Towards Unsupervised Deep Image Enhancement With Generative Adversarial Network

Zhangkai Ni, *Graduate Student Member, IEEE*, Wenhan Yang, *Member, IEEE*, Shiqi Wang, *Member, IEEE*, Lin Ma, *Member, IEEE*, and Sam Kwong, *Fellow, IEEE*

Abstract—**Improving the aesthetic quality of images is challenging and eager for the public. To address this problem, most existing algorithms are based on *supervised* learning methods to learn an automatic photo enhancer for *paired* data, which consists of low-quality photos and corresponding expert-retouched versions. However, the style and characteristics of photos retouched by experts may not meet the needs or preferences of general users. In this paper, we present an *unsupervised* image enhancement generative adversarial network (UEGAN), which learns the corresponding *image-to-image* mapping from a set of images with desired characteristics in an *unsupervised* manner, rather than learning on a large number of paired images. The proposed model is based on *single* deep GAN which embeds the *modulation* and *attention* mechanisms to capture richer global and local features. Based on the proposed model, we introduce two losses to deal with the unsupervised image enhancement: (1) *fidelity loss*, which is defined as a $\ell 2$ regularization in the feature domain of a pre-trained VGG network to ensure the content between the enhanced image and the input image is the same, and (2) *quality loss* that is formulated as a relativistic hinge adversarial loss to endow the input image the desired characteristics. Both quantitative and qualitative results show that the proposed model effectively improves the aesthetic quality of images. Our code is available at: https://github.com/eezkni/UEGAN.**

Index Terms—**Unsupervised learning, image enhancement, global attention, generative adversarial network.**

I. Introduction

With the rapid development of mobile Internet, smart electronic devices, and social networks, it is becoming more and more popular to record and upload the wonderful lives of people through social media and online sharing communities. However, due to the high cost of high-quality hardware devices and the lack of professional photography skills, the aesthetic quality of photos taken by the general public is often unsatisfactory. Professional image-editing is expensive, and it is hard to provide such services in an automated manner as aesthetic feelings and preferences are usually a personal issue. Therefore, the *automatic image enhancement techniques* providing the *user-oriented* image beautification are preferred.

Compared with high-quality images, low-quality images usually suffer from multiple degradations in visual quality, such as poor colors, low contrast, and intensive noises *et al*. Therefore, the image enhancement process needs to address this degradation with a series of enhancement operations, such as contrast enhancement, color correction, and details adjustment *et al*. The earliest conventional image enhancement approaches mainly focused on contrast enhancement of low-quality image [1]–[3]. The most common histogram adjustment transfers the luminance histogram of a low-quality image to a given distribution (may be provided by other reference images) to stretch the contrast of the low-quality image. According to the transformation scope, this kind of method can be further classified into two categories: *global* histogram equalization (GHE) [2], [4] and *local* histogram equalization (LHE) [3], [5]. The former uses a single histogram transformation function to adjust all pixels of the entire image. It may lead to improper enhancement results in some local regions, such as under-exposure, over-exposure, color distortion, *et al*. To address this issue, the LHE derives the content adaptive transform functions based on the statistical information in local region and applies these transforms locally. However, the LHE is computationally complex and not always powerful because the extracted transformation depends on the dominating information in the local region. Therefore, they are also easy to generate visually unsatisfactory texture details, dull or over-saturated color.

For the past few years, deep convolutional neural networks (CNN) have made significant progress in low-level vision tasks [6]–[8]. In order to improve the modeling capacity and adaptivity, deep learning-based models are built to introduce the excellent expressive power of deep networks to facilitate automatic image enhancement with the knowledge of big data. Ignatov *et al.* [7] designed an end-to-end deep learning network that improves photos from mobile devices to the quality of digital single-lens reflex (DSLR) photos.

Manuscript received March 16, 2020; revised August 6, 2020; accepted September 1, 2020. Date of publication September 22, 2020; date of current version September 29, 2020. This work was supported in part by the Natural Science Foundation of China under Grant 61772344 and Grant 61672443, in part by the Hong Kong Research Grants Council (RGC) General Research Funds under Grant 9042816 (CityU 11209819) and Grant 9042957 (CityU 11203220), in part by the Hong Kong Research Grants Council (RGC) Early Career Scheme under Grant 9048122 (CityU 21211018), and in part by the Key Project of Science and Technology Innovation 2030 supported by the Ministry of Science and Technology of China under Grant 2018AAA0101301. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nikolaos Mitianoudis. *(Corresponding authors: Shiqi Wang; Sam Kwong.)*

Zhangkai Ni, Wenhan Yang, and Shiqi Wang are with the Department of Computer Science, City University of Hong Kong, Hong Kong (e-mail: eezkni@gmail.com; yangwenhan@pku.edu.cn; shiqwang@cityu.edu.hk).

Lin Ma is with the Meituan-Dianping Group, Beijing 100102, China (e-mail: forest.linma@gmail.com).

Sam Kwong is with the Department of Computer Science, City University of Hong Kong, Hong Kong, and also with the City University of Hong Kong, Shenzhen Research Institute, Shenzhen 518057, China (e-mail: cssamk@cityu.edu.hk).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/TIP.2020.3023615

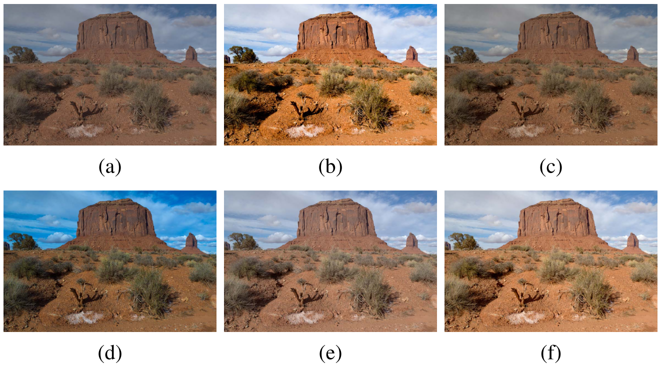


Fig. 1. An illustration of different expert-retouched versions of a low-quality photo in MIT-Adobe FiveK dataset [11]. (a) is the low-quality photo and (b) to (f) are different high-quality counterparts retouched by different experts. The obvious perceptual differences exist among different high-quality versions.

Ren *et al.* [9] present a hybrid loss to optimize the framework from three aspects (*i.e.,* color, texture, and content) to produce more visually pleasing results. Inspired by bilateral grid processing, Gharbi *et al.* [6] made real-time image enhancement possible, which dynamically generates the image transformation based on local and global information. To deal with low-light image enhancement, Wang *et al.* [10] established a large-scale under-exposed image dataset and learned an image-to-illumination mapping based on the Retinex model to enhance extremely low-light images.

However, these methods follow the route of *fully supervised* learning relying on large-scale datasets with *paired low/high*-quality images. First, paired data is usually expensive, and sometimes it takes a lot of effort and resources to build the dataset by professional photographers. Second, the judgment of image quality is usually closely related to the personality, aesthetics, taste, and experience of a person. “There are a thousand Hamlets in a thousand people’s eyes.” In other words, everyone has his/her different attitude towards the quality of the photography. To demonstrate this, a typical low-quality photo in MIT-Adobe FiveK Dataset [11] and its corresponding five high-quality versions retouched by five different experts in photo beautification, are shown in Fig. 1, respectively. It can be observed that the images processed by one expert are very different from the image retouched by another expert. Consequently, it is impractical to create a large-scale dataset with paired low and high-quality images to meets the preference of everyone. On the contrary, a more feasible way is to express the personal preferences of a user by providing a set of image collections that he/she loves. Therefore, an urgent demand is needed to build an enhancement model to learn the enhancement mapping from the low-quality dataset to a high-quality one even without the specific paired images. In this way, we can get rid of the burden of creating one-to-one paired data and rely only on the *target* dataset with the desired characteristics preferred by someone.

Benefit from the development of generative adversarial learning [12]–[14] and reinforcement learning (RL) [15], some works make attempts to handle the image enhancement tasks only with the help of unpaired data. The milestone work of transferring image style between unpaired data is CycleGAN [12]. It employs two generators and two discriminators and uses cycle consistency loss to achieve visually impressive results. Chen *et al.* [13] proposes to construct a bi-directional GAN with three improvements to transfer low-quality images into corresponding high-quality ones, and the experimental results show that this model is significantly better than CycleGAN. Hu *et al.* [15] design the first RL-based framework to train an effective photo post-processing model. Jiang *et al.* [16] carry out the first study on the task of low-light enhancement with an unsupervised framework. The method applies a self-regularized attention generator and dual discriminators to guide the generator globally and locally.

Rather than utilizing a *cyclic* generative adversarial network (GAN) to learn bi-directional mappings between the low-quality photos and high-quality ones, we build a *unidirectional* GAN to address the image *aesthetic* quality enhancement task, called the unsupervised image enhancement GAN (UEGAN). Inspired by the properties as mentioned above, our network consists of a joint global and local generator and a multi-scale discriminator with effective constraints. 1) The generator consists of an encoder and decoder with a *global attention module* and a *modulation module* embedded, which adjusts the features at different scales locally and globally. The *multi-scale* discriminator also inspects the results at different levels of granularity and guides the generator to produce better results to obtain global consistency and finer details. 2) To keep the content invariance, a *fidelity loss* is introduced to regularize the consistency between the input content and resulting content. 3) The global features extracted from the entire image reveal high-level information such as lighting conditions and color distributions. To capture these properties, the *global attention module* is designed to adjust pixels according to the information of a local neighborhood to meet both local adaptivity and global consistency. 4) For preventing over-enhancement, an *identity loss* is introduced to constrain the consistency between the enhanced result of the input high-quality image and the input one. This benefits controlling the enhancement procedure to be more quality-free and thus prevents over-enhancement. The main contributions of this work are summarized as follows:

- We design a single GAN framework that gets rid of the needs of *paired* training data for image *aesthetic* quality enhancement. To the best of our knowledge, this is the first trial to employ a *unidirectional* GAN framework to apply *unsupervised* learning to enhance the aesthetic quality of images (instead of low-light image enhancement).
- We propose a *global attention module* and a *modulation module* to construct the joint global and local generator to capture global features and adaptively adjust the local features. Together with the proposed multi-scale discriminator to inspect the quality of the generated results at different scales, well-enhanced results in perception and aesthetics are produced with both global consistency and finer details.
- We propose to jointly use *quality loss*, *fidelity loss*, and *identity loss* to train our model to make it

towards extracting quality-free features and controlling the enhancement procedure to be more robust to the quality change. Thus, our method can obtain more reasonable results and prevent over-enhancement. Extensive experimental results on various datasets demonstrate the superiority of the proposed model quantificationally and qualitatively.

The remaining of this paper is organized as follows. In Section II, the related work is succinctly described. In Section III, the proposed unsupervised image aesthetic quality enhancement model is presented in detail. In Section IV, extensive experimental results of the proposed are reported. In Section V, the ablation studies and analysis are presented. Finally, Section VI draws the conclusion.

II. Related Work

A. Traditional Image Enhancement

Extensive research has been conducted over the past few decades to improve the quality of photos. Most existing conventional image enhancement algorithms aim to stretch contrast and improve sharpness. The following three types of approaches are the most representative: *histogram adjustment*, *unsharp masking*, and *Retinex-based approaches*. These approaches are succinctly described as follows.

1) Histogram Adjustment: Based on the basic idea of mapping the luminance histogram to a specific distribution, many methods estimate the mapping function based on the statistical information of the entire image [2], [4], [17], while the details usually tend to be over-enhanced due to the dominance of some high-frequency information. Instead of estimating a single mapping function for the entire image, other approaches dynamically adjust the histogram based on *local* statistical information [3], [5], [18]. However, higher computational complexity limits the applicability of this method.

2) Unsharp Masking: Unsharp masking (UM) aims to improve image sharpness [19]. The framework of the UM approach can be summarized into the following two phases: First, the input image is decomposed into a *base layer* and a *detail layer* by applying a low or high pass filter. Second, all pixels in the detail layer are scaled by a *single* global weighting factor, or different pixels are adaptively scaled by pixel-wise weighting factors, and then added back to the base layer to obtain an enhanced version. Various works have been proposed to improve the performance of UM from two aspects: 1) design a more reasonable layer decomposition method to decouple different frequency bands [20], [21]; and 2) propose a better estimation algorithm for the adjustment scaling factor [19], [22].

3) Retinex-Based Approaches: Many researchers are working on Retinex-based image enhancement due to clear physical meaning. The basic assumption of the Retinex model is that the observed photo can be decomposed into reflection and illumination [23]. The enhanced image depends on the decomposed layer, *i.e.*, illumination and reflectance layers. Therefore, the Retinex-based model is usually approached as an illumination estimation problem [24]–[26]. However, such approaches might generate unnatural results due to the ambiguity and difficulty in accurately estimating the illumination and reflection map.

B. Learning-Based Image Enhancement

1) Supervised Learning Approaches: Given the explosive growth of CNN, image enhancement models based on learning methods have emerged in large numbers with impressive results. Yan *et al.* [8] took the first step in exploring the use of CNN for photo editing. Ignatov *et al.* [7] build a large-scale DSLR Photo Enhancement Dataset (*i.e.*, DPED), which consists of 6K photos captured simultaneously by a DSLR camera and three smartphones, respectively. With the paired data, it is easy to learn a mapping function between the low-quality photos captured by smartphones and the high-quality photos captured by the professional DSLR camera. Ren *et al.* [9] proposed a hybrid framework to address the low-light enhancement problem by jointly considering the content and structure. However, the promising performance of these models is inseparable from the premise of a large number of pairs of degraded images and corresponding high-quality counterparts.

2) Unsupervised Learning Approaches: Different from super-resolution, deraining, and denoising, the high-quality images are usually already present, and their low-quality versions can be easily generated by degrading them. In most cases, the image enhancement requires generating high-quality counterparts from low-quality images if need paired low-/high-quality during the training phase. High-quality photos are usually obtained by experts using professional photo editing programs (*i.e.*, Adobe Photoshop and Lightroom) to retouch low-quality photos. This is expensive, time-consuming, and the editing style might depend heavily on the expert rather than the real users. In order to get rid of paired training data, a few works attempted to address the image enhancement issue with unsupervised learning. Inspired by the well-known CycleGAN [12], Chen *et al.* [13] designed a dual GAN model to learn a bi-directional mapping between the source domain and target domain. Specifically, the learned transformation from the source domain to the target domain is first used to generate the high-quality image, and then the inverse mapping from the target domain to the source domain is learned to translate the generated high-quality image back to the source domain. The cycle consistency loss is constrained to enforce the closeness between the input low-quality photos and those generated by the reverse translation. The cycle consistency works well if both bi-directional generators provide an ideal mapping between the two domains. However, the instability of GAN increases training difficulty and risk to local minima when the cycle consistency is applied.

III. Proposed Unsupervised GAN for Image Enhancement

A. Motivations and Objectives

We observe that professional photographer usually follows these instincts when performing image editing:

- *Combination of global transformation and local adjustment*. The content and intrinsic semantic information should be kept the same between the low-quality and retouched versions. The expert might first perform a global transformation based on the overall lighting conditions (*e.g.*, well-exposure or under/over-exposure) and tone (*e.g.*, cool or warm colors) in the scenes. The local corrections then make finer adjustments based on the joint consideration of both global information and local content.
- *Over-enhancement prevention*. The trade-off between *fidelity* and *quality* is crucial. *Over-enhancement* donates the visual effects caused by excessively enhancing the properties of images related to the aesthetic feeling, such as very warm colors, high contrast, and over-exposure, *etc*. However, this can also make the results to deviate from fidelity and produce unnatural results. That is, a good automatic photo enhancer should be aware of over-enhancement while producing good visual effects.

Base on the observations mentioned above, we are dedicated to learning an *image-to-image* mapping function $\mathscr{F}$ to generate the high-quality counterpart x_g of a given low-quality photo x_l, which can be modeled as follows,

$$x_g = \mathscr{F}(x_l). \tag{1}$$

One critical issue in image enhancement tasks is how to define quality as well as high quality. Any user can easily provide a collection of images expressing their personal preferences without explicitly stating the quality he/she loves. Therefore, rather than defining $\mathscr{F}$ as various clearly defined rules, it is better to formulate it as a process of transforming low-quality image distribution under the guidance of the desired high-quality image distribution. This promotes us to learn a *user-oriented* photo enhancer based on *unpaired* data in an unsupervised manner. Based on this consideration, we make efforts in utilizing the set-level supervision of GAN to achieve our goals through adversarial learning.

B. Network Architecture

1) Joint Global and Local Generator: The generator plays a crucial role in our proposed UEGAN as it directly affects the quality of the final generated photos. The expert might first perform a global transform based on the overall lighting conditions or tone in the scenes. Therefore, the global features act as an image prior to guiding the generation and adjusting the local features. Based on this observation, we first propose a *global attention module* (GAM) to exploit the global attention of local features. Each channel of feature maps is extracted from the local neighborhood by the convolution layer. The focus of global attention is the 'holistic' understanding of each channel. In order to model the global attention of the intermediate features $z \in \mathbb{R}^{C\times H\times W}$, our proposed method can be summarized as the following three steps as shown in Fig. 3: 1) extracting global statistics information $f^{\mathrm{m}}_{\mathrm{pool}}(\cdot)$ of each channel via Eqn. (2); 2) digging the inter-channel relationship ρ using the extracted g_{mean} via the multi-layer perceptron $f_{\mathrm{FC}}(\cdot)$ in Eqn. (3); 3) fusing global and local features via Eqn. (4).

$$g_{\mathrm{mean}} = f^{\mathrm{m}}_{\mathrm{pool}}(z), \tag{2}$$

$$\rho = f_{\mathrm{FC}}(g_{\mathrm{mean}}), \tag{3}$$

$$\hat{z} = \mathrm{Conv}(C(E(\rho), z)), \tag{4}$$

where $f^{\mathrm{m}}_{\mathrm{pool}}(\cdot)$ means the average pooling operation, $f_{\mathrm{FC}}(\cdot)$ is two fully-connected layers, $E(\cdot)$ represents expanding the spatial dimension of ρ to that of z, $C(\cdot)$ is the concatenation operation, and $Conv(\cdot)$ is a convolution layer.

Fig. 2 shows the proposed *modulation module* (MM) in the joint global and local generator. In particular, we use skip connections between encoder and decoder at different scales locally and globally to prevent the information loss caused by resolution change. Unlike traditional U-Net [27], the features of the encoder are concatenated to those of the symmetric decoder at each stage (*i.e.*, four stages in our model). Our proposed modulation module learns to generate two branches of features and then merge them together with the multiplication operation. In our model, to further reuse the features, the learned modulation layer multiples the features of the first stage of the encoder and those of the penultimate layer by element-wise multiplication. Learning global features and feature modulation can effectively enhance the visual effect of the resulting image. The global features can also guide to penalize some low-quality features that might lead to visual artifacts or poorly reconstructed details. Complex image processing can be approximated by a set of simple local smoothing curves [28], the proposed joint global and local generator G is more capable than traditional U-Net for learning complex mappings from low-quality images to high-quality ones.

2) Multi-Scale Discriminator: In order to distinguish between real high-quality image and generated "pseudo" high-quality image, the discriminator requires a large receptive field to capture the global characteristics. This directly leads to the need for deeper networks or larger convolution kernels. The last layer of the discriminator usually captures the information from a larger region of the image and can guide the generator to produce the image with better global consistency. However, the intermediate layer of the discriminator with a smaller receptive field can force the generator to pay more attention to finer details. Based on this observation, as shown in Fig. 2, we propose a multi-scale discriminator D that uses multi-scale features to guide the generator to produce images with both global consistency and finer details.

C. Loss Function

1) Quality Loss: We use quality loss to adapt the distribution of enhanced results to that of high-quality images. The quality loss guides the generator to produce more visually pleasing results. In the previous GAN frameworks, the discriminator aims at distinguishing between real samples and the generated ones. However, we observe that simply applying the discriminator D to separate generated images and real high-quality images is not enough to obtain a good generator that transfers low-quality images into high-quality ones. The

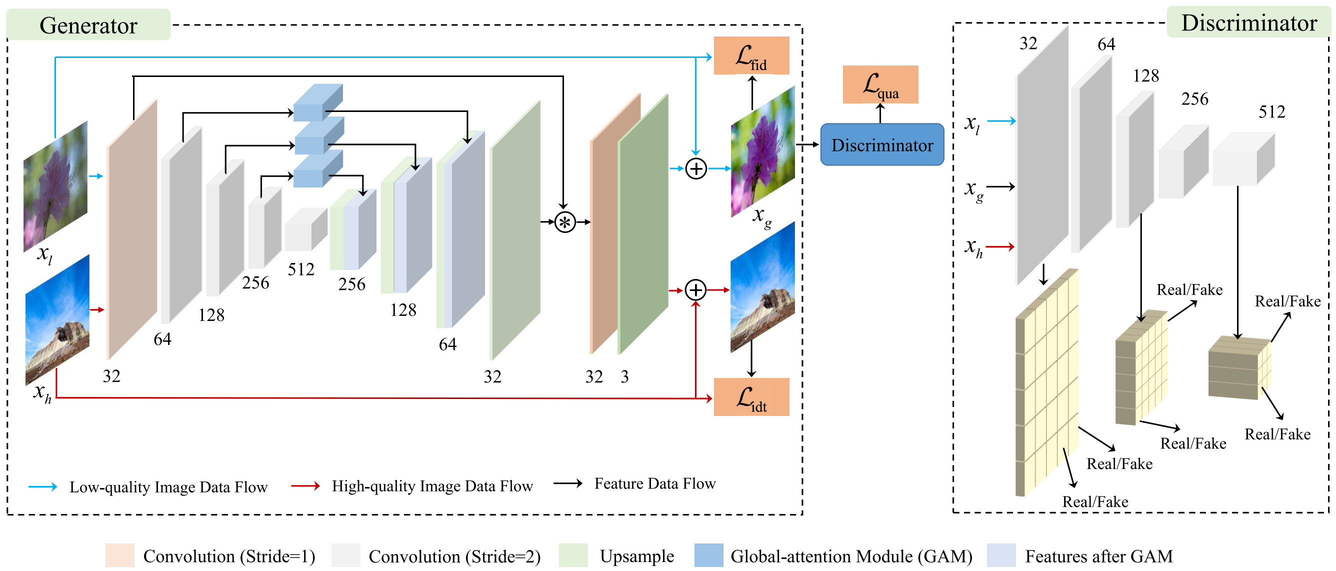


Fig. 2. The framework of the proposed UEGAN for image enhancement. The blue, red, and black lines indicate the low-quality images data flow, high-quality images data flow, and features data flow, respectively. The generator only inputs low-quality or high-quality images at a time.

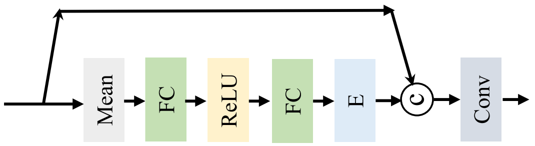


Fig. 3. The structure of the global attention module, where E and C denote the expanding and concatenation operations, respectively.

reason might be lies in that the quality ambiguity between low/high-quality images, some images in the low-quality image set are better than those in the high-quality image set. To address this issue, we also train the discriminator to distinguish between real low-quality images and real high-quality images as shown in Fig. 2.

Specifically, our proposed discriminator is based on the recently proposed relativistic discriminator structure [29], which not only assesses the probability that the real data (*i.e.*, real high-quality image) is more authentic than the fake data (*i.e.*, generated high-quality image or real low-quality image), but also guides the generator to produce high-quality images more realistic than real high-quality images. In addition, we employ an improved form of the relativistic discriminator, Relativistic average HingeGAN (RaHingeGAN) [29], [30] as follows:

$$\begin{aligned}\mathcal{L}^{D} = {} & \mathbb{E}_{x_l\sim P_l}\left[\max(0, 1 + (D(x_l) - E_{x_h\sim P_h} D(x_h)))\right]\\ & + \mathbb{E}_{x_h\sim P_h}\left[\max(0, 1 - (D(x_h) - E_{x_l\sim P_l} D(x_l)))\right]\\ & + \mathbb{E}_{x_g\sim P_g}\left[\max(0, 1 + (D(x_g) - E_{x_h\sim P_h} D(x_h)))\right]\\ & + \mathbb{E}_{x_h\sim P_h}\left[\max(0, 1 - (D(x_h) - E_{x_g\sim P_g} D(x_g)))\right], \quad (5)\\ \mathcal{L}^{G}_{\text{qua}} = {} & \mathbb{E}_{x_h\sim P_h}\left[\max(0, 1 + (D(x_h) - E_{x_g\sim P_g} D(x_g)))\right]\\ & + \mathbb{E}_{x_g\sim P_g}\left[\max(0, 1 - (D(x_g) - E_{x_h\sim P_h} D(x_h)))\right], \quad (6)\end{aligned}$$

where x_l, x_h, and x_g denote the real low-quality image, real high-quality image, and generated high-quality image, respectively.

2) Fidelity Loss: Since we train our model for image enhancement in an unsupervised manner, the quality loss itself might not ensure that the generated image has similar content to that of the input low-quality image. The simplest way is to measure the distance between the input and output images in the pixel domain. However, we cannot employ this strategy because the generated high-quality image is typically different from the input low-quality image in the pixel domain due to contrast stretching and color rendering. Therefore, we use fidelity loss to constrain the training of the generator, so as to achieve the purpose of generated high-quality images and inputting low-quality images with similar content. The fidelity loss is defined as the ℓ_2 norm between the feature maps of the input low-quality image and those of the generated high-quality images extracted by the pre-trained VGG network [31] as follows:

$$\mathcal{L}_{\text{fid}} = \sum\nolimits_{j=1}^{J}\{\mathbb{E}_{x_l\sim P_l}\left[\left\|\phi_j(x_l) - \phi_j(G(x_l))\right\|_2\right]\}, \quad (7)$$

where $\phi_j(\cdot)$ indicates the process of extracting the feature maps obtained by the j^{th} layer of the VGG network and J is the total number of layers used. Specifically, the *Relu*_1_1, *Relu*_2_1, *Relu*_3_1, *Relu*_4_1, and *Relu*_5_1 layers of VGG-19 network are adopted in this work.

3) Identity Loss: The identity loss is defined as ℓ_1 distance between the input high-quality image and the corresponding output of the generator G as follows:

$$\mathcal{L}_{\text{idt}} = \mathbb{E}_{x_h\sim P_h}\left[\|x_h - G(x_h)\|_1\right]. \quad (8)$$

The identity loss is calculated based on high-quality input images. Therefore, if the color distribution and contrast of the input image meet the characteristics of the high-quality image set, the identity loss intends to encourage preservation of the color distributions and contrast between the input and output. It ensures that the generator should make almost no changes to the image in content, contrast, and color during the image enhancement process. As a result, the identity loss makes it possible to simultaneously maintain the content, color rendering, and contrast of the input high-quality image.

3) Total Loss: By jointly considering *quality loss*, *fidelity loss*, and *identity loss*, our final loss is defined as the weighted sum of these losses which as follows:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{qua}}\mathcal{L}_{\text{qua}}^{G} + \lambda_{\text{fid}}\mathcal{L}_{\text{fid}} + \lambda_{\text{idt}}\mathcal{L}_{\text{idt}}, \tag{9}$$

where λ_{qua}, λ_{fid}, and λ_{idt} are weighting parameters to balance the relative importance of $\mathcal{L}_{\text{qua}}^{G}$, $\mathcal{L}_{\text{fid}}$ and $\mathcal{L}_{\text{idt}}$.

IV. Experimental Results

A. Dataset

1) MIT-Adobe FiveK Dataset: This dataset was constructed by Bychkovsky *et al.* [11] for the image enhancement task, where high-quality images are generated by experts retouching. It consists of 5000 raw photos and 25,000 retouched photos generated from those raw photos by five experienced photographers. Therefore, this dataset includes five subsets, each with 5,000 raw and corresponding retouched photo pairs. Following the works in [13], [15], we select the retouched photos generated by photographer C as the target photos (*i.e.*, ground truth) since the user rates this subset best. In order to generate unpaired training data, the subset is randomly divided into three partitions: 1) the first partition has 2,250 raw photos as low-quality input; 2) the second partition consists of retouched version of another 2,250 raw photos and served as the desired high-quality photos; 3) the last partition is the remaining 500 raw photos used for validation (100 images) and testing (400 images). These three parts have no overlaps with each other.

2) Flickr Dataset: In addition to training on the photographer results of the MIT-Adobe FiveK Dataset, we also collected a high-quality image collection from Flickr for unpaired training. These images are crawled from the Flickr images tagged with "High Dynamic Range" to ensure relatively consistent quality and then manually selected by the authors. Finally, we select 2,000 images as the desired high-quality labels.

B. Implementation Details

We built our network in Pytorch and train it for 150 epochs on an NVidia GeForce RTX 2080 Ti GPU with a mini-batch size of 10. The entire network is optimized from scratch using Adam optimizer [32] with a learning rate of 0.0001. The leaning rate is fixed at the first 75 epochs and then linearly decays to zero in the next 75 epochs. For the MIT-Adobe FiveK Dataset, we use Lightroom to decode the images into the png format and resize the long side of the images to 512 resolution. For data augmentation, we randomly cropped 256 × 256 patches from images.

In the MIT-Adobe FiveK Dataset, we set the hyper-parameters λ_{qua}, λ_{fid}, and λ_{idt} as 0.05, 1, and 0.1, respectively as empirically these values provide the best performance in quantitative and qualitative performance. When coming to the Flickr Dataset, the hyper-parameters λ_{qua}, λ_{fid}, and λ_{idt} are also empirically set as 0.05, 1 and 0.1.

TABLE I

Quantitative Comparison Between Our Proposed Method and State-of-the-Art Methods on MIT-Adobe FiveK Dataset [11]

Method	PSNR	SSIM	NIMA
Input	17.42	0.8037	4.46
CycleGAN [12]	20.72	0.7825	4.37
Exposure [15]	19.74	0.8442	4.62
EnlightenGAN [16]	16.96	0.7562	4.25
DPE [13]	22.36	0.8674	4.54
Ours	**22.88**	**0.8882**	**4.76**

C. Evaluation Metrics

The most commonly-used full-reference image quality assessment metrics (*i.e.*, PSNR and SSIM) focus only on *signal fidelity* but may not accurately reflect aesthetic and perceptual quality. Although the evaluation of aesthetic quality is challenging, we still have the tool to measure the enhancement quality to an extent with the quantitative evaluation. To this end, the NIMA [33] score is used to quantify the aesthetic quality. The NIMA is an effective CNN-based image aesthetic quality assessment method trained on the large-scale aesthetic dataset AVA [34]. It is predicts the distribution of human opinion scores rather than the mean opinion scores (*i.e*, MOS). Therefore, we use PSNR, SSIM, and NIMA to compare our proposed method with the state-of-the-art methods at the pixel level, structural level, and aesthetics level, where the first two metrics are performed in terms of the similarity between the enhanced results and the corresponding expert-retouched (*i.e.*, ground truth). In general, higher PSNR, SSIM and NIMA values correspond to reasonably better results.

D. Quantitative Comparison

Most previous methods for automatic photo quality enhancement are based on supervised learning that requires paired data [6], [7], [9]–[11]. Recently, a series of works based on GANs or reinforcement learning (RL) attempted to use only unpaired data to solve this tasks. We compared our proposed method with CycleGAN [12], and three unpaired photo enhancement methods: Deep Photo Enhancer (DPE) [13], EnlightenGAN [16] and Exposure [15]. CycleGAN, DPE, and EnlightenGAN are GAN-based methods and Exposure is an RL and filter-based method.

Table I lists the quantitative comparison results of various models on MIT-Adobe FiveK dataset [11]. In this table, the best performance of each evaluation metric (*i.e.*, PSNR, SSIM, and NIMA) is boldfaced in black. Please note that the program codes of all models under comparison are downloaded from the link provided by the corresponding authors. Specifically, we used the codes provided by the corresponding authors to retrain the CycleGAN and EnlightenGAN on MIT-Adobe FiveK dataset. We test the Exposure and DPE using the models pre-trained on MIT-Adobe FiveK dataset provided by the corresponding authors, because it achieved better performance than our retrained model. Besides, the Flickr

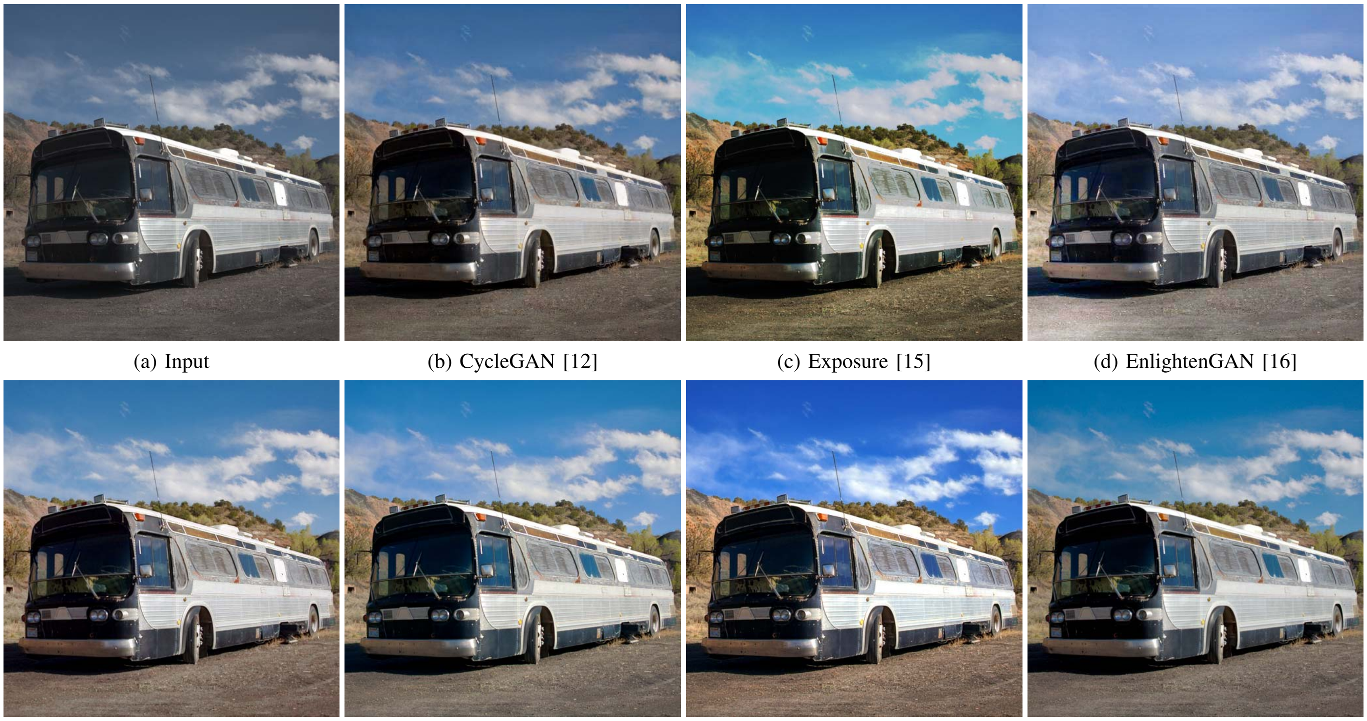


Fig. 4. Visual quality comparison with state-of-the-art methods (*i.e.*, CycleGAN, DPE, EnlightenGAN, and Exposure) on a test image from the MIT-Adobe FiveK [11] dataset.

dataset we collected has no ground truth, thus we can only perform qualitative experiments on it. From Table I, one can observe that our proposed UEGAN achieves the best performance in terms of PSNR, SSIM, and NIMA compared with other state-of-the-art image quality enhancement methods trained with unpaired data.

From the experimental results listed in Table I, the following conclusions can be drawn. 1) Our proposed UEGAN, DPE, and Exposure are ranked in the top three in the quantitative comparison and are superior to inputs on all three evaluation metrics. Specifically, our proposed UEGAN has consistently achieved the best performance. 2) Compared with the input, CycleGAN has been obtained worse performance in SSIM and NIMA, which is mainly due to the existence of blocking artifacts in the generated results. 3) Similar to CycleGAN, EnlightenGAN even performed worse on all three evaluation metrics than the input, which may be caused by significant changes in contrast.

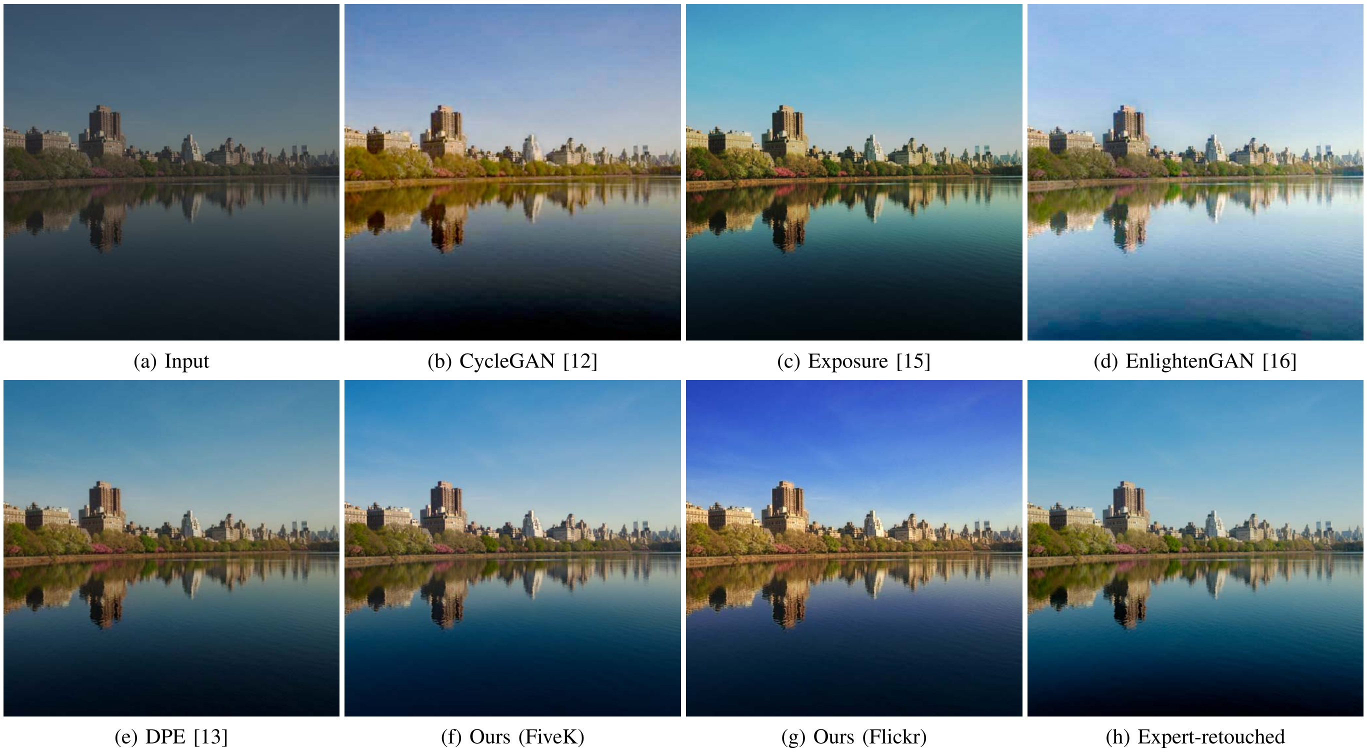
E. Qualitative Comparison

Besides the superiority in quantitative evaluation, our proposed UEGAN method is also superior to other enhancement methods in qualitative comparison. As shown in Fig. 4-7, four representative test images were selected from MIT-Adobe FiveK dataset for conducting visual comparisons. One can observe that the input images are diverse and challenging, including: 1) Fig. 4 (a) is an outdoor scene with normal lighting condition; 2) Fig. 5 (a) is a landscape image with under-exposed lake surface and buildings; 3) Fig. 6 (a) is a sky scene with a tiny airplane; 4) Fig. 7 (a) is a globally under-exposed outdoor scene with little portrait details. Compared to their respective expert retouched versions shown in Fig. 4 (d) - Fig. 7 (d), all input images have significantly worse visual experiences. *Additional results are provided in the supplementary material.*

As shown, we obtain some interesting insights. First, the proposed UEGAN trained on our collected Flickr dataset shows the best visual quality among all methods as it generates vivid colors and clear textures. Besides, the results of our proposed UEGAN trained on MIT-Adobe FiveK dataset are satisfactory in enhancing the input image. Second, CycleGAN is less effective in generating vivid colors and also leads to blocking artifacts, which degrade the image quality. In contrast, our method generates visually pleasing results with clear details and sharp structures. Third, Exposure is a filter-based method, which tends to produce over-saturated results and falsely remove textures. However, the results of our proposed UEGAN look natural with good color rendition. Fourth, EnlightenGAN significantly changes the contrast but makes the resulting images dull. On the contrary, our proposed UEGAN can generate satisfactory contrast and natural appearance with the appropriate saturation. Last, DPE produces competitive results compared with ours in structure and contrast enhancement, while it may generate unrealistically looking results. In short, our proposed UEGAN generates natural and pleasing results with satisfactory contrast, vibrant colors and clear details, which is superior to the state-of-the-art methods compared and comparable to the corresponding expert-retouched results.

F. User Study

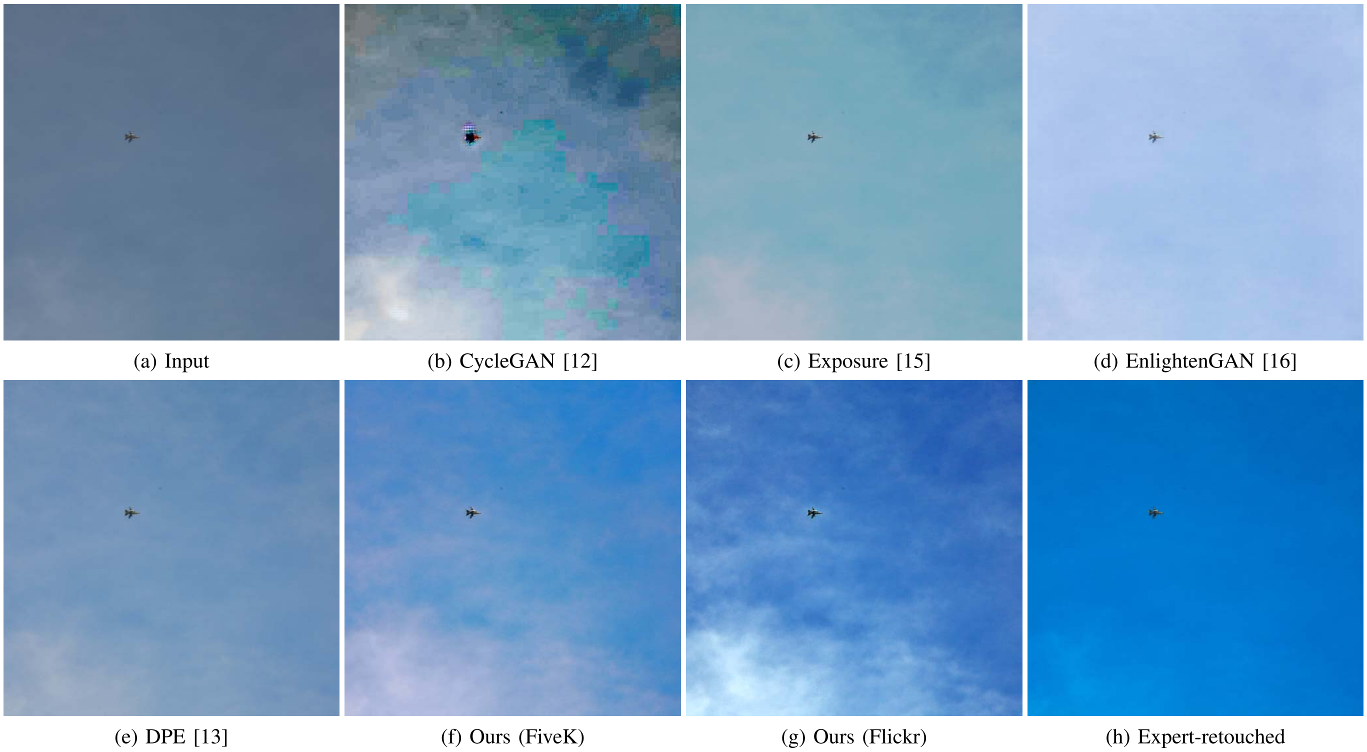
Our ultimate goal is to learn the implicit characteristics of the target domain to generate high-quality images with



(a) Input (b) CycleGAN [12] (c) Exposure [15] (d) EnlightenGAN [16]

(e) DPE [13] (f) Ours (FiveK) (g) Ours (Flickr) (h) Expert-retouched

Fig. 5. Visual quality comparison with state-of-the-art methods (*i.e.*, CycleGAN, DPE, EnlightenGAN, and Exposure) on a test image from the MIT-Adobe FiveK [11] dataset.

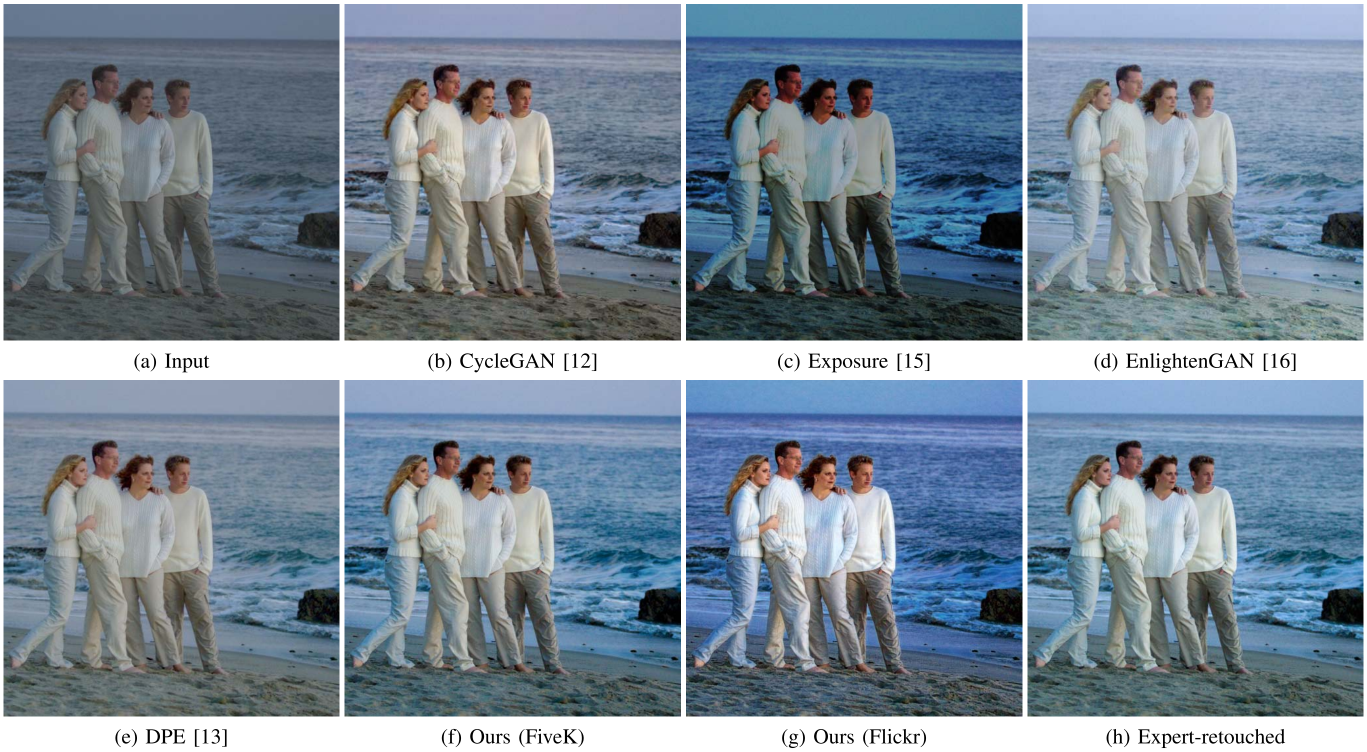


(a) Input (b) CycleGAN [12] (c) Exposure [15] (d) EnlightenGAN [16]

(e) DPE [13] (f) Ours (FiveK) (g) Ours (Flickr) (h) Expert-retouched

Fig. 6. Visual quality comparison with state-of-the-art methods (*i.e.*, CycleGAN, DPE, EnlightenGAN, and Exposure) on a test image from the MIT-Adobe FiveK [11] dataset.

similar properties. To measure the subjective quality, we have performed a user study with 28 participants and 40 image sets (*e.g.*, each image set contains 1 test image and the corresponding six generated versions) using pairwise comparisons on six methods (including two versions of our method). The participants are asked to choose his/her favorite result



(a) Input (b) CycleGAN [12] (c) Exposure [15] (d) EnlightenGAN [16]

(e) DPE [13] (f) Ours (FiveK) (g) Ours (Flickr) (h) Expert-retouched

Fig. 7. Visual quality comparison with state-of-the-art methods (*i.e.*, CycleGAN, DPE, EnlightenGAN, and Exposure) on a test image from the MIT-Adobe FiveK [11] dataset.

TABLE II

THE PAIRWISE COMPARISON PREFERENCE MATRIX IN USER STUDY. EG DENOTES ENLIGHTENGAN

	Input	CycleGAN [12]	Exposure [13]	EG [16]	DPE [15]	Ours (FiveK)	Ours (Flickr)	Total
Input	-	387	157	171	78	31	16	840
CycleGAN	733	-	185	263	95	74	32	1382
Exposure	963	935	-	692	328	253	141	3312
EG	949	857	428	-	223	153	92	2702
DPE	1042	1025	792	897	-	401	214	4371
Ours (FiveK)	1089	1046	867	967	719	-	327	5015
Ours (Flickr)	1104	1088	979	1028	906	793	-	5898

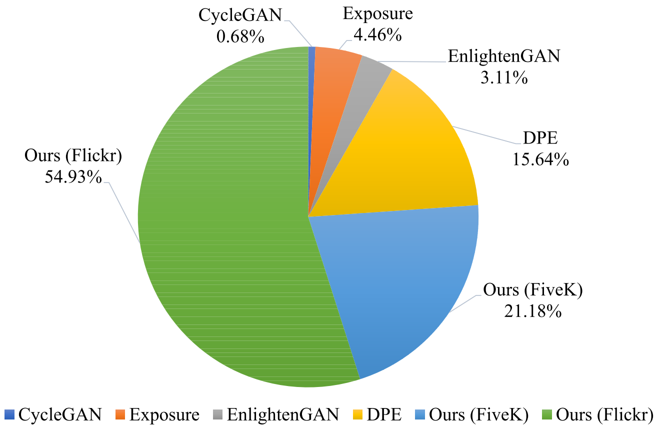


Fig. 8. User preference results of different aesthetic quality enhancement algorithms.

from the displayed pair and the generated images are presented randomly to avoid subjective bias. The corresponding pairwise comparison results are shown in Table II, where each figure indicates the number of times the method in that row outperforms the method in that column. It can be seen that, in all cases, the results of DPE and our proposed UEGAN are preferred much more frequently than the results of other models (*i.e.*, CycleGAN, Exposure, and EnlightenGAN). Among all the comparison methods, the preferred percentages of the proposed UEGAN trained on MIT-Adobe FiveK Dataset [8] over CycleGAN, Exposure, EnlightenGAN, and DPE are respectively 93.39%, 77.41%, 86.34%, and 64.20%, and the preferred percentages of our UEGAN trained on our collect Flickr dataset compared with CycleGAN, Exposure, EnlightenGAN, and DPE are 97.14%, 87.41%, 91.78%, and 80.90%, respectively. It can be seen that the proposed model is selected more frequently than the compared models, which means that the proposed UEGAN can produce more visually pleasing results than all state-of-the-art models in the comparison.

To measure the overall quality, we again randomly selected 100 test images and the corresponding 100 generated results for each model. Each time, six enhanced versions of a test image are present randomly to the participants and asked them to select their favorite one. Finally, 2800 subjective votes are obtained in total and the results are shown in Fig. 8.

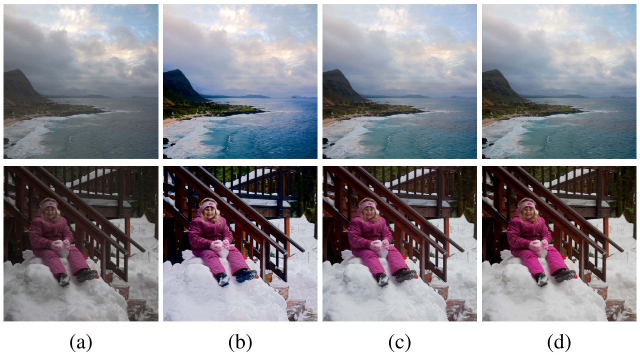


Fig. 9. Visual quality comparison results of our proposed UEGAN trained with different loss. (a) Inputs. (b) $\mathcal{L}_{\text{qua}}^{G} + \mathcal{L}_{\text{fid}}$. (c) $\mathcal{L}_{\text{qua}}^{G} + \mathcal{L}_{\text{fid}} + \mathcal{L}_{\text{idt}}$. (d) Expert-retouched (*i.e.*, Ground Truth).

TABLE III
AVERAGE PSNR, SSIM, AND NIMA RESULTS OF ENHANCED RESULTS ON MIT-ADOBE FIVEK DATASET [11]

Method	PSNR	SSIM	NIMA
Ours w/ $\mathcal{L}_{\text{qua}}^{G}$, w/ $\mathcal{L}_{\text{fid}}$, w/o $\mathcal{L}_{\text{idt}}$	22.56	0.8773	4.68
Ours w/ $\mathcal{L}_{\text{qua}}^{G}$, w/ $\mathcal{L}_{\text{fid}}$, w/ $\mathcal{L}_{\text{idt}}$	**22.88**	**0.8882**	**4.76**

The results show that the enhanced results obtained by our proposed UEGAN are preferred more frequently than those by other methods in the comparison. This further reveals that the proposed UEGAN is superior to all state-of-the-art models in improving the aesthetic quality of the photos.

V. ANALYSIS AND DISCUSSIONS

A. Ablation Studies

1) Loss Analysis: In this section, we study the effect of *quality loss*, *fidelity loss*, and *identity loss* quantitatively and qualitatively. Table III shows the PSNR, SSIM, and NIMA results achieved by using $\mathcal{L}_{\text{qua}}^{G} + \mathcal{L}_{\text{fid}}$ and $\mathcal{L}_{\text{qua}}^{G} + \mathcal{L}_{\text{fid}} + \mathcal{L}_{\text{idt}}$. We can observe that using only $\mathcal{L}_{\text{qua}}^{G} + \mathcal{L}_{\text{fid}}$ loss achieves better perfomance than the state-of-the-art DPE [13], EnlightenGAN [16] and Exposure [15], but adding identity loss could further improve quantization performance (*i.e.*, PSNR, SSIM, and NIMA). Fig. 9 shows two visual comparisons between the results of our proposed UEGAN trained with $\mathcal{L}_{\text{qua}}^{G} + \mathcal{L}_{\text{fid}}$ and $\mathcal{L}_{\text{qua}}^{G} + \mathcal{L}_{\text{fid}} + \mathcal{L}_{\text{idt}}$, respectively. It can be observed that the two results generated by our model are more visually pleasing than the input in Fig. 9 (a). However, compared with the ground truth in Fig. 9 (d), adding identity loss can suppress over-enhancement to some extent to produce more realistic colors and contrast, as shown in Fig. 9 (b) and (c).

Fig. 10 shows the results generated by our proposed UEGAN by fixing the weighting parameters of $\mathcal{L}_{\text{fid}}$ and $\mathcal{L}_{\text{idt}}$ at 1.0 and 0.1, respectively, and increasing that of $\mathcal{L}_{\text{qua}}^{G}$ from 0.05 to 0.4, respectively. We can observe that if we increase the weight of the $\mathcal{L}_{\text{qua}}^{G}$, the contrast becomes higher and the colors will be more vivid, but the result tends to be over-enhanced and thus loses fidelity. Therefore, we jointly consider fidelity loss, quality loss, and identity loss to improve the visual effect as much as possible while keeping the content the same and avoiding over-enhancement.

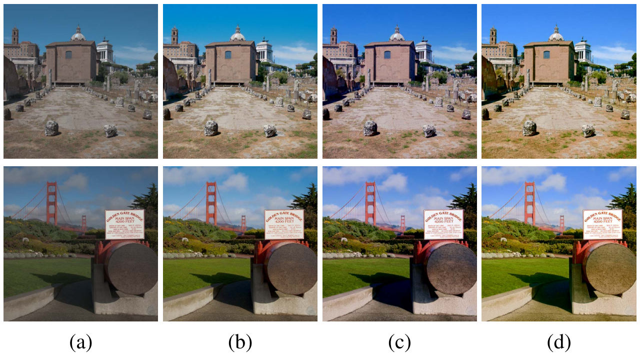


Fig. 10. Visual quality comparison results of fidelity loss vs. quality loss. (a) Inputs. (b) - (d) are results obtained by fixing the weighting parameters of $\mathcal{L}_{\text{fid}}$ and $\mathcal{L}_{\text{idt}}$ at 1.0 and 0.1, respectively, and setting $\mathcal{L}_{\text{qua}}^{G}$ to 0.05, 0.2, and 0.4, respectively.

TABLE IV
COMPARISON OF AVERAGE PSNR, SSIM, AND NIMA PERFORMANCE OF DIFFERENT NETWORK ARCHITECTURES ON MIT-ADOBE FIVEK DATASET [11]

Method	PSNR	SSIM	NIMA
GAM + U-Net	22.44	0.8756	4.66
GAM + MM-P	17.41	0.8037	4.46
UEGAN w/o GAM	22.54	0.8790	4.73
UEGAN w/o GAM and MM	21.87	0.8653	4.51
UEGAN	**22.88**	**0.8882**	**4.76**

2) Architecture Analysis: In this section, we investigate the effect of each individual component (*i.e.*, global attention module (GAM) and modulation module (MM)) in our proposed UEGAN described in Section III-B. We conduct ablation studies by comparing the proposed UEGAN with the following UEGAN variants: 1) GAM + U-Net: removing the MM and concatenating the features of the first stage of the encoder to those of the penultimate layer; 2) GAM + MM-P: we apply the MM at the pixel level. That is, the generator learns a modulation layer that multiplies the input image with the features of the last layer; 3) UEGAN w/o GAM: removing the GAM from the proposed generator. 4) UEGAN w/o GAM and MM: removing both the GAM and MM from the generator. The quantitative comparison results of all the different architectures are shown in Table IV. It can be observed that, compared with the traditional U-Net (*i.e.*, GAM+U-Net), our proposed UEGAN achieves the best improvements. Using MM at the feature level can significantly improve the performance than that at the pixel level (*i.e.*, GAM+MM-P). Both GAM or MM lead to better PSNR, SSIM, and NIMA, and combining them can further improve the quantitative performance to achieve the best.

B. Limitations

The proposed method is completely unsupervised and inevitably has limitations. A typical artifact that is present on

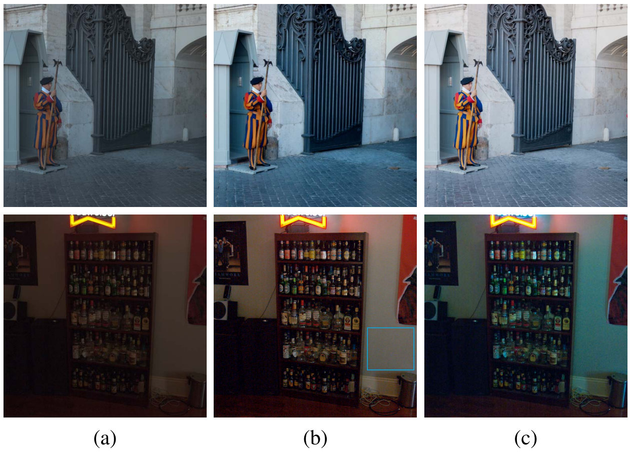


Fig. 11. Failure cases generated by our method compared with the ground truth. (a) Inputs. (b) Our results. (c) Ground truth.

the resulting image is color deviation. For example, the color of the ground of the second image in the first row of Fig. 11 is different from that of the input and ground truth. Even though they might produce more pleasing results sometimes coincidentally, this kind of adjustment changes the content and makes the results look unreal. In addition, as shown by the blue box in the second row of Fig. 11, our method cannot remove noise from the generated results. However, this kind of noise is common in under-exposed images.

VI. Conclusion

In this paper, we present an *unsupervised* deep *generative adversarial network* model developed for image enhancement, call the *Unsupervised image Enhancement GAN* (UEGAN). The proposed model is able to learn the corresponding *image-to-image* mapping from a set of images provided by public users with desired characteristics in an *unsupervised* manner, which makes it possible to learn a *user-oriented* automatic photo enhancer. We embed the global attention module (GAM) and modulation module (MM) into the generator to capture global features and adjust the features adaptively. In addition, we combine fidelity loss, quality loss, and identity loss with the proposed network to improve the visual quality of the enhanced results. The quantitative and qualitative experimental results show that our proposed method UEGAN is superior to the four state-of-the-art methods.

References

[1] T. Arici, S. Dikbas, and Y. Altunbasak, "A histogram modification framework and its application for image contrast enhancement," *IEEE Trans. Image Process.*, vol. 18, no. 9, pp. 1921–1935, Sep. 2009.

[2] G. Thomas, D. Flores-Tapia, and S. Pistorius, "Histogram specification: A fast and flexible method to process digital images," *IEEE Trans. Instrum. Meas.*, vol. 60, no. 5, pp. 1565–1578, May 2011.

[3] C. Lee, C. Lee, and C.-S. Kim, "Contrast enhancement based on layered difference representation of 2D histograms," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5372–5384, Dec. 2013.

[4] D. Coltuc, P. Bolon, and J.-M. Chassery, "Exact histogram specification," *IEEE Trans. Image Process.*, vol. 15, no. 5, pp. 1143–1152, May 2006.

[5] M. Abdullah-Al-Wadud, M. H. Kabir, M. A. A. Dewan, and O. Chae, "A dynamic histogram equalization for image contrast enhancement," *IEEE Trans. Consum. Electron.*, vol. 53, no. 2, pp. 593–600, May 2007.

[6] 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.

[7] A. Ignatov, N. Kobyshev, R. Timofte, and K. Vanhoey, "DSLR-quality photos on mobile devices with deep convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3277–3285.

[8] Z. Yan, H. Zhang, B. Wang, S. Paris, and Y. Yu, "Automatic photo adjustment using deep neural networks," *ACM Trans. Graph.*, vol. 35, no. 2, p. 11, 2016.

[9] W. Ren *et al.*, "Low-light image enhancement via a deep hybrid network," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4364–4375, Sep. 2019.

[10] R. Wang, Q. Zhang, C.-W. Fu, X. Shen, W.-S. Zheng, and J. Jia, "Underexposed photo enhancement using deep illumination estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 6849–6857.

[11] V. Bychkovsky, S. Paris, E. Chan, and F. Durand, "Learning photographic global tonal adjustment with a database of input/output image pairs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 97–104.

[12] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2223–2232.

[13] 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 Conf. Comput. Vis. And Pattern Recognit.*, Jun. 2018, pp. 6306–6314.

[14] C. H. Lin, C.-C. Chang, Y.-S. Chen, D.-C. Juan, W. Wei, and H.-T. Chen, "COCO-GAN: Generation by parts via conditional coordinating," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 4512–4521.

[15] Y. Hu, H. He, C. Xu, B. Wang, and S. Lin, "Exposure: A white-box photo post-processing framework," *ACM Trans. Graph.*, vol. 37, no. 2, p. 26, 2018.

[16] Y. Jiang *et al.*, "EnlightenGAN: Deep light enhancement without paired supervision," 2019, *arXiv:1906.06972*. [Online]. Available: http://arxiv.org/abs/1906.06972

[17] H. Ibrahim and N. P. Kong, "Brightness preserving dynamic histogram equalization for image contrast enhancement," *IEEE Trans. Consum. Electron.*, vol. 53, no. 4, pp. 1752–1758, Nov. 2007.

[18] J. A. Stark, "Adaptive image contrast enhancement using generalizations of histogram equalization," *IEEE Trans. Image Process.*, vol. 9, no. 5, pp. 889–896, May 2000.

[19] W. Ye and K.-K. Ma, "Blurriness-guided unsharp masking," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4465–4477, Sep. 2018.

[20] S. K. Mitra, H. Li, I.-S. Lin, and T.-H. Yu, "A new class of nonlinear filters for image enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1991, pp. 2525–2528.

[21] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, Jun. 2013.

[22] A. Polesel, G. Ramponi, and V. J. Mathews, "Image enhancement via adaptive unsharp masking," *IEEE Trans. Image Process.*, vol. 9, no. 3, pp. 505–510, Mar. 2000.

[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] 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.

[25] 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.

[26] Z. Ying, G. Li, Y. Ren, R. Wang, and W. Wang, "A new low-light image enhancement algorithm using camera response model," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Oct. 2017, pp. 3015–3022.

[27] 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*, 2015, pp. 234–241.

[28] J. Chen, A. Adams, N. Wadhwa, and S. W. Hasinoff, "Bilateral guided upsampling," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–8, Nov. 2016.

[29] 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

[30] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.

[31] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[32] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[33] H. Talebi and P. Milanfar, “NIMA: Neural image assessment,” *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.

[34] N. Murray, L. Marchesotti, and F. Perronnin, “AVA: A large-scale database for aesthetic visual analysis,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2408–2415.

Zhangkai Ni (Graduate Student Member, IEEE) received the M.E. degree in communication engineering from the School of Information Science and Engineering, Huaqiao University, Xiamen, China, in 2017. He is currently pursuing the Ph.D. degree with the Department of Computer Science, City University of Hong Kong, Hong Kong. He was a Research Engineer with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, from 2017 to 2018. His current research interests include computer vision, image processing, unsupervised learning, and quality assessment.

Wenhan Yang (Member, IEEE) received the B.S. and Ph.D. degrees (Hons.) in computer science from Peking University, Beijing, China, in 2012 and 2018, respectively. He was a Visiting Scholar with the National University of Singapore from 2015 to 2016. He is currently a Postdoctoral Research Fellow with the Department of Computer Science, City University of Hong Kong. His current research interests include deep-learning-based image processing, bad weather restoration, and related applications and theories.

Shiqi Wang (Member, IEEE) received the B.S. degree in computer science from the Harbin Institute of Technology in 2008, and the Ph.D. degree in computer application technology from Peking University in 2014. From March 2014 to March 2016, he was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. From April 2016 to April 2017, he was a Research Fellow with the Rapid-Rich Object Search Laboratory, Nanyang Technological University, Singapore. He is currently an Assistant Professor with the Department of Computer Science, City University of Hong Kong. He has proposed over 40 technical proposals to ISO/MPEG and ITU-T and AVS standards, and authored/coauthored more than 200 refereed journal/conference papers. He is the coauthor of an article that received the Best Student Paper Award in the IEEE International Conference on Image Processing (ICIP) 2018. His research interests include video compression, image/video quality assessment, and image/video search and analysis. He received the Best Paper Award of the IEEE Multimedia 2018, the IEEE International Conference on Multimedia and Expo (ICME) 2019, the IEEE International Conference on Visual Communications and Image Processing (VCIP) 2019, and the Pacific-Rim Conference on Multimedia (PCM) 2017.

Lin Ma (Member, IEEE) received the B.E. and M.E. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2006 and 2008, respectively, and the Ph.D. degree from the Department of Electronic Engineering, The Chinese University of Hong Kong, in 2013.

He was a Researcher with Huawei Noah’s Ark Laboratory, Hong Kong, from 2013 to 2016. He was a Principal Researcher with Tencent AI Laboratory, Shenzhen, China, from 2016 to 2020. He is a currently a Principal Researcher with Meituan-Dianping Group, Beijing, China. His current research interests include computer vision, multimodal deep learning, specifically for image and language, image/video understanding, and quality assessment.

Dr. Ma received the Best Paper Award from the Pacific-Rim Conference on Multimedia in 2008. He was a recipient of the Microsoft Research Asia Fellowship in 2011. He was a Finalist of the HKIS Young Scientist Award in engineering science in 2012.

Sam Kwong (Fellow, IEEE) received the B.Sc. degree in electrical engineering from the State University of New York, Buffalo, NY, USA, in 1983, the M.Sc. degree in electrical engineering from the University of Waterloo, Waterloo, ON, Canada, in 1985, and the Ph.D. degree from the University of Hagen, Germany, in 1996. From 1985 to 1987, he was a Diagnostic Engineer with the Control Data Canada, Mississauga, ON, Canada. He later joined Bell-Northern Research, Ottawa, ON, Canada, as a Member of Scientific Staff and as a Lecturer with the Department of Electronic Engineering, City University of Hong Kong (CityU), Hong Kong, in 1990. He is currently a Chair Professor with the Department of Computer Science, CityU. His research interests include video coding, pattern recognition, and evolutionary algorithms. He is also the Vice-President of conferences and meetings with the IEEE Systems, Man, and Cybernetics Society. He also serves as an Associate Editor for the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, and the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, and the *Journal of Information Science*.