EnlightenGAN: Deep Light Enhancement Without Paired Supervision

Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang, *Member, IEEE*

***Abstract*— Deep learning-based methods have achieved remarkable success in image restoration and enhancement, but are they still competitive when there is a lack of paired training data? As one such example, this paper explores the low-light image enhancement problem, where in practice it is extremely challenging to simultaneously take a low-light and a normal-light photo of the same visual scene. We propose a highly effective unsupervised generative adversarial network, dubbed *EnlightenGAN*, that can be trained without low/normal-light image pairs, yet proves to generalize very well on various real-world test images. Instead of supervising the learning using ground truth data, we propose to regularize the unpaired training using the information extracted from the input itself, and benchmark a series of innovations for the low-light image enhancement problem, including a global-local discriminator structure, a self-regularized perceptual loss fusion, and the attention mechanism. Through extensive experiments, our proposed approach outperforms recent methods under a variety of metrics in terms of visual quality and subjective user study. Thanks to the great flexibility brought by unpaired training, EnlightenGAN is demonstrated to be easily adaptable to enhancing real-world images from various domains. Our codes and pre-trained models are available at: https://github.com/VITA-Group/EnlightenGAN.**

***Index Terms*— Low-light enhancement, generative adversarial networks, unsupervised learning.**

I. Introduction

IMAGE captured in low-light conditions suffer from low contrast, poor visibility and high ISO noise. Those issues challenge both human visual perception that prefers high-visibility images, and numerous intelligent systems relying on computer vision algorithms such as all-day autonomous driving and biometric recognition [1]. To mitigate the degradation, a large number of algorithms have been proposed, ranging from histogram or cognition-based ones [2], [3] to learning-based approaches [4], [5]. The state-of-the-art image restoration and enhancement approaches using deep learning heavily rely on either synthesized or captured corrupted and clean image pairs to train, such as super-resolution [6], denoising [7] and deblurring [8].

However, the availability assumption of paired training images has raised more difficulties, when it comes to enhancing images from more uncontrolled scenarios, such as dehazing, deraining or low-light enhancement: 1) it is very difficult or even impractical to simultaneously capture corrupted and ground truth images of the same visual scene (e.g., low-light and normal-light image pairs at the same time); 2) synthesizing corrupted images from clean images could sometimes help, but such synthesized results are usually not photo-realistic enough, leading to various artifacts when the trained model is applied to real-world low-light images; 3) specifically for the low-light enhancement problem, there may be no unique or well-defined high-light ground truth given a low-light image. For example, any photo taken from dawn to dusk could be viewed as a high-light version for the photo taken over the midnight at the same scene. Taking into account the above issues, our overarching goal is to enhance a low-light photo with spatially varying light conditions and over/under-exposure artifacts, while the paired training data is unavailable.

Inspired by [9], [10] for unsupervised image-to-image translation, we adopt generative adversarial networks (GANs) to build an unpaired mapping between low and normal light image spaces without relying on exactly paired images. That frees us from training with only synthetic data or limited real paired data captured in controlled settings. We introduce a lightweight yet effective one-path GAN named **EnlightenGAN**, without using cycle-consistency as prior works [11]–[14] and therefore enjoying the merit of much shorter training time.

Due to the lack of paired training data, we incorporate a number of innovative techniques. We first propose a dual-discriminator to balance global and local low-light enhancement. Further, owing to the absence of ground-truth supervision, a self-regularized perceptual loss is proposed to constrain the feature distance between the low-light input image and its enhanced version, which is subsequently adopted both locally and globally together with the adversarial loss for training EnlightenGAN. We also propose to exploit the illumination information of the low-light input as a self-regularized attentional map in each level of deep features to regularize

Manuscript received September 11, 2019; revised April 14, 2020 and August 22, 2020; accepted December 22, 2020. Date of publication January 22, 2021; date of current version January 28, 2021. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jingyi Yu. *(Corresponding author: Zhangyang Wang.)*

Yifan Jiang, Xinyu Gong, and Zhangyang Wang are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: yifanjiang97@utexas.edu; xinyu.gong@utexas.edu; atlaswang@utexas.edu).

Ding Liu, Chen Fang, Xiaohui Shen, and Jianchao Yang are with Bytedance Inc., Mountain View, CA 94041 USA (e-mail: liuding@bytedance.com; fangchen@bytedance.com; shenxiaohui.kevin@bytedance.com; yangjianchao@bytedance.com).

Yu Cheng is with the Microsoft AI and Research, Redmond, WA 98052 USA (e-mail: yu.cheng@microsoft.com).

Pan Zhou is with the Department of Electronic Information and Communication, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: panzhou@hust.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2021.3051462, provided by the authors.

Digital Object Identifier 10.1109/TIP.2021.3051462

the unsupervised learning. Thanks to the unsupervised setting, we show that EnlightenGAN can be very easily adapted to enhancing real-world low-light images from different domains.

We highlight the notable innovations of EnlightenGAN:

- EnlightenGAN is the **first work** that successfully introduces unpaired training to low-light image enhancement. Such a training strategy removes the dependency on paired training data and enables us to train with larger varieties of images from different domains. It also avoids overfitting any specific data generation protocol or imaging device that previous works [5], [15], [16] implicitly rely on, hence leading to notably improved real-world generalization.
- EnlightenGAN gains remarkable performance by imposing (i) a global-local discriminator structure that handles spatially-varying light conditions in the input image; (ii) the idea of self-regularization, implemented by both the self feature preserving loss and the self-regularized attention mechanism. The self-regularization is critical to our model success, because of the unpaired setting where no strong form of external supervision is available.
- EnlightenGAN is compared with several state-of-the-art methods via comprehensive experiments. The results are measured in terms of visual quality, no-referenced image quality assessment, and human subjective survey. All results consistently endorse the superiority of EnlightenGAN. Moreover, in contrast to existing paired-trained enhancement approaches, EnlightenGAN proves particularly easy and flexible to be adapted to enhancing real-world low-light images from different domains.

II. Related Works

A. Paired Datasets: Status Quo

There exist several options to collect a paired dataset of low/normal-light images, but unfortunately none is efficient nor easily scalable. One may fix a camera and then reduce the exposure time in normal-light condition [5] or increase exposure time in low-light condition [16]. The LOL dataset [5] is so far the only dataset of low/normal-light image pairs taken from real scenes by changing exposure time and ISO. Due to the tedious experimental setup, e.g. the camera needs to be fixed and the object cannot move, etc., it consists of only 500 pairs. Moreover, it may still deviate from the true mapping between natural low/normal-light images. Especially under spatially varying lights, simply increasing/decreasing exposure time may lead to local over-/under-exposure artifacts.

In the high-dynamic-ranging (HDR) field, a few works first capture several images at different imperfect light conditions, then align and fuse them into one high-quality image [15], [17]. However, they are not designed for the purpose of post-processing only one single low-light image.

B. Traditional Approaches

Low-light image image enhancement has been actively studied as an image processing problem for long, with a few classical methods such as the adaptive histogram equalization (AHE) [3], Retinex [2] and multi-scale Retinex model [18]. More recently, [19] proposed an enhancement algorithm for non-uniform illumination images, utilizing a bi-log transformation to make a balance between details and naturalness. Based on the previous investigation of the logarithmic transformation, Fu *et al.* proposed a weighted variational model [20] to estimate both the reflectance and the illumination from an observed image with imposed regularization terms. In [21], a simple yet effective low-light image enhancement (LIME) was proposed, where the illumination of each pixel was first estimated by finding the maximum value in its RGB channels, then the illumination map was constructed by imposing a structure prior. Reference [22] introduced a joint low-light image enhancement and denoising model via decomposition in a successive image sequence. Reference [23] further proposed a robust Retinex model, which additionally considered a noise map compared with the conventional Retinex model, to improve the performance of enhancing low-light images accompanied by intensive noise.

C. Deep Learning Approaches

Existing deep learning solutions mostly rely on paired training, where most low-light images are synthesized from normal images. Reference [4] proposed a stacked auto-encoder (LL-Net) to learn joint denoising and low-light enhancement on the patch level. Retinex-Net in [5] provided an end-to-end framework to combine the Retinex theory and deep networks. HDR-Net [24] incorporated deep networks with the ideas of bilateral grid processing and local affine color transforms with pairwise supervision. A few multi-frame low-light enhancement methods were developed in the HDR domain, such as [15], [17], [25].

Lately, [16] proposed a "learning to see in the dark" model that achieves impressive visual results. However, this method operates directly on raw sensor data, in addition to the requirement of paired low/normal-light training images. Besides, it focuses more on avoiding the amplified artifacts during low-light enhancement by learning the pipeline of color transformations, demosaicing and denoising, which differs from EnlightenGAN in terms of settings and goal.

D. Adversarial Learning

GANs [26], [27] have proven successful in image synthesis and translation. When applying GANs to image restoration and enhancement, most existing works use paired training data as well, such as super resolution [28], artistic style transfer and image editing [29], [30], deraining [31] and dehazing [32]. Several unsupervised GANs are proposed to learn inter-domain mappings using adversarial learning and are adopted for many other tasks. References [9], [10] adopted a two-way GAN to translate between two different domains by using a cycle-consistent loss with unpaired data A handful of latest works followed their methodology and applied unpaired training with cycle-consistency to several low-level vision tasks, e.g. dehazing, deraining, super-resolution and mobile photo enhancement [33]–[36]. Different from them, EnlightenGAN refers to unpaired training but with a lightweight

Fig. 1. Representative visual examples by enhancing low-light images using EnlightenGAN. From left to right: columns 1, 3 and 5 are the low-light input images; while columns 2, 4, an 6 their corresponding enhanced images by EnlightenGAN.

one-path GAN structure (i.e., without cycle-consistency), which is stable and easy to train.

III. Method

As shown in Fig. 2, our proposed method adopts an attention-guided U-Net as the generator and uses the dual-discriminator to direct the global and local information. We also use a self feature preserving loss to guide the training process and maintain the textures and structures. In this section we first introduce two important building blocks, i.e., the global-local discriminators and the self feature preserving loss, then the whole network in details. The detailed network architectures are in the supplementary materials.

A. Global-Local Discriminators

We adopt the adversarial loss to minimize the distance between the real and output normal light distributions. However, we observe that an image-level vanilla discriminator often fails on spatially-varying light images; if the input image has some local area that needs to be enhanced differently from other parts, e.g., a small bright region in an overall dark background, the global image discriminator alone is often unable to provide the desired adaptivity.

Inspired by previous work [37], to enhance local regions adaptively in addition to improving the light globally, we propose a novel global-local discriminator structure, both using PatchGAN for real/fake discrimination. In addition to the image-level global discriminator, we add a local discriminator by taking randomly cropped local patches from both output and real normal-light images, and learning to distinguish whether they are *real* (from real images) or *fake* (from enhanced outputs). Such a global-local structure ensures all local patches of an enhanced images look like realistic normal-light ones, which proves to be critical in avoiding local over- or under-exposures as our experiments will reveal later.

Furthermore, for the global discriminator, we utilize the recently proposed relativistic discriminator structure [38] which estimates the probability that real data is more realistic than fake data and also directs the generator to synthesize a fake image that is more realistic than real images. The standard function of relativistic discriminator is:

$$D_{Ra}(x_r, x_f) = \sigma(C(x_r) - \mathbb{E}_{x_f \sim \mathbb{P}_{\text{fake}}}[C(x_f)]), \quad (1)$$

$$D_{Ra}(x_f, x_r) = \sigma(C(x_f) - \mathbb{E}_{x_r \sim \mathbb{P}_{\text{real}}}[C(x_r)]), \quad (2)$$

where C denotes the network of discriminator, x_r and x_f are sampled from the real and fake distribution, σ represents the sigmoid function. We slight modify the relativistic discriminator to replace the sigmoid function with the least-square GAN (LSGAN) [39] loss. Finally, the loss functions for the global discriminator D and the generator G are:

$$\mathcal{L}_D^{Global} = \mathbb{E}_{x_r \sim \mathbb{P}_{\text{real}}}[(D_{Ra}(x_r, x_f) - 1)^2] + \mathbb{E}_{x_f \sim \mathbb{P}_{\text{fake}}}[D_{Ra}(x_f, x_r)^2], \quad (3)$$

$$\mathcal{L}_G^{Global} = \mathbb{E}_{x_f \sim \mathbb{P}_{\text{fake}}}[(D_{Ra}(x_f, x_r) - 1)^2] + \mathbb{E}_{x_r \sim \mathbb{P}_{\text{real}}}[D_{Ra}(x_r, x_f)^2], \quad (4)$$

For the local discriminator, we randomly crop 5 patches from the output and real images each time. Here we adopt the original LSGAN as the adversarial loss, as follows:

$$\mathcal{L}_D^{Local} = \mathbb{E}_{x_r \sim \mathbb{P}_{\text{real-patches}}}[(D(x_r) - 1)^2] + \mathbb{E}_{x_f \sim \mathbb{P}_{\text{fake-patches}}}[(D(x_f) - 0)^2], \quad (5)$$

$$\mathcal{L}_G^{Local} = \mathbb{E}_{x_r \sim \mathbb{P}_{\text{fake-patches}}}[(D(x_f) - 1)^2], \quad (6)$$

B. Self Feature Preserving Loss

To constrain the perceptual similarity, Johnson *et al.* [40] proposed *perceptual loss* by adopting a pre-trained VGG to model feature space distance between images, which was widely adopted to many low-level vision tasks [28], [41]. The common practice constrains the extracted feature distance between the output image and its ground truth.

In our unpaired setting, we propose to instead constrain the VGG-feature distance between the input low-light and its enhanced normal-light output. This is based on our empirical observation that the classification results by VGG models are

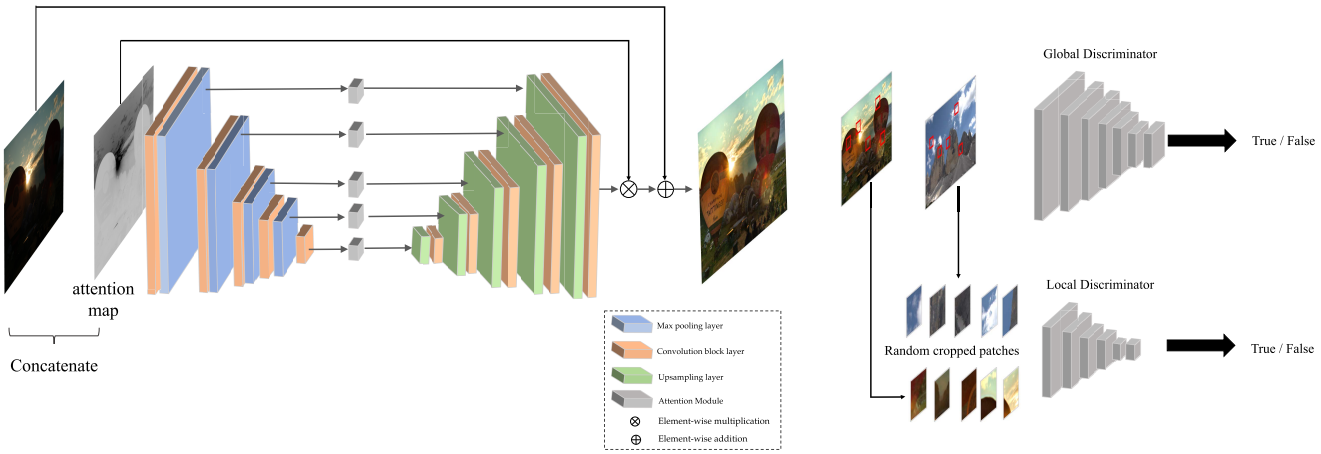


Fig. 2. The overall architecture of EnlightenGAN. In the generator, each convolutional block consists of two 3 × 3 convolutional layers followed by batch normalization and LeakyRelu. Each attention module has the feature map multiply with a (resized) attention map.

not very sensitive when we manipulate the input pixel intensity range, which is concurred by another recent study [42]. We call it *self feature preserving loss* to stress its self-regularization utility to preserve the image content features to itself, before and after the enhancement. That is distinct from the typical usage of the perceptual loss in (paired) image restoration, and is motivated from our unpaired setting too. Concretely, the self feature preserving loss L_{SFP} is defined as:

$$\mathcal{L}_{SFP}(I^L) = \frac{1}{W_{i,j} H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I^L) - \phi_{i,j}(G(I^L)))^2, \tag{7}$$

where I^L denotes the input low-light image and $G(I^L)$ denotes the generator's enhanced output. $\phi_{i,j}$ denotes the feature map extracted from a VGG-16 model pre-trained on ImageNet. i represents its i-th max pooling, and j represents its j-th convolutional layer after i-th max pooling layer. $W_{i,j}$ and $H_{i,j}$ are the dimensions of the extracted feature maps. By default we choose $i = 5$, $j = 1$.

For our local discriminator, the cropped local patches from input and output images are also regularized by a similarly defined self feature preserving loss, L_{SFP}^{Local}. Furthermore, We add an instance normalization layer [43] after the VGG feature maps before feeding into L_{SFP} and L_{SFP}^{Local} in order to stabilize training. The overall loss function for training EnlightenGAN is thus written as:

$$Loss = \mathcal{L}_{SFP}^{Global} + \mathcal{L}_{SFP}^{Local} + \mathcal{L}_{G}^{Global} + \mathcal{L}_{G}^{Local}, \tag{8}$$

C. U-Net Generator Guided With Self-Regularized Attention

U-Net [44] has achieved huge success on semantic segmentation, image restoration and enhancement [45]. By extracting multi-level features from different depth layers, U-Net preserves rich texture information and synthesizes high quality images using multi-scale context information. We adopt U-Net as our generator backbone.

We further propose an easy-to-use attention mechanism for the U-Net generator. Intuitively, in a low-light image of spatially varying light condition, we always want to enhance the dark regions more than bright regions, so that the output image has neither over- nor under-exposure. We take the illumination channel I of the input RGB image, normalize it to [0,1], and then use $1-I$ (element-wise difference) as our self-regularized attention map. We then resize the attention map to fit each feature map and multiply it with all intermediate feature maps as well as the output image. We emphasize that our attention map is also a form of self-regularization, rather than learned with supervision. Despite its simplicity, the attention guidance shows to improve the visual quality consistently.

Our attention-guided U-Net generator is implemented with 8 convolutional blocks. Each block consists of two 3 × 3 convolutional layers, followed by LeakyReLu and a batch normalization layer [46]. At the upsampling stage, we replace the standard deconvolutional layer with one bilinear upsampling layer plus one convolutional layer, to mitigate the checkerboard artifacts. The final architecture of EnlightenGAN is illustrated in the left of Fig. 2. The detailed configuration could be found in the supplementary materials.

IV. Experiments

A. Dataset and Implementation Details

Because EnlightenGAN has the unique ability to be trained with unpaired low/normal light images, we are enabled to collect a larger-scale unpaired training set, that covers diverse image qualities and contents. We assemble a mixture of 914 low light and 1016 normal light images from several datasets released in [5], [47] and also HDR sources [15], [25], without the need to keep any pair.[1] Manual inspection and selection are performed to remove images of medium brightness. All these photos are converted to PNG format and resized to 600 × 400 pixels. For testing images, we choose those

[1]The LOL dataset by [5] was a small paired dataset, but we did not use them as pairs for training. An exception is that, we hold out a subset of 50 low/normal light image pairs from LOL [5], as the validation set.

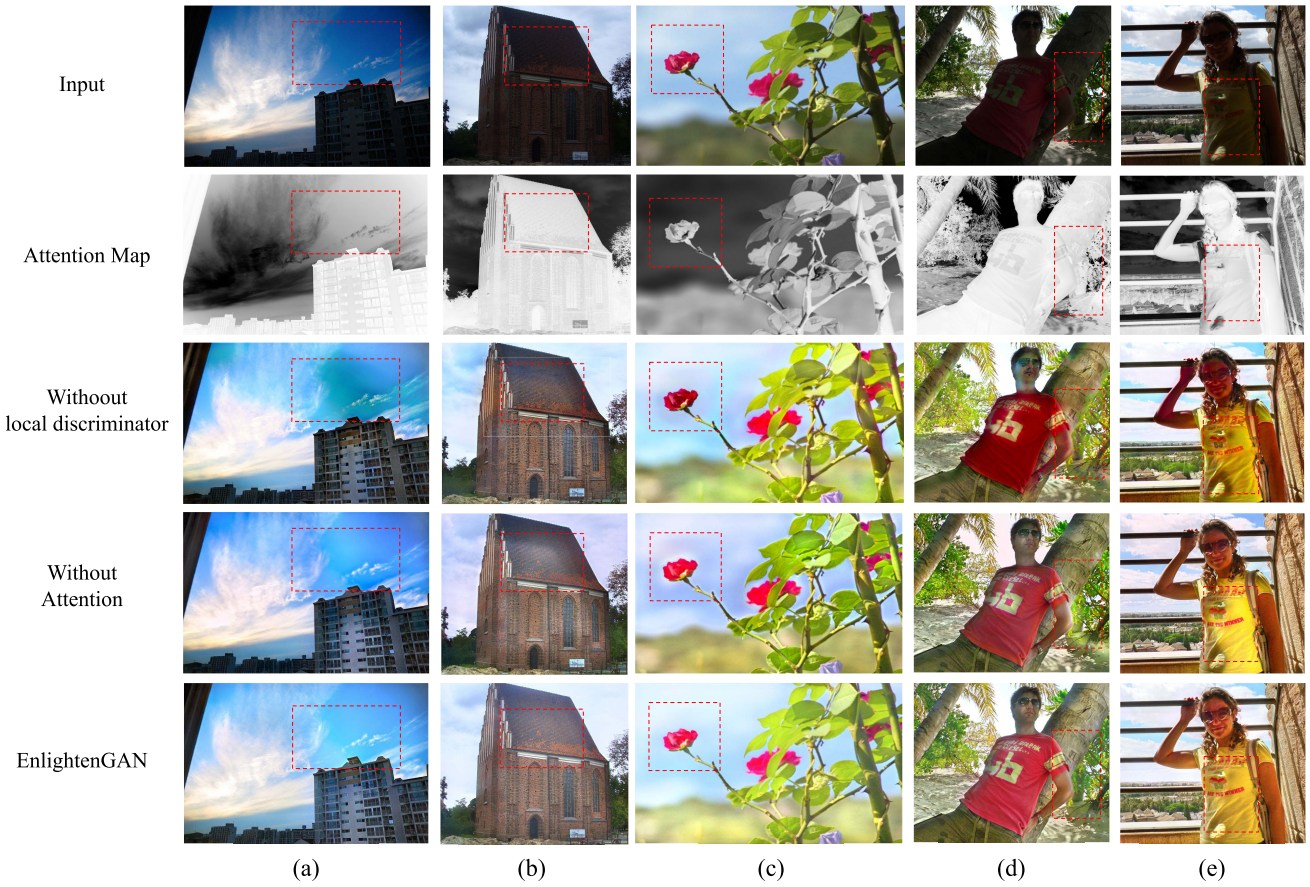


Fig. 3. Visual comparison from the ablation study of EnlightenGAN. Row 1∼5 display the low-light image inputs, the attention map of input, results from EnlightenGAN with only global discriminator, results from EnlightenGAN without self-regularized attention mechanism, and results from the final version of EnlightenGAN, respectively. Images in Row 3 and 4 suffer from severe color distortion or inconsistency, which are highlighted by bounding boxes. The final version of EnlightenGAN is able to mitigate the above issues and gains the most visually pleasing results. Please zoom in to see the details.

standard ones used in previous works (NPE [19], LIME [21], MEF [48], DICM [49], VV,[2] etc.).

EnlightenGAN is first trained from the scratch for 100 epochs with the learning rate of 1e-4, followed by another 100 epochs with the learning rate linearly decayed to 0. We use the Adam optimizer and the batch size is set to be 32. Thanks to the lightweight design of one-path GAN without using cycle-consistency, the training time is much shorter than cycle based methods. The whole training process takes 3 hours on 3 Nvidia 1080Ti GPUs.

B. Ablation Study

To demonstrate the effectiveness of each component proposed in Sec. III, we conduct several ablation experiments. Specifically, we design two experiments by removing the components of local discriminator and attention mechanism, respectively. As shown in Fig. 3, the first row shows the input images. The second row shows the attention map of the input images, we can easily observe that the attention map gives a good guideline to the algorithm by which region should be enhanced more while others should be enhanced less. The third row shows the image produced by EnlightenGAN with only global discriminator to distinguish between low-light and normal-light images. The fourth row is the result produced by EnlightenGAN which does not adopt self-regularized attention mechanism and uses U-Net as the generator instead. The last row is produced by our proposed version of EnlightenGAN.

The enhanced results in the third row and the fourth row tend to contain local regions of severe color distortion or under-exposure, namely, the sky over the building in Fig.3(a), the roof region in Fig.3(b), the left blossom in Fig.3(c), the boundary of tree and bush in Fig.3(d), and the T-shirt in Fig.3(e). In contrast, the results of the full EnlightenGAN contain realistic color and thus more visually pleasing, which validates the effectiveness of the global-local discriminator design and self-regularized attention mechanism. More images are in the supplementary materials.

C. Comparison With State-of-the-Arts

In this section we compare the performance of EnlightenGAN with current state-of-the-art methods. We conduct a list of experiments including visual quality comparison, human subjective review and no-referenced image quality assessment (IQA), which are elaborated on next.

1) Visual Quality Comparison: We first compare the visual quality of EnlightenGAN with several recent competing methods. Results are demonstrated in Fig. 4, where the first

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

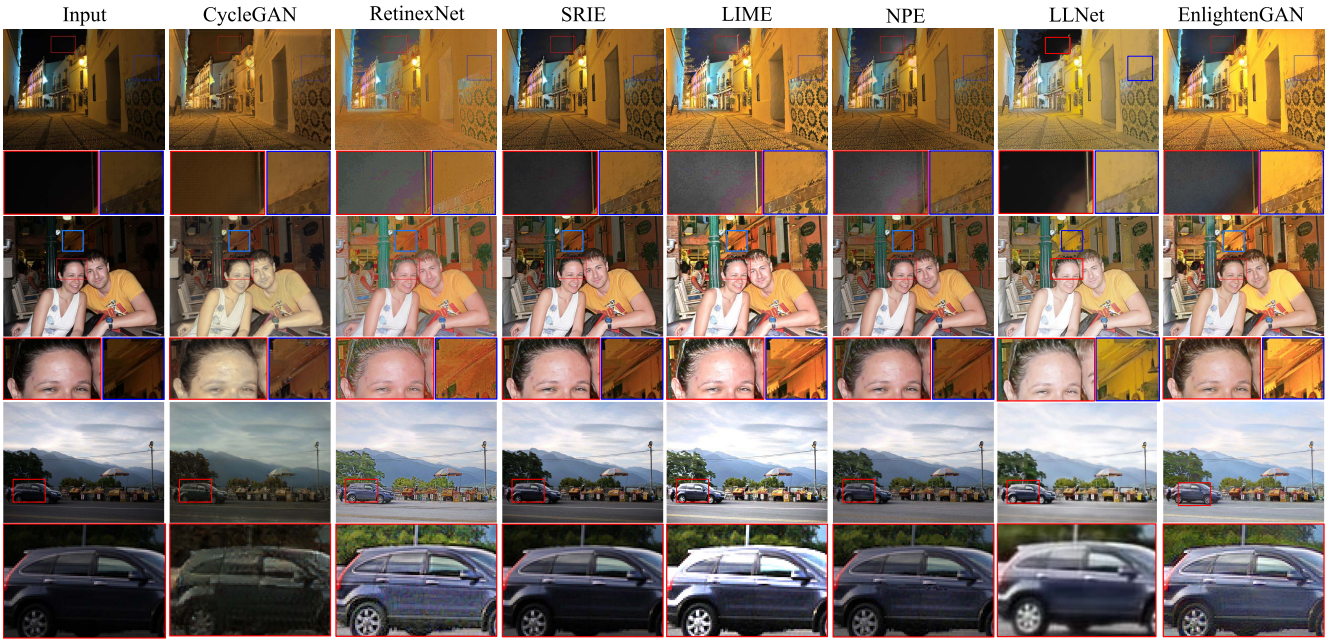


Fig. 4. Comparison with other state-of-the-art methods. Zoom-in regions are used to illustrate the visual differences. Three examples are listed from the top to the bottom rows. First example: EnlightenGAN successfully suppresses the noise in black sky and produces the best visible details of yellow wall. Second example: NPE and SRIE fail to enhance the background details. LIME introduces over-exposure on the woman's face. LLNet generate severe color distortion. However, EnlightenGAN not only restores the background details but also avoids over-exposure artifacts, distinctly outperforming other methods. Third example: EnlightenGAN produces a visually pleasing result while avoiding over-exposure artifacts in the car and cloud. Others either do not enhance dark details enough or generate over-exposure artifacts. Please zoom in to see more details.

column shows the original low-light images, and the second to fifth columns are the images enhanced by: a vanilla CycleGAN [9] trained using our unpaired training set, RetinexNet [5], SRIE [20], LIME [21], NPE [19], LLNet [4], and CycleGAN [9]. The last column shows the results produced by EnlightenGAN.

We next zoom in on some details in the bounding boxes. LIME easily leads to over-exposure artifacts, which makes the results distorted and glaring with the some information missing. The results of SRIE and NPE are generally darker compared with others. CycleGAN and RetinexNet generate unsatisfactory visual results in terms of both brightness and naturalness. In contrast, EnlightenGAN successfully not only learns to enhance the dark area but also preserves the texture details and avoids over-exposure artifacts. More results are shown in the supplementary materials.

2) No-Referenced Image Quality Assessment: We adopt Natural Image Quality Evaluator (NIQE) [50], a well-known no-reference image quality assessment for evaluating real image restoration without ground-truth, to provide quantitative comparisons. The NIQE results on five publicly available image sets used by previous works (MEF, NPE, LIME, VV, and DICM) are reported in Table I: a lower NIQE value indicates better visual quality. EnlightenGAN wins on three out of five sets, and is the best in terms of overall averaged NIQE. This further endorses the superiority of EnlightenGAN over current state-of-the-art methods in generating high-quality visual results.

3) Human Subjective Evaluation: We conduct a human subjective study to compare the performance of EnlightenGAN and other methods. We randomly select 23 images from the testing set. For each image, it is first enhanced by five methods (LIME, RetinexNet, NPE, SRIE, and EnlightenGAN). We then ask 9 subjects to independently compare the five outputs in a pairwise manner. Specifically, each time a human subject is displayed with a pair of images randomly drawn from the five outputs, and is asked to evaluated which one has better quality. The human subjects are instructed to consider the: 1) whether the images contain visible noise; 2) whether the images contain over- or under-exposure artifacts; and 3) whether the images show nonrealistic color or texture distortions. Next, we fit a Bradley-Terry model [51] to estimate the numerical subjective scores so that the five methods can be ranked, using the exactly same routine as described in previous works [52]. As a result, each method is assigned with rank 1-5 on that image. We repeat the above for all 23 images.

TABLE I

NIQE SCORES ON THE WHOLE TESTING SET (ALL) AND EACH SUBSET (MEF, LIME, NPE, VV, DICM) RESPECTIVELY. SMALLER NIQE INDICATES MORE PERCEPTUALLY FAVORED QUALITY

Image set	MEF	LIME	NPE	VV	DICM	All
Input	4.265	4.438	4.319	3.525	4.255	4.134
LLNet	4.845	4.940	4.78	4.446	4.809	4.751
CycleGAN	3.782	3.276	4.036	3.343	3.560	3.554
RetinexNet	4.149	4.420	4.485	2.602	4.200	3.920
LIME	3.720	4.155	4.268	**2.489**	3.846	3.629
SRIE	3.475	3.788	3.986	2.850	3.899	3.650
NPE	3.524	3.905	**3.953**	2.524	3.760	3.525
EnlightenGAN	**3.232**	**3.719**	4.113	2.581	**3.570**	**3.385**

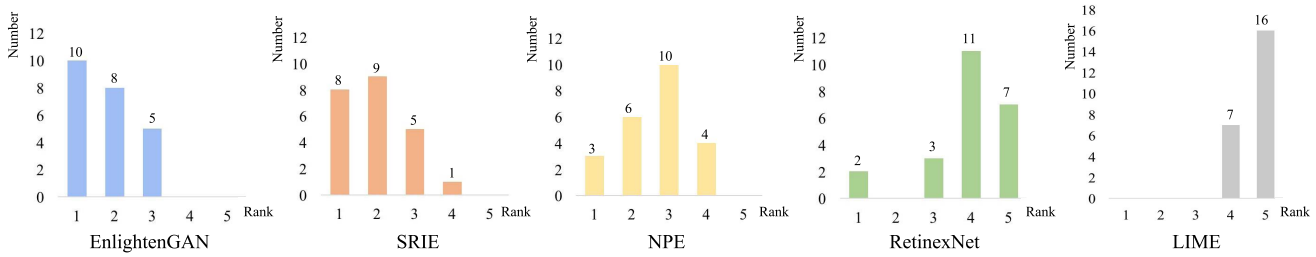


Fig. 5. The result of five methods in the human subjective evaluation. In each histogram, x-axis denotes the ranking index ($1 \sim 5$, 1 represents the highest), and y-axis denotes the number of images in each ranking index. EnlightenGAN produces the most top-ranking images and gains the best performance with the smallest average ranking value.

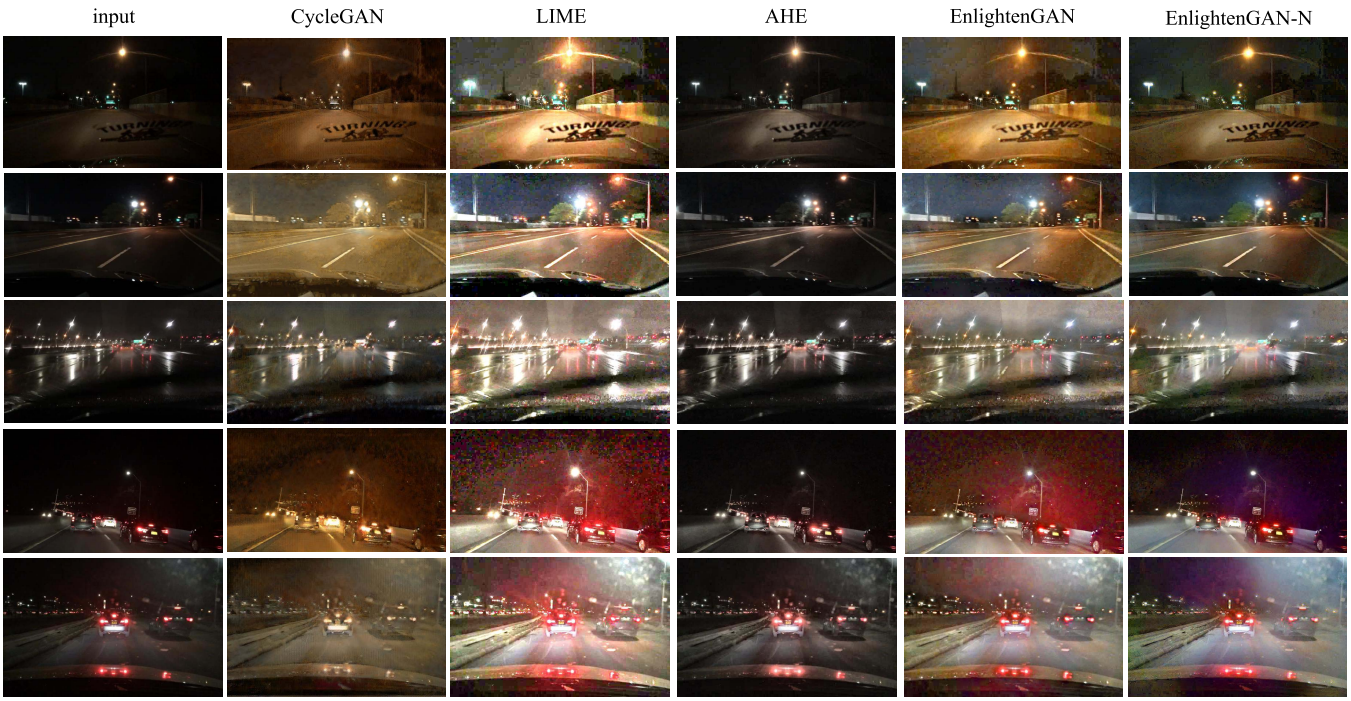


Fig. 6. Visual comparison of the results on the BBD-100k dataset [1]. EnlightenGAN-N is the domain-adapted version of EnlightenGAN, which generates the most visually pleasing results with noise suppressed. Please zoom in to see the details.

Fig. 5 displays the five histograms, each of which depicts the rank distributions that a method receives on the 23 images. For example, EnlightGAN has been ranked the 1st (i.e., the highest subjective score) on 10 out of 23 images, the 2nd for 8 images, and the 3rd for 5 images. By comparing the five histograms, it is clear that EnlightenGAN produces the overall most favored results by human subjects, with an average ranking of 1.78 over 23 images. RetinexNet and LIME are not well scored, because of causing many over-exposures and sometimes amplifying the noise.

D. Adaptation on Real-World Images

Domain adaptation is an indispensable factor for real-world generalizable image enhancement. The unpaired training strategy of EnlightenGAN allows us to directly learn to enhance real-world low-light images from various domains, where there is **no paired normal-light** training data or even **no normal-light** data from the same domain available. We conduct experiments using low-light images from a real-world driving dataset, Berkeley Deep Driving (BBD-100k) [1], to showcase this **unique advantage** of EnlightenGAN in practice.

We pick 950 night-time photos (selected by mean pixel intensity values smaller than 45) from the BBD-100k set as the low-light training images, plus 50 low-light images for hold-out testing. Those low-light images suffer from severe artifacts and high ISO noise. We then compare two EnlightenGAN versions trained on different normal-light image sets, including: 1) the pre-trained EnlightenGAN model as described in Sec. IV-A, without any adaptation for BBD-100k; 2) **EnlightenGAN-N**: a domain-adapted version of EnlightenGAN, which uses BBD-100k low-light images from the BBD-100k dataset for training, while the normal-light images are still the high-quality ones from our unpaired dataset in Sec. IV-A. We also include a traditional method, Adaptive histogram equalization (AHE), and a pre-trained LIME model for comparison, and an unsupervised approach CycleGAN.

As shown in Fig. 6, the results from LIME suffer from severe noise amplification and over-exposure artifacts, while AHE does not enhance the brightness enough. The unsupervised approach CycleGAN generate very low quality due to its unstability. The original EnlightenGAN also leads to noticeable artifacts on this unseen image domain. In comparison, EnlightenGAN-N produces the most visually pleasing results, striking an impressive balance between brightness and artifact/noise suppression. Thanks to the unpaired training, EnlightenGAN could be easily adapted into EnlightenGAN-N without requiring any supervised/paired data in the new domain, which greatly facilitates its real-world generalization.

E. Pre-Processing for Improving Classification

Image enhancement as pre-processing for improving subsequent high-level vision tasks has recently received increasing attention [41], [53]–[55], with a number of benchmarking efforts [52], [56]–[58]. We investigate the impact of light enhancement on the *extremely dark* (**ExDark**) dataset [59], which was specifically built for the task of low-light image recognition. The classification results after light enhancement could be treated as an indirect measure on semantic information preservation, as [41], [52] suggested.

The ExDark dataset consists of 7,363 low-light images, including 3000 images in training set, 1800 images in validation set and 2563 images in testing set, annotated into 12 object classes. We use its testing set only, applying our pretrained EnlightenGAN as a pre-processing step, followed by passing through another ImageNet-pretrained ResNet-50 classifier. Neither domain adaption nor joint training is performed. The high-level task performance serves as a fixed semantic-aware metric for enhancement results.

In the low-light testing set, using EnlightenGAN as pre-processing improves the classification accuracy from 22.02% (top-1) and 39.46% (top-5), to 23.94% (top-1) and 40.92% (top-5) after enhancement. That supplies a side evidence that EnlightenGAN preserves semantic details, in addition to producing visually pleasing results. We also conduct experiment using LIME and AHE. LIME improves the accuracy to 23.32% (top-1) and 40.60% (top-5), while AHE obtains to 23.04% (top-1) and 40.37% (top-5).

V. Conclusion

In this paper, we address the low-light enhancement problem with a novel and flexible unsupervised framework. The proposed EnlightenGAN operates and generalizes well without any paired training data. The experimental results on various low light datasets show that our approach outperforms multiple state-of-the-art approaches under both subjective and objective metrics. Furthermore, we demonstrate that EnlightenGAN can be easily adapted on real noisy low-light images and yields visually pleasing enhanced images. Our future work will explore how to control and adjust the light enhancement levels based on user inputs in one unified model. Due to the complicacy of light enhancement, we also expect integrate algorithm with sensor innovations.

References

[1] F. Yu *et al.*, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," 2018, *arXiv:1805.04687*. [Online]. Available: http://arxiv.org/abs/1805.04687

[2] E. H. Land, "The Retinex theory of color vision," *Sci. Amer.*, vol. 237, no. 6, pp. 108–129, 1977.

[3] S. M. Pizer *et al.*, "Adaptive histogram equalization and its variations," *Comput. Vis., Graph., Image Process.*, vol. 39, no. 3, pp. 355–368, 1987.

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

[5] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep Retinex decomposition for low-light enhancement," 2018, *arXiv:1808.04560*. [Online]. Available: http://arxiv.org/abs/1808.04560

[6] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[7] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[8] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8174–8182.

[9] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, 2017, pp. 2223–2232.

[10] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 700–708.

[11] T. M. Nimisha, K. Sunil, and A. Rajagopalan, "Unsupervised class-specific deblurring," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 353–369.

[12] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 172–189.

[13] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain Image-to-Image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[14] J. Hoffman *et al.*, "CyCADA: Cycle-consistent adversarial domain adaptation," 2017, *arXiv:1711.03213*. [Online]. Available: http://arxiv.org/abs/1711.03213

[15] N. K. Kalantari and R. Ramamoorthi, "Deep high dynamic range imaging of dynamic scenes," *ACM Trans. Graph.*, vol. 36, no. 4, p. 144, 2017.

[16] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," 2018, *arXiv:1805.01934*. [Online]. Available: http://arxiv.org/abs/1805.01934

[17] S. Wu, J. Xu, Y.-W. Tai, and C.-K. Tang, "Deep high dynamic range imaging with large foreground motions," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 117–132.

[18] D. J. Jobson, Z. Rahman, and G. A. Woodell, "A multiscale Retinex for bridging the gap between color images and the human observation of scenes," *IEEE Trans. Image Process.*, vol. 6, no. 7, pp. 965–976, Jul. 1997.

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

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

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

[22] X. Ren, M. Li, W.-H. Cheng, and J. Liu, "Joint enhancement and denoising method via sequential decomposition," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2018, pp. 1–5.

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

[24] M. Gharbi, J. Chen, J. T. Barron, S. W. Hasinoff, and F. Durand, "Deep bilateral learning for real-time image enhancement," *ACM Trans. Graph.*, vol. 36, no. 4, p. 118, 2017.

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

[26] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[27] X. Gong, S. Chang, Y. Jiang, and Z. Wang, "AutoGAN: Neural architecture search for generative adversarial networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3224–3234.

[28] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jul. 2017, pp. 4681–4690.

[29] S. Yang, Z. Wang, Z. Wang, N. Xu, J. Liu, and Z. Guo, "Controllable artistic text style transfer via shape-matching GAN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4442–4451.

[30] S. Yang, Z. Wang, J. Liu, and Z. Guo, "Deep plastic surgery: Robust and controllable image editing with human-drawn sketches," 2020, *arXiv:2001.02890*. [Online]. Available: http://arxiv.org/abs/2001.02890

[31] R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu, "Attentive generative adversarial network for raindrop removal from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2482–2491.

[32] R. Li, J. Pan, Z. Li, and J. Tang, "Single image dehazing via conditional generative adversarial network," *Methods*, vol. 3, p. 24, Sep. 2018.

[33] X. Yang, Z. Xu, and J. Luo, "Towards perceptual image dehazing by physics-based disentanglement and adversarial training," in *Proc. AAAI*, vol. 32, no. 1, Apr. 2018.

[34] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 30–32.

[35] Y.-S. Chen, Y.-C. Wang, M.-H. Kao, and Y.-Y. Chuang, "Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6306–6314.

[36] X. Jin, Z. Chen, J. Lin, Z. Chen, and W. Zhou, "Unsupervised single image deraining with self-supervised constraints," 2018, *arXiv:1811.08575*. [Online]. Available: http://arxiv.org/abs/1811.08575

[37] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.

[38] A. Jolicoeur-Martineau, "The relativistic discriminator: A key element missing from standard GAN," 2018, *arXiv:1807.00734*. [Online]. Available: http://arxiv.org/abs/1807.00734

[39] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2813–2821.

[40] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2016, pp. 694–711.

[41] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "DeblurGAN: Blind motion deblurring using conditional adversarial networks," 2017, *arXiv:1711.07064*. [Online]. Available: http://arxiv.org/abs/1711.07064

[42] B. R. Webster, S. E. Anthony, and W. J. Scheirer, "PsyPhy: A psychophysics driven evaluation framework for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2280–2286, Sep. 2019.

[43] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jul. 2017, pp. 6924–6932.

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

[45] D. Liu, B. Wen, X. Liu, Z. Wang, and T. Huang, "When image denoising meets high-level vision tasks: A deep learning approach," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 842–848.

[46] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[47] D.-T. Dang-Nguyen, C. Pasquini, V. Conotter, and G. Boato, "Raise: A raw images dataset for digital image forensics," in *Proc. 6th ACM Multimedia Syst. Conf.*, 2015, pp. 219–224.

[48] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3345–3356, Nov. 2015.

[49] C. Lee, C. Lee, and C.-S. Kim, "Contrast enhancement based on layered difference representation," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 965–968.

[50] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[51] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, nos. 3–4, p. 324, Dec. 1952.

[52] B. Li *et al.*, "Benchmarking single-image dehazing and beyond," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 492–505, Jan. 2019.

[53] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "AOD-net: All-in-one dehazing network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4770–4778.

[54] Y. Liu *et al.*, "Improved techniques for learning to dehaze and beyond: A collective study," 2018, *arXiv:1807.00202*. [Online]. Available: http://arxiv.org/abs/1807.00202

[55] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8878–8887.

[56] Y. Yuan, W. Yang, W. Ren, J. Liu, W. J. Scheirer, and Z. Wang, "UG^{2+} track 2: A collective benchmark effort for evaluating and advancing image understanding in poor visibility environments," 2019, *arXiv:1904.04474*. [Online]. Available: http://arxiv.org/abs/1904.04474

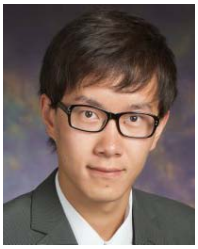
[57] S. Li *et al.*, "Single image deraining: A comprehensive benchmark analysis," 2019, *arXiv:1903.08558*. [Online]. Available: http://arxiv.org/abs/1903.08558

[58] W. Yang *et al.*, "Advancing image understanding in poor visibility environments: A collective benchmark study," *IEEE Trans. Image Process.*, vol. 29, pp. 5737–5752, 2020.

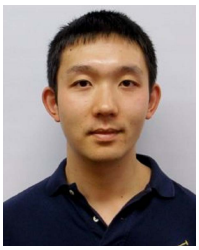
[59] Y. P. Loh and C. S. Chan, "Getting to know low-light images with the exclusively dark dataset," *Comput. Vis. Image Understand.*, vol. 178, pp. 30–42, Jan. 2019.

Yifan Jiang received the bachelor's degree from the Huazhong University of Science and Technology, in 2019. He is currently pursuing the Ph.D. degree with The University of Texas at Austin. His research interests include computer vision and deep learning. He mainly works on the area of generative models and image enhancement.

Xinyu Gong received the bachelor's degree from the University of Electronic Science and Technology of China in 2018. He is currently pursuing the Ph.D. degree in electrical and computer engineering with The University of Texas at Austin. His research interests include AutoML and generative models.



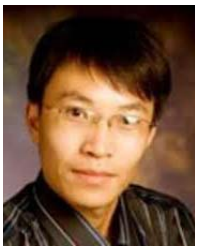
Ding Liu received the Ph.D. degree from the University of Illinois at Urbana–Champaign, USA, in 2018. He is currently a Research Scientist with Bytedance Inc., Mountain View, CA, USA. His research experience encompasses low-level vision problems, including image/video restoration, and enhancement. He has broad research interests include in the area of computer vision, image processing, and deep learning.



Yu Cheng received the bachelor's degree from Tsinghua University in 2010, and the Ph.D. degree from Northwestern University in 2015. Before that, he spent three years as a Research Staff Member at the IBM T. J. Watson Research Center. He is currently a Researcher with Microsoft. His research interests include deep learning in general, with specific interests in the deep generative model, model compression, and reinforcement learning. He is also interested in solving real-world problems of computer vision and natural language processing. He regularly serves on the program committees of top-tier AI conferences, such as NIPS, ICML, ICLR, CVPR, and ACL.

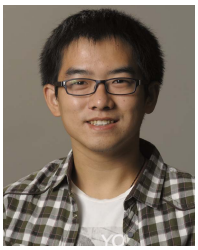
Chen Fang received the bachelor's degree from Hunan University in 2010, and the Ph.D. degree from Dartmouth in 2015. He was a Research Scientist with Adobe Research. He is currently a Researcher with Bytedance Inc. His research interests include image understanding, image search, image editing, and generative models.

Xiaohui Shen received the B.S. and M.S. degrees from the Department of Automation, Tsinghua University, China, and the Ph.D. degree from the Department of EECS, Northwestern University. He was a Senior Research Scientist with Adobe Research. He is currently a Researcher with the Bytedance AI Laboratory. His research interests include computer vision and deep learning.



Jianchao Yang received the M.S. and Ph.D. degrees from the ECE Department, University of Illinois at Urbana–Champaign, under the supervision of Prof. T. Huang. He was a Research Scientist with Adobe Research. He is currently a Director with Bytedance Inc. He has authored more than 80 technical papers over a wide variety of topics on top tier conferences and journals, with Google scholar citation more than 12 000 times. His research interests include computer vision, deep learning, and image and video processing. He received the Best Student Paper Award from ICCV 2010, the Classification Task Prize in PASCAL VOC 2009, first position for object localization using external data for ILSVRC ImageNet 2014, and Third Place in the WebVision Challenge 2017. He serves as the Workshop Chair for the ACM MM 2017.

Pan Zhou received the Ph.D. degree from the School of Electrical and Computer Engineering, Georgia Institute of Technology, Georgia Tech., Atlanta, GA, USA, in 2011. He is currently a Full Professor with the Hubei Engineering Research Center on Big Data Security, School of Cyber Science and Engineering, Huazhong University of Science and Technology. He was a Senior Technical Member with Oracle Inc., from 2011 to 2013. His current research interests include security and privacy, big data analytics, machine learning, and information networks.



Zhangyang Wang (Member, IEEE) received the Ph.D. degree the ECE Department, University of Illinois at Urbana–Champaign, under the supervision of Prof. T. Huang. He was an Assistant Professor with Texas A&M University, from 2017 to 2020. He is currently an Assistant Professor with The University of Texas at Austin. He is broadly interested in the fields of machine learning, computer vision, optimization, and their interdisciplinary applications. His latest research interests include automated machine learning (AutoML), learning-based optimization, machine learning robustness, and efficient deep learning.