019)

Fig. 6 Examples of enhanced results on a real low-light image from VE-LOL-L-Cap

Table 9 Comparison of average per-image running time (second) on images in VE-LOL-H (Resolution: 1080 × 720)

Method	MSR	Dehazing	BPDHE	NPE	LIME	MF	SRIE	BIMEF	JED
Running time (Second)	1.4161	0.9574	0.7506	8.1812	1.2454	1.5136	6.7943	0.1761	1.9646
Method	LLNet	RetinexNet	CVC	DHECI	LDR	Robust	SICE	WAHE	KinD
Running time (Second)	4.0213	0.4690	1.2660	25.3356	0.3602	44.6751	0.8075	1.4023	3.0031


(a) Input **(b)** MSR (1997a) **(c)** BPDHE (2007) **(d)** WAHE (2009)

(e) GT **(f)** CVC (2011) **(g)** LDR (2013a) **(h)** NPE (2013b)

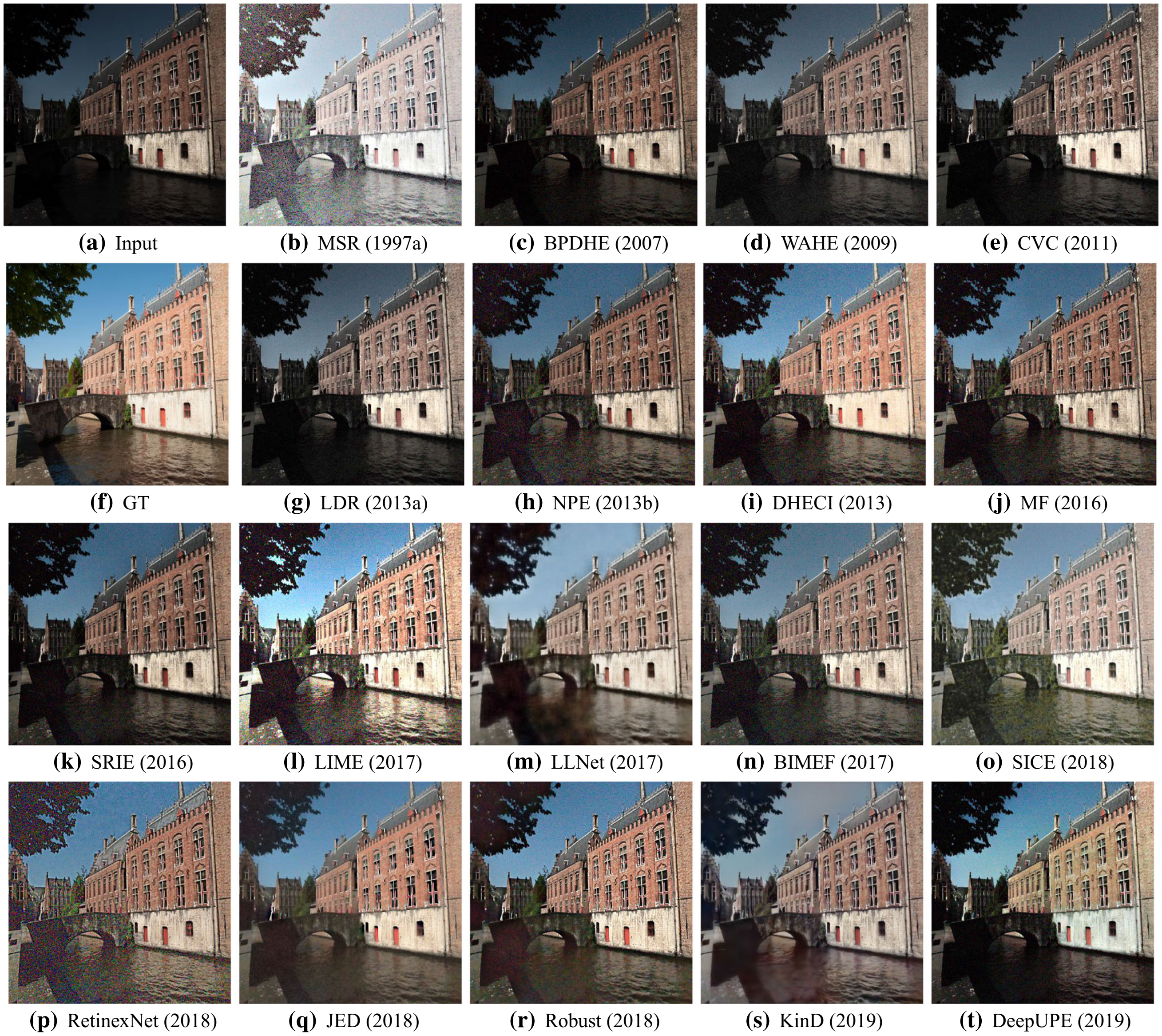
(i) DHECI (2013) **(j)** MF (2016) **(k)** SRIE (2016) **(l)** LIME (2017)

(m) LLNet (2017) **(n)** BIMEF (2017) **(o)** SICE (2018) **(p)** RetinexNet (2018)

(q) JED (2018) **(r)** Robust (2018) **(s)** KinD (2019) **(t)** DeepUPE (2019)

Fig. 7 Examples of enhanced results on a real low-light image from VE-LOL-L-Cap

detection by improving the quality of the degraded images and should not impair detection for good quality images. Following this intuition, we use enhancement methods to pre-process the VE-LOL-H dataset and two state-of-the-art face detection methods, *i.e.* DSFD, PyramidBox, and SRN, to detect the processed data. The visual quality of the enhanced images is better and the detectors indeed perform superiorly. As shown in Figs. 9b–d, and 13, the precision of the detectors notably increases compared to that of the data without enhancement in Fig. 9a. From Fig. 9b–d, it is observed that, BIMEF, MSR, and MF most significantly improve the performance evaluated using both DSFD and MSR. However, for PyramidBox, MF, Dehazing, and LIME achieve the most significant gains. Thus, in general, MF is a good method for machine vision. Our results also demonstrate that, a simple cascade of low-light enhancement and state-of-the-art face detectors present superior results to pretrained models. In Fig. 9, Proposed denotes our full version method. w/o Degrade, w/o MDL, w/o Skip, and w/o Fusion denote the versions without the half cyclic constraint, multiple detection losses, skip connection from the enhancement module to the detection module, and dual-path fusion, respectively. It is observed that, our method largely outperforms existing baselines, the joint results of existing low-light enhancement


(a) Input **(b)** MSR (1997a) **(c)** BPDHE (2007) **(d)** WAHE (2009) **(e)** CVC (2011)

(f) GT **(g)** LDR (2013a) **(h)** NPE (2013b) **(i)** DHECI (2013) **(j)** MF (2016)

(k) SRIE (2016) **(l)** LIME (2017) **(m)** LLNet (2017) **(n)** BIMEF (2017) **(o)** SICE (2018)

(p) RetinexNet (2018) **(q)** JED (2018) **(r)** Robust (2018) **(s)** KinD (2019) **(t)** DeepUPE (2019)

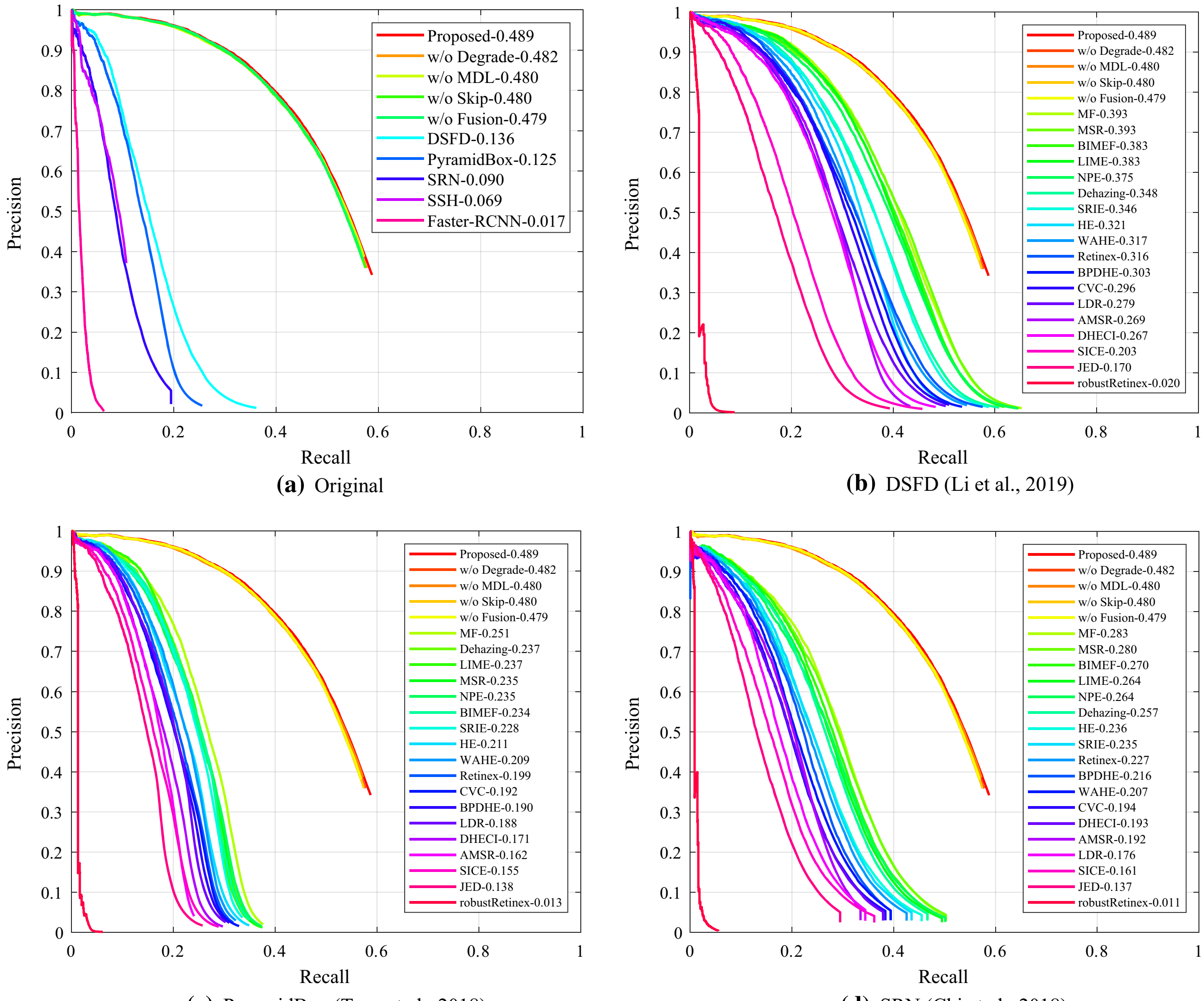
Fig. 8 Examples of enhanced results on a synthetic low-light image from VE-LOL-L-Syn

methods and face detectors. However, we also expect that, joint optimization of enhancement and detection may utilize more information in low-light images, which is explored in the following section.

4.6 Analysis of Noisy Labels in VE-LOL-H

In (2018), Wang et al. discussed performance changes caused by noisy labels in face recognition. However, the case changes when meeting face detection, as there are multiple labels belonging to a given image. In this section, we evaluate the performance of state-of-the-art face detectors trained with noisy labels in low-light conditions.

The noisy labels are collected from the annotation errors in the real human labeling process. We perform two rounds of annotations. A more careful check is employed in the second round based on the annotations in the first round. In the last round labeling, we conduct a very rigorous validation process on the labeled data. For the testing data, we carefully check every labeled image. In theory, until no wrong labeled face is found, the labeling process can be stopped. For the training data, we random select 100 images from every 1000 images. The labeling process of the 1000 images is stopped until less than 3 faces in these 100 images are found to be wrongly labeled. Over ten thousand bounding boxes were adjusted, resulting in an update from 83,885 bounding boxes to 91,330 bounding boxes on the whole VE-LOL-H. Such a clean and



(a) Original

(b) DSFD (Li et al., 2019)

(c) PyramidBox (Tang et al., 2018)

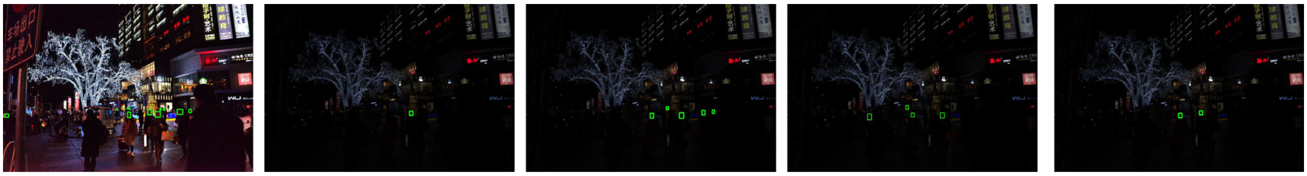
(d) SRN (Chi et al., 2018)

Fig. 9 Comparison of detection accuracies for different face scale in VE-LOL-H

large-scale dataset should enable an unbiased analysis of low-light enhancement methods by using face detection accuracy.

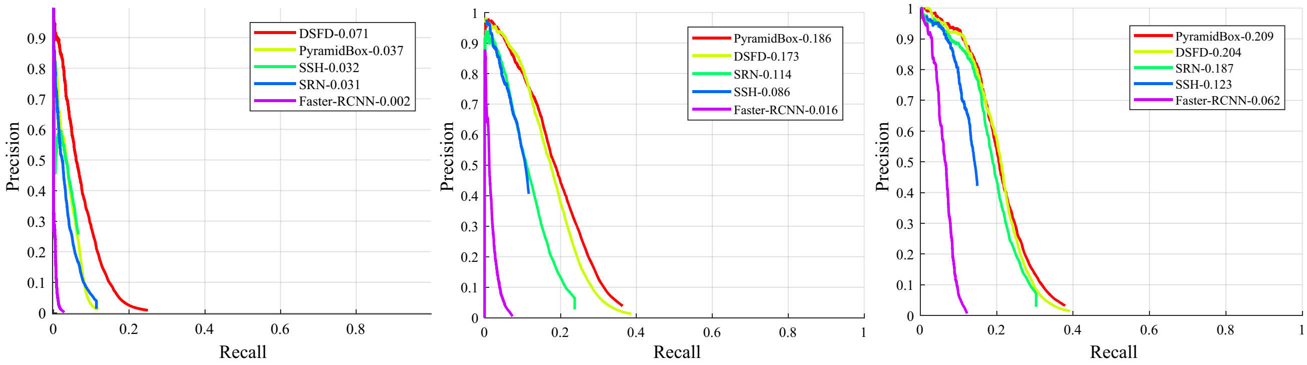
By comparing the labels of two rounds, noisy labels in the real labeling process are obtained. There are various kinds of label noise in the dataset. For example, bounding boxes might be shifted and scaled incorrectly, and sometimes annotations might be missing for a certain instance. It is costly to label a large-scale dataset with high accuracy because human annotators tend to make tiny mistakes from time to time. In low-light conditions, the labels are acquired by annotating on enhanced images. Preprocessors usually bring unpleasant artifacts that might disturb the judgment of human annotators. Additionally, the criteria for bounding boxes vary among different annotators. We expect the bounding boxes tightly fit the recognizable forehead, chin, and cheek, *etc*. However, for the occluded faces, some bounding boxes include hats and scarfs, while some recognizable faces were ignored.

In our experiment, we examine how noisy labels would degrade the performance of state-of-the-art detectors. Specifically, DSFD is selected as the baseline for the experiments. During the experiments, we control the noise ratio, the portion of images with noisy labels. The trained detector is further tested on a carefully refined test set. As shown in Fig. 14, noisy labels largely degrade the performance of detectors. As the portion of noisy labels grows, the performance of detectors drops severely.



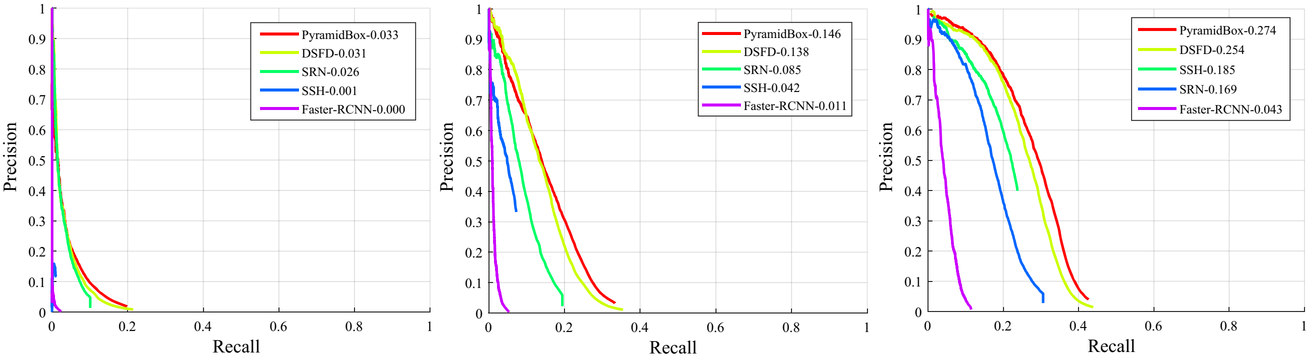
(a) GT **(b)** SSH (2017) **(c)** PyramidBox (2018) **(d)** SRN (2018) **(e)** DSFD (2019)

Fig. 10 Sample face detection results of pretrained models on the original low-light images of the proposed VE-LOL-H dataset. For better visualization, The ground truth is enlightened by LIME



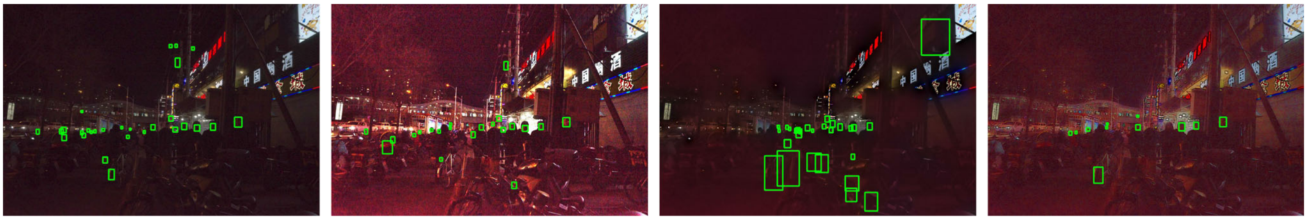
(a) Smallface **(b)** Mediumface **(c)** Largeface

Fig. 11 Comparison of detection accuracies for different face scale in VE-LOL-H



(a) Low illumination **(b)** Medium illumination **(c)** High illumination

Fig. 12 Comparison of detection accuracies for different face brightness for VE-LOL-H



(a) BIMEF (2017) **(b)** LIME (2017) **(c)** JED (2018) **(d)** RetinexNet (2018)

Fig. 13 Sample face detection results of an image by DSFD in the proposed VE-LOL-H dataset enhanced by different methods

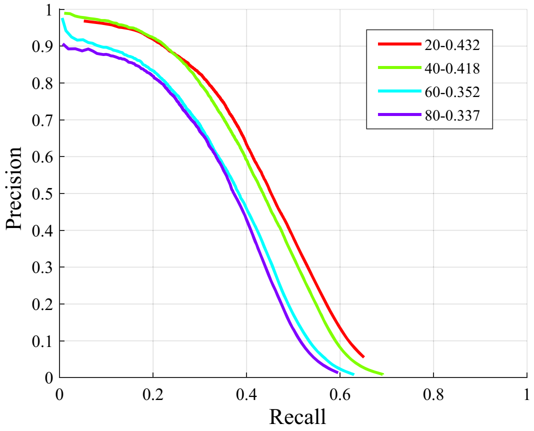


Fig. 14 Evaluation results of DSFD trained on different noise level of labels

5 ED-TwinsNet for joint Low-Light Enhancement and Face Detection

Based on the wealth of VE-LOL, we further explore joint low-light enhancement and face detection. An Enhancement and Detection Twins Network is proposed to improve the performance of face detection in the low-light condition. To fully utilize image priors in both paired and unpaired data, we introduce an additional half cyclic constraint in the unpaired case to better train a low-light enhancement module. After that, the low-light image enhancement module serves as a learnable preprocessing module of face detection. By connecting multi-scale features from different modules across enhancement and detection phases, robust and discriminative features are learned for face detection in low-light conditions. A dual-path fusion network is built in the end to take as input the intermediate features extracted from the original and enhanced images and fuse them adaptively for the final prediction of the bounding box locations.

5.1 Problem Formulation

Low-light face detection aims to accurately predict face bounding boxes ω (locations $\{[a_1, b_1], [a_2, b_2], ..., [a_T, b_T]\}$ and sizes $\{[h_1, w_1], [h_2, w_2], ..., [h_T, w_T]\}$) based on the image x captured in the low-light condition. Low-light face detection can be formulated as the joint optimization of low-light image enhancement $P(\cdot)$ and face detection $Q(\cdot)$ as follows,

$$\begin{aligned}\left[\hat{x}_1, \hat{x}_2, ..., \hat{x}_n, f_x\right] &= P(x),\\ \hat{\omega} &= Q(x, \hat{x}_1, \hat{x}_2, ..., \hat{x}_n, f_x),\end{aligned} \tag{1}$$

where f_x is the extracted feature from the enhancement stage. $\left\{\hat{x}_i\right\}_{i=1,...,n}$ are the intermediate enhanced results. In the second face detection stage, $Q(\cdot)$ takes as the input the original input, intermediate enhanced results, and the extracted feature at the first stage and generates $\hat{\omega}$, which aims to predict ω. The face detection performance is usually measured with IoU (intersection over union) given an overlap threshold (usually 0.5).

5.2 Motivations

To address the problem in Eq. (1), we design the model architecture with the following motivations:

– *Utilization of information at both original and enhanced exposure levels* After the enhancement process, dark details might be revealed. However, the process may generate artifacts and over-exposed details. Therefore, we hope that, the detection stage can exploit both information of the original input and enhanced results. In our work, we introduce the *dual path fusion* module that takes both the extracted features from the original image and those from the enhanced image as the input to utilize the information of the well-exposed regions from two inputs selectively.
– *Exploiting both paired and unpaired data* Paired data provides effective priors to guide the restoration of pixel-level structure. While unpaired data provides useful cues to infer face locations with high-level semantics. It is beneficial in theory to exploit both kinds of data to infer the enhanced results and benefit the successive detection performance. Therefore, in our work, we create a half cyclic constrained low-light enhancement method, where paired data is only used to train the enhancement module and the unpaired data is utilized to train both the enhancement and degradation modules. The joint utilization of two kinds of data boosts the enhancement capacity of the model.
– *Task correlation at both input and feature ends* The two stages, low-light image enhancement and face detection, should be considered jointly and they should have communication as much as possible. Therefore, we consider feed-forwarding the features of the enhancement module at different levels to the detection model to make the two-stage models tie together and the features of the enhancement module benefit the detection model.

In summary, our detection module as shown in Fig. 15 takes as input the original feature and the enhanced feature (*dual-path fusion*), and our enhancement module trained with both paired and unpaired data (*half cyclic constrained low-light enhancement*). Furthermore, to create abundant connections between the enhancement and detection modules, the extracted features at the enhancement module are extracted to facilitate the detection of different grain sizes (*feature extractor and skip connections*). The motivations and

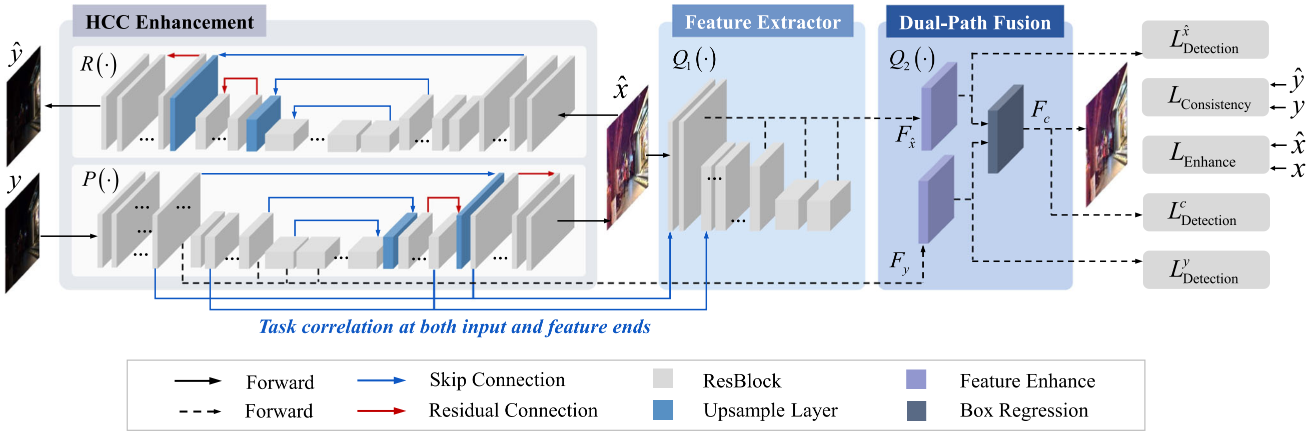


Fig. 15 The proposed Enhancement and Detection Twins Network (EDTNet) for joint low-light enhancement and face detection. The features extracted by the enhancement module are fed into the same level of the detection module. Thus, these features are interwined and unitedly learn useful information across two phases for face detection in low-light conditions. HCC Enhancement enables *exploiting both paired and unpaired data*, while Dual-path fusion helps *utilize of information at both original and enhanced exposure levels*”

Table 10 Summary of our motivations and modules

Module	Motivation	Method
HCC enhancement	Utilization of information at both original and enhanced exposure levels	Both the extracted features from the original image and those from the enhanced image are taken as the input.
Feature extractor and Skip connection	Exploiting both paired and unpaired data	Paired data is only used to train the enhancement module and the unpaired data is utilized to train both the enhancement and degradation modules.
Dual-path fusion	Task correlation at both input and feature ends	Features of the enhancement module are feed-forwarded at different levels to the detection model.

the corresponding model design are summarized in Table 10 and Fig. 15.

5.3 Model Architecture

Enhancement and Detection Twins Network (EDT-Net) The whole architecture of our method is shown in Fig. 15. It jointly optimizes the enhancement and detection phases by tightly connecting their intermediate features together at different levels. It consists of three main parts: a Half Cyclic Constrained low-light enhancement module (denoted as $P(\cdot)$ and $R(\cdot)$), a Feature Extraction Module (denoted as $Q_1(\cdot)$), and Dual-Path Fusion Module (denoted as $Q_2(\cdot)$). In the first part, besides the enhancement module $P(\cdot)$ usually trained with paired images, for unpaired low-light images, we also train a degradation module $R(\cdot)$ to further project the enhanced results back into low-light images, which confirms the signal and information fidelity of the enhanced results. In the second part, we first extract multi-scale features from the previously enhanced results via $Q_1(\cdot)$. Note that, features from different levels across the enhancement and detection phases are connected. As a result, robust and discriminative features are learned to boost the final predictions. In the last part $Q_2(\cdot)$, we feed-forward the extracted features of the original low-light images and the enhanced results at the same time to fully make use of their hidden potentials to facilitate face detection.

Half Cyclic Constrained (HCC) Low-Light Enhancement Module HCC Low-Light Enhancement Module aims to excavate the image priors of natural images and those about the mapping from low-light images to normal-light ones. The enhancement module $P(\cdot)$ consists of an encoder and a decoder. The encoder learns to extract features for both detection and low-light enhancement, and the decoder learns a mapping from feature space to enhanced images. Skip connections are used between encoder and decoder for detail reconstruction. In order to fully exploit the wealthy information of natural images, we use the images in both VE-LOL-L and VE-LOL-H for training. When using the paired images in VE-LOL-L, we can use the ground truth x to directly constrain the model training. When using the unpaired images in VE-LOL-H, the adversarial loss and consistency loss play a part in guiding the model learning. The loss to train the enhancement model $P(\cdot)$ is defined as follows,

$$\begin{aligned} L_{\text{Enhance}} &= \gamma\left(\left\|\hat{x}-x\right\| - \alpha\text{SSIM}\left(\hat{x},x\right)\right) + L_{\text{Adv}}(\hat{x},x), \\ L_{\text{Adv}}(\hat{x},x) &= -logD(x) - log\left(1-D(\hat{x})\right), \\ \hat{x} &= P(y), \end{aligned} \tag{2}$$

where y is the low-light input image, x is the ground truth (if available), and $\hat{x}$ is the enhanced result of y. $\gamma = 0$ denotes training with images in VE-LOL-H. $\gamma = 1$ denotes training with images in VE-LOL-L. $L_{\text{Adv}}(\cdot)$ is the adversarial loss (Goodfellow et al. 2014). D is the discriminator that is optimized to distinguish between x and $\hat{x}$. α is a weighting parameter, which is set to 0.5. For the unpaired low-light images in VE-LOL-H, we also hope to exploit their potential image priors. Therefore, we introduce a half cyclic constraint. That is, after enhancement, we use a learned degradation module $R(\cdot)$ to project $\hat{x}$ back into a low-light image $\hat{y}$ and use an L_1 loss to enforce the consistency between $\hat{y}$ and y as follows,

$$L_{\text{Consistency}} = \left\|\hat{y}-y\right\|, \tag{3}$$

$$\hat{y} = R(\hat{x}), \tag{4}$$

where y is regarded as the ground truth here and this loss confirms that the enhancement process not only improves the visual quality but keeps the information fidelity as well. Additionally, since many previous enhancement methods aim to improve human perception instead of machine vision, critical features for detection might be distorted when the light condition is adjusted. In order to solve the aforementioned problems, during the training phase, HCC is jointly trained with some Feature Enhance layers which are similar to those in DSFD (Li et al. 2019). We make use of the multi-task learning strategy, and optimize the modules for both detection and enhancement tasks. Specifically, we define L_{HCC} as

$$L_{\text{HCC}} = \lambda_1 L^{y}_{\text{Detection}} + L_{\text{Enhance}} + L_{\text{Consistency}}, \tag{5}$$

where λ_1 is a weighting parameter, set to 1 by default. $L^{y}_{\text{Detection}}$ is the detection loss using the feature of the encoder of the enhancement network extracted from y.

Feature Extraction Module The Feature Extraction Module adopts the same structure as the encoder in HCC. We feed-forward the previously enhanced result $\hat{x}$ into this module and extract robust features for face detection. Since the down-sampling layers usually lead to losing essential low-level information, skip connections are also adopted in order to share information between the first two stages. By bringing in some features from multiple layers in encoder and decoder, the Feature Extraction process and the previous Enhancement Module learn jointly to extract robust and discriminative features. The learnt features aggregate both low-level and high-level information, which is beneficial for the final prediction of face bounding boxes.

Dual-Path Fusion Module This design originates from the fact that the enhancement operation will magnify visual information but at the same time inevitably remove some structure details due to blurring, over-exposure or signal distortion. Thus, it cannot guarantee to preserve the desirable information for face detection. From the previous sub-modules, we obtain features F_y extracted from the original image y and $F_{\hat{x}}$ from its enhanced result $\hat{x}$. Then, we concatenate them together into a combined feature F_c , which is further fed into a box regression network to predict the bounding box results constrained by multiple detection losses (MDL) as follows,

$$L_{\text{Detection}} = L^{y}_{\text{Detection}} + L^{\hat{x}}_{\text{Detection}} + \lambda_2 L^{\text{c}}_{\text{Detection}}, \tag{6}$$

where the three terms denote the detection losses of the paths taking F_y, $F_{\hat{x}}$, and F_c as input, respectively. λ_2 is a weighting parameter, which is set to 1 by default. Each detection loss includes both Softmax loss over face and background, and smooth L_1 loss between the parameterizations of the predicted boxes and ground-truth box ones, as DSFD (Li et al. 2019) does. MDL is designed to serve as a regularization term and forces both branches to make their own efforts for better detection, facilitating to utilize more information at the feature level.

5.4 Evaluations

5.4.1 Implementation Details

Network Details Our HCC low-light enhancement network takes a gradual down and up-sampling structure. Residual blocks are cascaded to transform features and enrich representational information at each scale. The backbone of dual-path fusion network adopts a similar structure to DSFD (Li et al. 2019), using the front end of the VGG network (Simonyan and Zisserman 2014) as a coarse Feature Extraction Module, connected with extra feature refinement module, feature fusion module, and box regression module. The dual-path fusion network is initialized with the weights pretrained on WIDER FACE training set (Yang et al. 2016). All the other layers are initialized by "Xavier" method (Glorot and Bengio 2010).

Optimization In our training, we first pretrain the HCC low-light enhancement network with both the images in VE-LOL-L and VE-LOL-H. Then, we train the Feature

Table 11 The mAP scores of different methods

Method	Mean average precision (mAP) (%)
Pretrained DSFD	13.6
Finetuned DSFD	44.3
MF + pretrained DSFD	39.3
MF + finetuned DSFD	46.8
EnlightenGAN + finetuned DSFD	35.8
DeepUPE + finetuned DSFD	32.3
KinD + finetuned DSFD	29.6
SICE + finetuned DSFD	27.0
Proposed w/o half cyclic constraint (w/o degrade)	48.2
Proposed w/o skip connection (w/o Skip)	48.0
Proposed w/o dual-path fusion (w/o Fusion)	48.0
Proposed w/o multiple detection losses (w/o MDL)	47.9
Proposed	48.9

Extraction Module with images from VE-LOL-H. Finally, we finetune the whole network with images from VE-LOL-H. In the pretraining phase, we use RMSprop optimizer (Tieleman and Hinton 2012) and set the learning rate to 0.00001. We allow at most 20 epochs in the pretraining phase. After that, we use SGD optimizer to fine-tune our dual-path fusion network and set the learning rate, momentum and weight decay to 0.0001, 0.9, and 0.0005, respectively. The fine-tune phase is for at most 5 epochs. The gradient in the paths related to F_c is only allowed to train its corresponding box regression module and is forbidden to back-propagate to the front-end Feature Extractor Module. We implement the whole framework using the PyTorch library (Paszke et al. 2017).

Data Augmentation We adopt RandomCrop, HorizontalFlip, RandomSizedBBoxSafeCrop from the Albumentation Library (Buslaev et al. 2018) to prevent over-fitting and construct a more robust model. The first two augmentation strategies are applied with a probability 0.5, while the last one is applied with 0.3 possibility. During the RandomCrop process, input images are cropped into patches of a size 640 × 640. RandomSizedBBoxSafeCrop is similar to random cropping and then rescaling patches to 640 × 640. However, it will make sure there exists at least one bounding box in the resulted image patch. Bounding box labels are also cropped and rescaled correspondingly. A further filtering process is also adopted, to make sure that only those patches with bounding box labels inside are further fed in to the detection modules, while other patches are used to train the HCC module only.

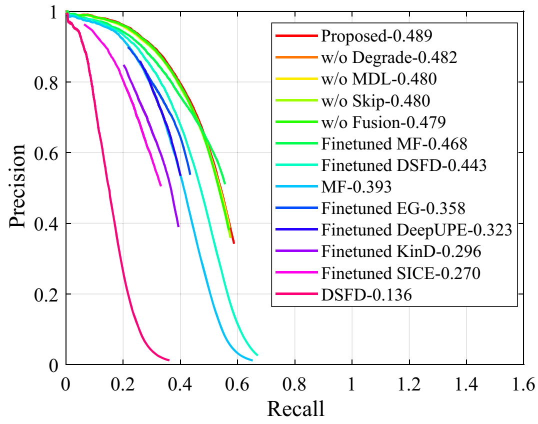


Fig. 16 Evaluation results of different algorithms on the proposed VE-LOL dataset

Inference During inference, all those paths related to F_y and $F_{\hat{x}}$ are ignored and we only use the path related to F_c to detect faces in the low-light condition. Finally, non-maximum suppression is applied with 0.3 Jaccard overlap (Tan et al. 2005) and we use the top 750 confident bounding boxes per image as the final results.

5.4.2 Experimental Results

We evaluate our joint low-light enhancement and face detection module on our VE-LOL-H. We compare with the finetuned results of DSFD taking the input as the pre-processing results by EnlightenGAN, DeepUPE, KinD and SICE. We also conduct more ablation studies to analyze the effect of half cyclic constraint, skip connection from the enhancement module to the detection module, and dual-path fusion. The results are shown in Table 11 and Fig. 16. As shown in Table 11, our method achieves superior performance to previous well-trained baselines. Directly finetuning DSFD or incorporating a low-light enhancement method, i.e. MF, SICE, KinD, largely boost the performance. Multiple detection losses, dual-path fusion, the skip connection from the enhancement module to the detection module, and half cyclic constraint will also benefit the final performance gently. The precision-recall curve is provided in Fig. 16. It is clearly demonstrated that, after finetuning or retraining on VE-LOL-L, the models improve the performance a lot, such as the proposed one, finetuned DSFD and the version without MDL.

6 Conclusion and Lessons

In this paper, we systematically evaluate the state-of-the-arts single-image low-light enhancement. First, a large-scale low-light image dataset has been presented. The dataset highlights captured both paired low and normal-light images and unpaired low-light images with face annotations. Then, with rich resources of the dataset, different methods are evaluated with diverse testing criteria. Last but not least, to handle the challenge of joint enhancement and detection, we make a preliminary attempt and acquire better performance with the power of image prior modeling and dual-path fusion architecture. From our work, there are many insights we learn from it:

– There is no method achieving overwhelming superiority on all metrics. Deep learning-based methods tend to perform well on fidelity-driven metrics. Retinex-based models achieve better results on other metrics.
– An off-line low-light enhancement may largely improve the low-light face detection accuracy with the model pretrained on face images captured in the normal-light condition. However, the superiority of low-light enhancement methods combined with the face detection methods is also dependent on the latter.
– Putting the low-light enhancement method as a learnable preprocessing module sometime may deteriorate the performance of face detection.
– Half cycle constraint provides an effective tool to build the enhancement module making use of the unpaired data in the target domain.
– By connecting features from different levels across enhancement and detection phases, the two phases are jointly optimized to learn robust and discriminative features for the final prediction of the bounding boxes.
– A successful solution to joint enhancement and detection is to feed-forward different levels of both input and enhanced results to the model, and fuse their features together for the final face detection.

Although our attempts are preliminary, we hope to inspire the community and attract researchers to this problem. Our benchmark results reveal state-of-the-art performance and the limitations in various aspects, and inspire new future directions, e.g. perception-guided low-light enhancement, real-time enhancement, and more excellent and comprehensive metrics, as well as superior low-light enhancement algorithms that benefit both human perception and machine vision.

Acknowledgements This work was supported by the National Key Research and Development Program of China under Grant 2018AAA0102702, the Fundamental Research Funds for the Central Universities, and the National Natural Science Foundation of China under contract No. 61772043.

References

Abdullah-Al-Wadud, M., Kabir, M. H., Dewan, M. A. A., & Chae, O. (2007). A dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics*, *53*(2), 593–600.

Anaya, J., & Barbu, A. (2018). Renoir: A dataset for real low-light image noise reduction. *Journal of Visual Communication and Image Representation*, *51*, 144–154.

Arici, T., Dikbas, S., & Altunbasak, Y. (2009). A histogram modification framework and its application for image contrast enhancement. *IEEE Transactions on Image Processing*, *18*(9), 1921–1935.

Brooks, T., Mildenhall, B., Xue, T., Chen, J., Sharlet, D., & Barron, J. T. (2019). Unprocessing images for learned raw denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 11028–11037.

Buslaev, A., Parinov, A., Khvedchenya, E., Iglovikov, V. I., & Kalinin, A. A. (2018). Albumentations: Fast and flexible image augmentations. arXiv preprint arXiv:180906839

Cai, B., Xu, X., Guo, K., Jia, K., Hu, B., & Tao, D. (2017). A joint intrinsic-extrinsic prior model for retinex. In *Proceedings of the IEEE international conference on computer vision*, pp. 4020–4029.

Cai, J., Gu, S., & Zhang, L. (2018). Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, *27*(4), 2049–2062.

Celik, T., & Tjahjadi, T. (2011). Contextual and variational contrast enhancement. *IEEE Transactions on Image Processing*, *20*(12), 3431–3441.

Chang, Y., & Jung, C. (2016). Perceptual contrast enhancement of dark images based on textural coefficients. In *Proceedings of IEEE visual communication and image processing*, pp. 1–4.

Chen, C., Chen, Q., Xu, J., & Koltun, V. (2018). Learning to see in the dark. In *Proceedings of IEEE international conference on computer vision and pattern recognition*, pp. 3291–3300.

Chen, C., Chen, Q., Do, M., & Koltun, V. (2019). Seeing motion in the dark. In *Proceedings of IEEE international conference on computer vision*, pp. 3184–3193.

Chen, X., Zhang, Q., Lin, M., Yang, G., & He, C. (2018). No-reference color image quality assessment: From entropy to perceptual quality. arXiv preprint arXiv:181210695

Chen, Z., Abidi, B. R., Page, D. L., & Abidi, M. A. (2006). Gray-level grouping (glg): An automatic method for optimized image contrast enhancement—part II: the variations. *IEEE Transactions on Image Processing*, *15*(8), 2303–2314.

Chi, C., Zhang, S., Xing, J., Lei, Z., Li, S. Z,, & Zou, X. (2018). Selective refinement network for high performance face detection. arXiv preprint arXiv:180902693

Dabov, K., Foi, A., Katkovnik, V., & Egiazarian, K. (2007). Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, *16*(8), 2080–2095. https://doi.org/10.1109/TIP.2007.901238.

Dai, D., & Gool, L. V. (2018). Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *International conference on intelligent transportation systems*, pp. 3819–3824.

Dang-Nguyen, D. T., Pasquini, C., Conotter, V., & Boato, G. (2015). Raise: A raw images dataset for digital image forensics. In *Proceedings of ACM multimedia systems conference*, pp. 219–224.

Dong, X., Wang, G., Pang, Y., Li, W., Wen, J., Meng, W., Lu, Y. (2011). Fast efficient algorithm for enhancement of low lighting video.

In *Proceedings of IEEE international conference multimedia and expo*, pp 1–6.

Fan, M., Wang, W., Yang, W., & Liu, J. (2020). Integrating semantic segmentation and retinex model for low-light image enhancement. In *ACM transactions on multimedia*, pp. 2317–2325.

Fu, X., Sun, Y., LiWang, M., Huang, Y., Zhang, X., & Ding, X. (2014). A novel retinex based approach for image enhancement with illumination adjustment. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing*, pp. 1190–1194.

Fu, X., Zeng, D., Huang, Y., Liao, Y., Ding, X., & Paisley, J. (2016). A fusion-based enhancing method for weakly illuminated images. *Signal Processing*, *129*, 82–96.

Fu, X., Zeng,,D., Huang, Y., Zhang, X., Ding, X. (2016). A weighted variational model for simultaneous reflectance and illumination estimation. In *Proceedings of IEEE international conference on computer vision and pattern recognition*, pp. 2782–2790.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *AISTATS, JMLR*, *9*, 249–256.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courvillem, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680.

Guo, C.G., Li, C., Guo, J., Loy, C.C., Hou, J., Kwong, S., Cong, R. (2020). Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of IEEE international conference on computer vision and pattern recognition*, pp. 1780–1789.

Guo, X., Li, Y., & Ling, H. (2017). Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, *26*(2), 982–993.

Hordley, S. D., & Finlayson, G. D. (2004). Re-evaluating colour constancy algorithms. In *Proceedings of IEEE international conference on pattern recognition*, *1*, 76–79.

Hwang, S., Park, J., Kim, N., Choi, Y., & So Kweon, I. (2015). Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of IEEE international conference on computer vision and pattern recognition*, pp. 1037–1045.

Ibrahim, H., & Pik Kong, N. S. (2007). Brightness preserving dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics*, *53*(4), 1752–1758.

Jiang, H., & Learned-Miller, E. G. (2017). Face detection with the faster R-CNN. In *IEEE international conference on automatic face and gesture recognition*, pp. 650–657.

Jiang, H., & Zheng, Y. (2019). Learning to see moving objects in the dark. In *Proceedings of IEEE international conference on computer vision*, pp. 7323–7332.

Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., & Wang, Z. (2019). EnlightenGAN: Deep light enhancement without paired supervision. arXiv e-prints arXiv:1906.06972

Jobson, D. J., Rahman, Z., & Woodell, G. A. (1997a). A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image Processing*, *6*(7), 965–976.

Jobson, D. J., Rahman, Z., & Woodell, G. A. (1997b). Properties and performance of a center/surround retinex. *IEEE Transactions on Image Processing*, *6*(3), 451–462.

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of IEEE European conference on computer vision*.

Kim, G., & Kwon, J. (2019). LED2Net: Deep illumination-aware Dehazing with low-light and detail enhancement. arXiv e-prints arXiv:1906.05119

Kim, G., Kwon, D., & Kwon, J. (2019). Low-lightgan: Low-light enhancement via advanced generative adversarial network with task-driven training. In *Proceedings of IEEE international conference on image processing*, pp. 2811–2815.

Land, E., & McCann, J. (1971) Lightness and retinex theory. *Journal of the Optical Society of America*, pp. 1–11.

Land, E. H. (1977). The retinex theory of color vision. *Scientific American*, pp. 108–128.

Lee, C., Lee, C., & Kim, C. S. (2013a). Contrast enhancement based on layered difference representation of 2D histograms. *IEEE Transactions on Image Processing*, *22*(12), 5372–5384.

Lee, C., Kim, J., Lee, C., & Kim, C. (2014a). Optimized brightness compensation and contrast enhancement for transmissive liquid crystal displays. *IEEE Transactions on Circuits and Systems for Video Technology*, *24*(4), 576–590.

Lee, C. H., Shih, J. L., Lien, C.C., & Han, C. C. (2013b). Adaptive multiscale retinex for image contrast enhancement. In *Signal-image technology and internet-based systems (SITIS), 2013 international conference on, IEEE*, pp. 43–50.

Lee, J., Lee, C., Sim, J., & Kim, C. (2014b). Depth-guided adaptive contrast enhancement using 2D histograms. In *Proceedings of IEEE international conference on image processing*, pp. 4527–4531.

Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., et al. (2019). Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, *28*(1), 492–505.

Li, J., Wang, Y., Wang, C., Tai, Y., Qian, J., Yang, J., Wang, C., Li, J., & Huang, F. (2019). Dsfd: Dual shot face detector. In *Proceedings of IEEE international conference on computer vision and pattern recognition*.

Li, L., Wang, R., Wang, W., & Gao, W. (2015). A low-light image enhancement method for both denoising and contrast enlarging. In *Proceedings of IEEE international conference on image processing*, pp. 3730–3734.

Li, M., Liu, J., Yang, W., Sun, X., & Guo, Z. (2018). Structure-revealing low-light image enhancement via robust retinex model. *IEEE Transactions on Image Processing*, *27*(6), 2828–2841.

Li, M., Wu, X., Liu, J., & GUo, Z. (2018). Restoration of unevenly illuminated images. In *Proceedings of IEEE international conference on image processing*, pp. 1118–1122.

Liang, Z., Liu, W., & Yao, R. (2016). Contrast enhancement by nonlinear diffusion filtering. *IEEE Transactions on Image Processing*, *25*(2), 673–686.

Lim, J., Kim, J., Sim, J., & Kim, C. (2015). Robust contrast enhancement of noisy low-light images: Denoising-enhancement-completion. In *Proceedings of IEEE international conference on image processing*, pp. 4131–4135.

Liu, L., Liu, B., Huang, H., & Bovik, A. C. (2014). No-reference image quality assessment based on spatial and spectral entropies. *Signal Processing: Image Communication*, *29*(8), 856–863.

Liu, X., Cheung, G., Wu, X. (2015). Joint denoising and contrast enhancement of images using graph laplacian operator. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing*, pp. 2274–2278.

Loh, Y. P., & Chan, C. S. (2019). Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*.

Lore, K. G., Akintayo, A., & Sarkar, S. (2017). Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, *61*, 650–662.

Lv, F., Lu, F., Wu, J., & Lim, C. (2018). Mbllen: Low-light image/video enhancement using CNNs. In *Proceedings of British machine vision conference*.

Ma, K., Zeng, K., & Wang, Z. (2015). Perceptual quality assessment for multi-exposure image fusion. *IEEE Transactions on Image Processing*, *24*(11), 3345–3356.

Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In

Proceedings of IEEE international conference on computer vision, *2*, 416–423.

Mittal, A., Moorthy, A. K., & Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, *21*(12), 4695–4708.

Mittal, A., Soundararajan, R., & Bovik, A. C. (2013). Making a completely blind image quality analyzer. *IEEE Signal Processing Letters*, *20*, 209–212.

Nada, H., Sindagi, V.A., Zhang, H., & Patel, V. M. (2018). Pushing the limits of unconstrained face detection: A challenge dataset and baseline results. In *IEEE international conference on biometrics: Theory, applications, and systems*.

Najibi, M., Samangouei, P., Chellappa, R., & Davis, L. S. (2017). Ssh: Single stage headless face detector. In *Proceedings of IEEE international conference on computer vision*, pp. 4885–4894.

Nakai, K., Hoshi, Y., & Taguchi, A. (2013). Color image contrast enhacement method based on differential intensity/saturation graylevels histograms. In *International symposium on intelligent signal processing and communications systems*, pp. 445–449.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in PyTorch. In *NIPS autodiff workshop*.

Pierre, F., Aujol, J., Bugeau, A., Steidl, G., & Ta, V. (2016). Hue-preserving perceptual contrast enhancement. In *Proceedings of IEEE international conference on image processing*, pp. 4067–4071.

Pizer, S. M., Johnston, R. E., Ericksen, J. P., Yankaskas, B. C., & Muller, K. E. (1990). Contrast-limited adaptive histogram equalization: Speed and effectiveness. In *Proceedings of conference on visualization in biomedical computing*, pp. 337–345.

Ren, W., Liu, S., Ma, L., Xu, Q., Xu, X., Cao, X., et al. (2019). Low-light image enhancement via a deep hybrid network. *IEEE Transactions on Image Processing*, *28*(9), 4364–4375.

Ren, X., Li, M., Cheng, W. H., & Liu, J. (2018). Joint enhancement and denoising method via sequential decomposition. In *IEEE international symposium on circuits and systems*.

Saad, M. A., Bovik, A. C., & Charrier, C. (2011). Dct statistics model-based blind image quality assessment. In *IEEE international conference on image processing* (pp. 3093–3096). IEEE.

Sakaridis, C., Dai, D., & Van Gool, L. (2019). Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pp. 7373–7382.

Sakaridis, C., Dai, D., Van Gool, L. (2020). Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. arXiv e-prints arXiv:2005.14553

Sasagawa, Y., & Nagahara, H. (2020). Yolo in the dark-domain adaptation method for merging multiple models. In *Proceedings of IEEE European conference on computer vision*, pp. 345–359.

Schaefer, G., & Stich, M. (2004). Ucid: An uncompressed colour image database. In *Storage and retrieval methods and applications for multimedia, proceedings of SPIE*, *5307*, 472–480.

Sheikh, H. R., & Bovik, A. C. (2006). Image information and visual quality. *IEEE Transactions on Image Processing*, *15*(2), 430–444.

Shen, L., Yue, Z., Feng, F., Chen, Q., Liu, S., & Ma, J. (2017). MSR-net: Low-light image enhancement using deep convolutional network. ArXiv e-prints.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. In *Proceedings of international conference on learning representations*.

Su, H., & Jung, C. (2017). Low light image enhancement based on two-step noise suppression. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing*, pp. 1977–1981.

Tan, P. N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining* (1st ed.). Boston, MA: Addison-Wesley Longman Publishing Co., Inc.

Tang, X., Du, D. K., He, Z., & Liu, J. (2018). Pyramidbox: A context-assisted single shot face detector. In *Proceedings of IEEE European conference on computer vision*.

Tao, L., Zhu, C., Xiang, G., Li, Y., Jia, H., & Xie, X. (2017). Llcnn: A convolutional neural network for low-light image enhancement. In *Proceedings of IEEE visual communication and image processing*, pp. 1–4.

Tieleman, T., & Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. Tech. rep.

Vonikakis, V., Chrysostomou, D., Kouskouridas, R., & Gasteratos, A. (2013). A biologically inspired scale-space for illumination invariant feature detection. *Measurement Science and Technology*.

Vonikakis, V., Kouskouridas, R., & Gasteratos, A. (2018). On the evaluation of illumination compensation algorithms. *Multimedia Tools and Appllications*, *77*(8), 9211–9231.

Wang, F., Chen, L., Li, C., Huang, S., Chen, Y., Qian, C., & Change Loy, C. (2018). The devil of face recognition is in the noise. In *Proceedings of IEEE European conference on computer vision*, pp. 765–780.

Wang, L., Xiao, L., Liu, H., & Wei, Z. (2014). Variational bayesian method for retinex. *IEEE Transactions on Image Processing*, *23*(8), 3381–3396.

Wang, R., Zhang, Q., Fu, C. W., Shen, X., Zheng, W. S., & Jia, J. (2019). Underexposed photo enhancement using deep illumination estimation. In *Proceedings of IEEE international conference on computer vision and pattern recognition*.

Wang, S., Zheng, J., Hu, H. M., & Li, B. (2013a). Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE Transactions on Image Processing*, *22*(9), 3538–3548.

Wang, S., Zheng, J., Hu, H. M., & Li, B. (2013b). Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE Transactions on Image Processing*, *22*(9), 3538–3548.

Wang, X., & Zhang, D. (2010). An optimized tongue image color correction scheme. *IEEE Transactions on Information Technology in Biomedicine*, *14*(6), 1355–1364.

Wang, Y., Cao, Y., Zha, Z.J., Zhang, J., Xiong, Z., Zhang, W., Wu, F. (2019). Progressive retinex: Mutually reinforced illumination-noise perception network for low light image enhancement. In *ACM transactions on multimedia*.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.

Wei, C., Wang, W., Yang, W., & Liu, J. (2018). Deep retinex decomposition for low-light enhancement. In *British machine vision conference*.

Wu, X., Liu, X., Hiramatsu, K., & Kashino, K. (2017). Contrast-accumulated histogram equalization for image enhancement. In *2017 IEEE international conference on image processing (ICIP),* pp. 3190–3194.

Xu, J., Ye, P., Li, Q., Du, H., Liu, Y., & Doermann, D. (2016). Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing*, *25*(9), 4444–4457.

Xu, J., Hou, Y., Ren, D., Liu, L., Zhu, F., Yu, M., Wang, H., & Shao, L. (2019). STAR: A structure and texture aware retinex model. arXiv e-prints arXiv:1906.06690

Xu, K., & Jung, C. (2017). Retinex-based perceptual contrast enhancement in images using luminance adaptation. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing*, pp. 1363–1367.

Xu, K., Yang, X., Yin, B., & Lau, R. W. (2020). Learning to restore low-light images via decomposition-and-enhancement. In *Proceedings* *of IEEE international conference on computer vision and pattern recognition*.

Yan, W., Tan, R. T., Dai, D. (2020). Nighttime defogging using high-low frequency decomposition and grayscale-color networks. In *Proceedings of IEEE European conference on computer vision*, pp. 473–488.

Yang, B., Yan, J., Lei, Z., Li, S. Z. (2015). Fine-grained evaluation on face detection in the wild. In *IEEE international conference and workshops on automatic face and gesture recognition*, vol. 1, pp. 1–7.

Yang, Q., Jung, C., Fu, Q., Song, H. (2018). Low light image denoising based on poisson noise model and weighted tv regularization. In *Proceedings of IEEE international conference on image processing*, pp. 3199–3203.

Yang, S., Luo, P., Loy, C. C., Tang, X. (2016). Wider face: A face detection benchmark. In *Proceedings of IEEE international conference on computer vision and pattern recognition*, pp. 5525–5533.

Yang, W., Wang, S., Fang, Y., Wang, Y., Liu, J. (2020). From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *Proceedings of IEEE international conference on computer vision and pattern recognition*.

Ye, Z., Mohamadian, H., Ye, Y. (2007). Discrete entropy and relative entropy study on nonlinear clustering of underwater and arial images. In *IEEE international conference on control applications*, pp. 313–318.

Yeganeh, H., & Wang, Z. (2013). Objective quality assessment of tone-mapped images. *IEEE Transactions on Image Processing*, *22*(2), 657–667.

Ying, Z., Li, G., Gao, W. (2017). A bio-inspired multi-exposure fusion framework for low-light image enhancement. ArXiv e-prints.

Ying, Z., Li, G., Ren, Y., Wang, R., Wang, W. (2017). A new image contrast enhancement algorithm using exposure fusion framework. In Felsberg, M., Heyden, A., Krüger, N. (Eds.) *Computer analysis of images and patterns*, pp. 36–46.

Ying, Z., Li, G., Ren, Y., Wang, R., Wang, W. (2017). A new low-light image enhancement algorithm using camera response model. In *Proceedings of IEEE international conference on computer vision*.

Yu, S., & Zhu, H. (2019). Low-illumination image enhancement algorithm based on a physical lighting model. *IEEE Transactions on Circuits and Systems for Video Technology*, *29*(1), 28–37.

Zhang, L., Zhang, L., Mou, X., & Zhang, D. (2011). Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, *20*(8), 2378–2386.

Zhang, L., Zhang, L., & Bovik, A. C. (2015). A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, *24*(8), 2579–2591.

Zhang, Q., Nie, Y., Zhang, L., & Xiao, C. (2016). Underexposed video enhancement via perception-driven progressive fusion. *IEEE Transactions on Visualization and Computer Graphics*, *22*(6), 1773–1785.

Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., Li, S. Z. (2017). S3fd: Single shot scale-invariant face detector. In *Proceedings of IEEE international conference on computer vision*, pp. 192–201.

Zhang, X., Shen, P., Luo, L., Zhang, L., Song, J. (2012). Enhancement and noise reduction of very low light level images. In *Proceedings of IEEE international conference on pattern recognition*, pp. 2034–2037.

Zhang, Y., Zhang, J., Guo, X. (2019). Kindling the darkness: a practical low-light image enhancer. In *ACM international conference on multimedia*.

Zhu, M., Pan, P., Chen, W., Yang, Y. (2020) Eemefn: Low-light image enhancement via edge-enhanced multi-exposure fusion network. In *Proceedings of AAAI conference on artificial intelligence*.