Unpaired Image Enhancement with Quality-Attention Generative Adversarial Network

Zhangkai Ni
City University of Hong Kong
Hong Kong
eezkni@gmail.com

Wenhan Yang
City University of Hong Kong
Hong Kong
yangwenhan@pku.edu.cn

Shiqi Wang*
City University of Hong Kong
Hong Kong
shiqwang@cityu.edu.hk

Lin Ma
Meituan-Dianping Group
China
forest.linma@gmail.com

Sam Kwong*
City University of Hong Kong
Hong Kong
cssamk@cityu.edu.hk

Figure 1: Examples of our proposed QAGAN model trained on two different unpaired datasets. (*i.e.*, results of models trained using the MIT-Adobe FiveK dataset (red) and the crawled Flickr dataset (green).)

ABSTRACT

In this work, we aim to learn an unpaired image enhancement model, which can enrich low-quality images with the characteristics of high-quality images provided by users. We propose a quality attention generative adversarial network (QAGAN) trained on *unpaired* data based on the bidirectional *Generative Adversarial Network* (GAN) embedded with a *quality attention module* (QAM). The key novelty of the proposed QAGAN lies in the injected QAM for the generator such that it learns domain-relevant quality attention *directly* from the two domains. More specifically, the proposed QAM allows the generator to effectively select semantic-related characteristics from the spatial-wise and adaptively incorporate style-related attributes from the channel-wise, respectively. Therefore, in our proposed QAGAN, not only discriminators but also the generator can directly access both domains which significantly facilitate the generator to learn the mapping function. Extensive experimental results show that, compared with the state-of-the-art methods based on unpaired learning, our proposed method achieves better performance in both objective and subjective evaluations.

*Corresponding authors: Sam Kwong and Shiqi Wang

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413839

CCS CONCEPTS

• **Computing methodologies** → **Image processing; Computer vision; Unpaired image enhancement**.

KEYWORDS

Image enhancement, generative adversarial network (GAN); quality attention module

ACM Reference Format:
Zhangkai Ni, Wenhan Yang, Shiqi Wang, Lin Ma, and Sam Kwong. 2020. Unpaired Image Enhancement with Quality-Attention Generative Adversarial Network. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413839

1 INTRODUCTION

Image enhancement aims to improve the quality of images via various post-processing techniques, such as contrast enhancement, color rendition, and sharpening, to name a few.

Experienced photographers can freely generate their favorite visual pleasing images through the professional image-editing software (*e.g.*, Adobe Photoshop and Lightroom), which are desired by the general public who lacks professional image-editing skills. This contradiction highlights the importance of the *user-oriented* automatic image enhancement method for the general public to produce high-quality images they want. Furthermore, automatic image enhancement is already a build-in technology for displays, cameras, scanners, and photography applications to provide users with better customized services.

Significant progress has been made in improving the performance of automatic image enhancement methods [1, 5, 10, 20, 25, 28]. However, most of these models require a large-scale dataset with *paired* low/high-quality images in order to learn the enhancement mappings in a *full-supervised* manner. In general, such paired training images can either be manually edited by experienced photographers to generate corresponding high-quality counterparts from low-quality images [1], or both kinds of images are captured simultaneously and well-registered to ensure the pixel correspondence [10]. Numerous efforts have been devoted to collecting the low/high-quality training image pairs [1, 10, 25] to take advantage of data-driven learning to train automatic image quality enhancement models. Unfortunately, such supervised models have the following two drawbacks. On the one hand, it is difficult to collect paired training images in dynamic scenes with various motions. On the other hand, even if such paired images are collected, they cannot reflect the user preferences and cannot play the role in guiding the training/learning of the enhancement model. The user preference is a kind of property that supervised learning methods hardly possess. Since the perception of visual quality is a cognitive task that is highly relevant to the human personality, educational background, and aesthetic standard, whereas the quality of carefully prepared paired training images are usually highly dependent on the creator (*e.g.*, retoucher or photographer).

In this study, we make efforts in addressing the unpaired image enhancement task that only needs to be supervised by the unpaired image set (without any paired images) expressing personal user preferences. The nature of such unpaired learning makes it possible to easily create *personalized* training data for everyone to learn their own *individual-oriented* image enhancement model. Our proposed method is based on a bidirectional generative adversarial network (GAN) [6] between the source domain (*i.e.*, low-quality images domain) and the target domain (*i.e.*, high-quality images domain). However, the generator of most existing GAN-based models can only learn the features of the target domain under the guidance of discriminator because only the discriminator can access both domains. This increases the burden of learning the ideal mapping between different domains and also leads to instability in GAN training. To overcome this difficulty, we propose a *quality attention module* (QAM) which allows the generator to effectively select semantic-relevant characteristics from the spatial-wise and adaptively incorporate style-related natures from the channel-wise, respectively. In other words, in our proposed QAGAN model, not only the discriminator but also the generator can directly access both two domains, which significantly facilitates the generator to learn a more efficient mapping function that encodes richer context. This is the significant difference between our proposed QAGAN with the well-known CycleGAN [31] and the related follow-up works. In particular, we employ the VGG-19 network [22] as the encoder of the generator, such that the entire framework of our algorithm only needs to learn two decoders and two corresponding discriminators with unpaired images. The main contributions of our work can be summarized as follows:

- We propose a bidirectional QAGAN framework embedded with QAM which performs *user-oriented* unpaired image enhancement by feed-forward learning.
- We propose a *quality attention module* (QAM) that enables the generator to directly learn the domain-relevant quality adaptively rather than only under the guidance of the discriminator.
- Extensive experiments demonstrate the superiority of the proposed method compared to the state-of-the-art in qualitative and quantitative evaluations.

2 RELATED WORK

2.1 Automatic Image enhancement

Conventional Image Enhancement: Extensive efforts have been made over the past few decades to automatically improve the aesthetic quality of images. Conventional image enhancement algorithms can be roughly divided into the following three categories: *histogram equalization* [3, 24], *unsharp masking* [29], and *Retinex-based methods* [7, 15]. Most of these methods either extract pre-defined sophisticated handcrafted features or adjust the image locally or globally based on prior knowledge. In addition, most of these adjustments are only used for individual types of improvement, such as contrast enhancement, edge sharpening, or color adjustment. For instance, Thomas *et al.* [24] adjusted the luminance histogram to a specific distribution based on the statistical information of the entire image to stretch the contrast. Ye *et al.* [29] adaptively performed pixel-wise enhancement under the guidance of blurriness to sharpen the image. Moreover, some works tried to manipulate image from visual attention or saliency [17–19, 27]. Such as, Nguyen *et al.* [19] proposed an image re-attentionizing framework to make the target region in the image attract the attention of humans. Mechrez *et al.* [18] designed an optimization framework to change visual saliency based on internal patch statistics.

Data-driven Image Enhancement: Due to the excellent modeling capabilities of deep convolutional neural network (CNN), the state-of-the-art image enhancement algorithms are almost all based on data-driven models [2, 4, 5, 8, 10, 11, 20, 25, 28]. Most of these methods, however, heavily depend on fully supervised learning that requires a large number of training image pairs. For example, Ignatov *et al.* [10] presented a model for improving the image quality of mobile devices to make it close to that of high-quality

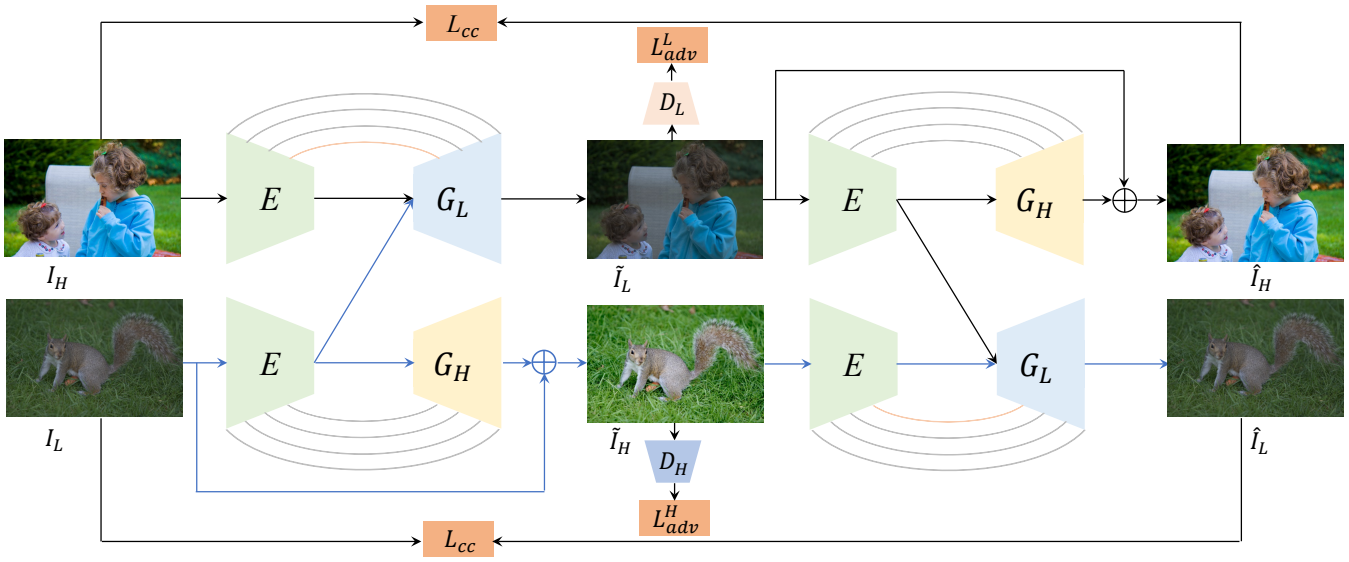

Figure 2: An overview of the proposed Quality Attention Generative Adversarial Network (QAGAN) for image enhancement. E is a pre-trained encoder (*i.e.*, VGG-19 [22] in our work) used to extract image features. G_L and G_H are low-quality and high-quality image generators, respectively. The discriminator D_L and D_H try to distinguish the generated low-quality image $\tilde{I}_L$ from real low-quality images, and distinguish the generated high-quality image $\tilde{I}_H$ from the real high-quality ones. The proposed framework uses the adversarial loss L_{adv} and cycle-consistency loss L_{cc} to train from the unpaired data in an end-to-end manner.

digital single-lens reflex photos. In order to enable the model to learn in a supervised manner, it takes a lot of time and efforts to create a paired training dataset DSLR Photo Enhancement Dataset (DPED). Gharbi *et al.* [5] realized the real-time image enhancement by jointly considering the local and global information to dynamically enhance the image. Wang *et al.* [25] proposed an image-to-illumination mapping model trained from a newly established under-exposed image dataset that consists of 3000 pairs of under-exposed images and corresponding retouched versions. However, the practicality of all these models is limited by the supervised models that rely heavily on the paired training data.

In recent years, the development of reinforcement learning (RL) and generative adversarial network (GAN) [6] has brought significant progress to data-driven image processing, especially the *image-to-image translation* tasks with *unpaired learning*. Hu *et al.* [8] proposed an RL-based framework that learns retouched images similar to those edited by professional photographers with sequentially apply differential filters. Kosugi *et al.* [14] developed an RL-based model to handle image enhancement by directly using image editing software (*e.g.*, Adobe Photoshop). Deng *et al.* [4] proposed a model based on adversarial learning, which is trained under the supervision of the aesthetic quality of binary tags. One of the most classical models dedicated to unpaired learning is the CycleGAN [31], which uses cycle consistency and adversarial losses to learn the transform mappings between two domains. The model is concise and explicit, but it makes unpaired image processing possible and achieve relative good results. Beyond CycleGAN, Chen *et al.* [2] made some improvements to stable the training of GAN to achieve better enhanced results. However, in these works, only the *discriminators* can access both domains while the *generators* can only learn the corresponding unidirectional mapping under the guidance of discriminators. On the contrary, the generator in our proposed model can also access both domains to directly learn the domain-relevant quality features which benefit the learning procedure significantly.

2.2 Attention Mechanism

Recently, attention mechanisms have been extensively studied for video classification [26], image generation [30], and image restoration [16]. The spatial attention mechanism for computer vision is proposed to make up for the defect that CNN is ineffective in capturing the long-range dependencies due to the locality of the convolution operations. Wang *et al.* [26] formulated the attention mechanism as the *non-local* operation to explore long-range dependencies. Zhang *et al.* [30] demonstrated that the quality of the generated image can be significantly improved by modeling the self-attention within an image. Liu *et al.* [16] proposed a non-local recurrent neural network (RNN) to use self-similarity prior in images for image denoising. All these attention models share the same idea, that is, correlation among non-local regions within an image/video is an effective prior for image processing. Our proposed *quality attention module* (QAM) is very close to this idea but we take advantage of the similarity between images from different domains.

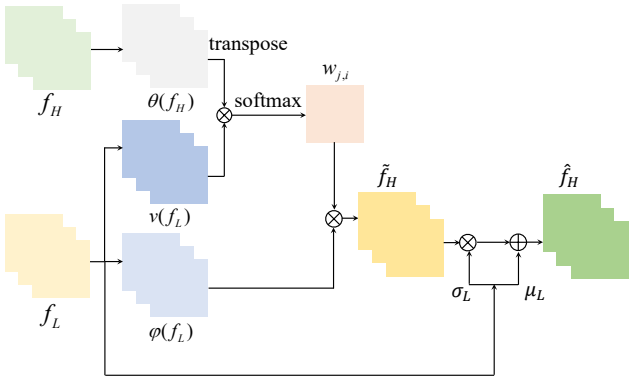

Figure 3: The structure of *quality attention module* (QAM). The QAM enables f_H to adaptively select content-relevant features from the spatial non-local inter-domain similarity and then incorporate the style attributes from the channel-wise statistics.

3 METHOD

An overview of the proposed unpaired learning model for image enhancement is shown in Fig. 2. Our proposed model is based on *bidirectional* GANs, which contains two generators: G_H and G_L are used to generate high-quality and low-quality images, respectively, and are paired with two discriminators D_H and D_L, respectively. The encoder E and generator G_H form a U-Net [21] structure dedicated to translating the low-quality image into a high-quality version. The generator G_L embeds the proposed *quality attention module* (QAM) to directly learn the domain-relevant quality features from the features of low-quality images. The discriminators D_L and D_H aim for distinguishing the generated data from the real low-quality images and high-quality images, respectively. The key novelty of the proposed QAGAN model lies in that we exploit the quality attention module of generator G_L to directly learn domain-relevant quality attention from the low-quality image domain. The proposed attention mechanism significantly benefits the generator G_L to achieve excellent performance in image mapping task from high-quality to low-quality, which therefore naturally leads to better performance in low-quality to high-quality image translation hereafter.

3.1 Quality Attention Module

Inspired by recent works on non-local self-similarity [26], we make the first attempt to introduce the non-local similarity of images from different domains into the generator in bidirectional GANs to achieve unpaired image enhancement. The proposed *quality attention module* (QAM) allows the generator G_L to effectively select semantic-relevant characteristics from the spatial-wise, and adaptively incorporate style-related attributes from the channel-wise, respectively. Therefore, the generator G_L can significantly be benefited from the QAM to translate the high-quality image to the low-quality version with learned domain-relevant quality features.

The detailed structure of the proposed QAM is shown in Fig. 3. In short, for given encoder features $f_L \in \mathbb{R}^{N\times C\times H\times W}$ of low-quality image I_L and encoder features $f_H \in \mathbb{R}^{N\times C\times H\times W}$ of high-quality image I_H. The proposed QAM first uses non-local inter-domain spatial similarity to generate spatially weighted features $\tilde{f}_H \in \mathbb{R}^{N\times C\times H\times W}$ and then derives the channel-wise recalibration features $\hat{f}_H \in \mathbb{R}^{N\times C\times H\times W}$ according to the style of f_L, where N is the value of batch size, C is the number of channels, H and W are high and width of the features, respectively. More specifically, we first generate two new features $\nu(f_L)$ and $\varphi(f_L)$ from f_L to reduce computational complexity, where $\{\nu(f_L), \varphi(f_L)\} \in \mathbb{R}^{N\times \frac{C}{8}\times H\times W}$. In a similar way, the $\theta(f_H) \in \mathbb{R}^{N\times \frac{C}{8}\times H\times W}$ is generated from f_H. After that, we reshape $\{\theta(f_H), \nu(f_L)\} \in \mathbb{R}^{N\times \frac{C}{8}\times H\times W}$ to $\{\theta(f_H), \nu(f_L)\} \in \mathbb{R}^{N\times \frac{C}{8}\times S}$, and then multiply $\nu(f_L)$ and the transpose of $\theta(f_H)$, where $S = H * W$. The inter-domain non-local spatial similarity map $w_{j,i}$ can be derived by:

$$w_{j,i} = \frac{e^{\theta(f_H)_i^T \cdot \nu(f_L)_j}}{\sum_{i=1}^{S} e^{\theta(f_H)_i^T \cdot \nu(f_L)_j}}, \tag{1}$$

where $w_{j,i}$ represents the dependence of the j-th position of f_L on the i-th position of f_H. The output of the inter-domain non-local spatial similarity can be formulated as the weighted sum of the features over all positions as follows:

$$e_i = \sum_{i=1}^{N} (w_{j,i} \cdot \varphi(f_L)). \tag{2}$$

The refined features $\tilde{f}_H$ can be obtained by simply reshaping the e_i. This inter-domain non-local spatial similarity allows the generator G_L to effectively learn the spatial content-relevant characteristics from I_L directly.

Inspired by [9], we adjust the channel-wise statistics (*i.e.*, mean and variance) of I_H to match that of I_L to inject the style of the low-quality image I_L to the generated low-quality version $\tilde{I}_L$. We adopt this strategy because its efficiency and effectiveness have been witnessed in many style transfer works. Specifically, the style features mean μ_L and variance σ_L can be computed as:

$$\mu_L = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} f_L, \tag{3}$$

$$\sigma_L = \left(\frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} (f_L - \mu_L)^2 \right)^{\frac{1}{2}}, \tag{4}$$

where, the μ_L and σ_L are regarded as the style information of the low-quality image I_L. Therefore, we can translate it to the generated low-quality image $\tilde{I}_L$ by simply scaling the normalized $\tilde{f}_H$ with σ_L and adding the bias μ_L, which can be expressed as:

$$\hat{f}_H = \sigma_L \left(\frac{\tilde{f}_H - \mu_H}{\sigma_H} \right) + \mu_L, \tag{5}$$

where μ_H and σ_H are the mean and variance of $\tilde{f}_H$, respectively, and the calculation is similar to that of f_L.

In summary, the proposed QAM allows the generator G_L to learn domain-relevant quality features from I_L in terms of

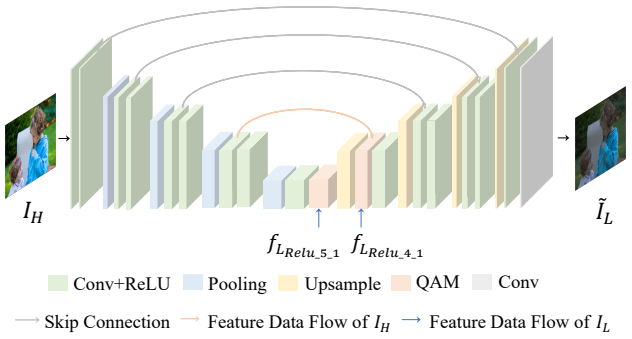

Figure 4: The structure of the encoder E and decoder G_L. We employ the pre-trained VGG19 [22] as the features extractor and use the features at five layers (*i.e.*, Relu_X_1, X=1, 2, 3, 4, 5). The $f_{L_{Relu\_5\_1}}$ and $f_{L_{Relu\_4\_1}}$ are the features of I_L at corresponding layers, respectively.

the spatial-wise based on non-local inter-domain similarity and channel-wise according to mean and variance.

3.2 Network Architecture

As shown in Fig. 2, our framework employs the pre-trained VGG [22] as the encoder E to transfer the I_L and I_H into the feature space. The generator G_L is a symmetric decoder incorporated with the proposed QAM to convert the low-quality images into high-quality versions. The generator G_H and E are combined to form a U-Net to map the low-quality images into the high-quality counterparts. The associated discriminators D_L and D_H are responsible for distinguishing the generated low-quality image $\tilde{I}_L$ from the real low-quality images and judging the generated high-quality image $\tilde{I}_H$ from the real high-quality ones.

As shown in Fig. 4, based on the pre-trained VGG [22], a symmetric decoder G_L and the proposed QAM, an encoder-decoder network is constructed to perform image quality degradation. We employ the pre-trained VGG [22] as the feature extractor and use features in five layers (*i.e.*, Relu_X_1, X=1, 2, 3, 4, 5). The deeper the network is, the larger the receptive field size is. Therefore, we use two layers (*i.e.*, Relu_5_1 and Relu_4_1) to capture the characteristics of low-quality features at different levels. To be specific, fed $\{f_{H_{Relu\_5\_1}}, f_{L_{Relu\_5\_1}}\}$ and $\{f_{H_{Relu\_4\_1}}, f_{L_{Relu\_4\_1}}\}$ into the first and second QAM, respectively, to generate the quality-aware features at the highest and second-highest level, respectively. Then fuse the two levels of quality-aware features after upsampling the spatial size of the former, which is then fed into the following modules to generate the low-quality images. The generator G_H is a symmetrical decoder of VGG-19 to form U-Net [21], which is used to translate the low-quality images into the high-quality versions.

Generally speaking, a larger receptive field is beneficial for the discriminator to capture the global features to judge the real image and the generated image. Although the low-level features of the discriminator only have a smaller receptive field, it can guide the generator to produce better details. We propose a multi-scale discriminator that uses not only the final layer of the discriminator but also multiple low-level features (*i.e.*, an intermediate layer of the discriminator) to guide the generator to generate the image with both finer details and global consistency.

3.3 Loss Function

1) Adversarial Loss: The adversarial loss used to train the mapping functions is a variation of the recently proposed relativistic discriminator structure [12]. On the one hand, this loss evaluates the probability that real data is more real than fake data, on the other hand, it also enforces the generated data is more realistic than the real data. The Relativistic average HingeGAN (RaHingeGAN) of mapping from low-quality to high-quality can be expressed as follows:

$$\begin{aligned} L_{\text{adv}}^{H} = &\mathbb{E}_{\tilde{I}_H \sim P_{\tilde{H}}}[\max(0, 1 + (D(\tilde{I}_H) - \mathbb{E}_{I_H \sim P_H} D_H(I_H)))] \\ &+\mathbb{E}_{I_H \sim P_H}[\max(0, 1 - (D(I_H) - \mathbb{E}_{\tilde{I}_H \sim P_{\tilde{H}}} D_H(\tilde{I}_H)))], \end{aligned} \tag{6}$$

where I_H and $\tilde{I}_H$ denote the real data from the high-quality image dataset P_H, and the generated data from the generator G_H, respectively.

Analogously, the RaHingeGAN of the mapping function from high-quality to low-quality can be formulated as:

$$\begin{aligned} L_{\text{adv}}^{L} = &\mathbb{E}_{\tilde{I}_L \sim P_{\tilde{L}}}[\max(0, 1 + (D(\tilde{I}_L) - \mathbb{E}_{I_L \sim P_L} D_L(I_L)))] \\ &+\mathbb{E}_{I_L \sim P_L}[\max(0, 1 - (D(I_L) - \mathbb{E}_{\tilde{I}_L \sim P_{\tilde{L}}} D_L(\tilde{I}_L)))], \end{aligned} \tag{7}$$

where I_L and $\tilde{I}_L$ denote the real data from the low-quality image dataset P_L, and the generated data from the generator G_L, respectively.

2) Cycle-Consistency Loss: The cycle-consistency loss is based on the assumption that an image should be mapped back to the original image after two opposite mappings. The cycle-consistency loss defined as the ℓ_2 norm between the feature maps of the input images I_H (or I_L) and those of the recycled images $\hat{I}_H$ (or $\hat{I}_L$) extracted by the pre-trained VGG network [22], as follows:

$$\begin{aligned} L_{\text{cc}} = &\frac{1}{J}\sum_{j=1}^{J}\{\mathbb{E}_{I_H \sim P_H}[\|\phi_j(I_H) - \mathbb{E}_{\hat{I}_H \sim P_{\hat{H}}}\phi_j(\hat{I}_H)\|_2] \\ &+ \mathbb{E}_{I_L \sim P_L}[\|\phi_j(I_L) - \mathbb{E}_{\hat{I}_L \sim P_{\hat{L}}}\phi_j(\hat{I}_L)\|_2]\}, \end{aligned} \tag{8}$$

where $\phi_j(\cdot)$ means to extract the feature maps of the j^{th} layer of the VGG network and J is the total number of layers used. Specifically, we use VGG-19 layers *Relu*_1_1, *Relu*_2_1, *Relu*_3_1, *Relu*_4_1, and *Relu*_5_1, to compute cycle-consistency loss.

3) Total Loss: By jointly considering the *adversarial loss* and *cycle-consistency loss*, we define the final loss as the weighted sum of these losses, as follows:

$$L_{\text{total}} = L_{\text{adv}}^{H} + L_{\text{adv}}^{L} + \lambda_{\text{cc}} L_{\text{cc}}, \tag{9}$$

where λ_{cc} is a weighting parameter to balance the relative importance of L_{cc}.

(a) Input (b) CycleGAN [31] (c) Exposure [8] (d) EnlightenGAN [11]

(e) DPE [2] (f) Ours (FiveK) (g) Ours (Flickr) (h) Expert-retouched

Figure 5: The visual quality comparison results between the proposed method against state-of-the-art methods.

4 EXPERIMENTS

In this section, extensive quantitative and qualitative experimental results are presented and discussed. Specifically, we first introduced two datasets for evaluating our proposed method and the methods being compared, and described the implementation details. Then, we compared our proposed model with state-of-the-art methods in terms of subjective and objective metrics. Last, we carried out user study to consolidate the conclusions of our subjective and objective evaluations.

4.1 Datasets and Evaluation Metrics

MIT-Adobe FiveK Dataset [1]: It consists of 5,000 raw photos taken by different photographers using various single-lens reflex cameras and 25,000 retouched photos. These retouched photos were generated from the raw photos by 5 experienced retouchers (*i.e.*, retoucher A, B, C, D, and E) and the results obtained by each retoucher have different visual experiences. According to [2] and [8], the photos generated by retoucher C were selected as the target photos because his results are the most popular with users. In order to meet the requirement of the proposed model for unpaired learning, the dataset was randomly divided into three non-overlapping subsets: 1) Subset 1 is a low-quality image subset consisting of 2,250 raw photos; 2) Subset 2 is a target high-quality image subset composing of 2,250 retouched photos of other raw photos; 3) Subset 3 contains the remaining 500 raw photos used for validation (100 images) and testing (400 images).

Flickr Dataset: In order to verify the performance of the proposed model, a high-quality image dataset crawled from Flickr to perform image enhancement under the supervision without paired data. The dataset contains 2,000 images with relatively high quality and served as the target images in unpaired training.

Evaluation Metrics: The widely used Peak Signal to Noise Ratio (PSNR) and Structural SIMilarity (SSIM) were used to evaluate the performance of each model in terms of *signal fidelity*. In addition, the NIMA [23] was adopted to quantify and compare the *aesthetic quality* of the results of each model.

4.2 Implementation Details

We implemented our model in Pytorch and used the Adam solver [13] to optimize the network with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The parameters of encoder (*i.e.*, VGG-19) were fixed, two generators and two discriminators were trained from scratch. The experiments were conducted on an NVIDIA GeForce RTX 2080 Ti GPU. The network was trained for 150 epochs and the mini-batch size was set as 12. The learning rates of generators and discriminators are both initialized to 1e-4 and then linearly decays to zero from 75 to 150 epochs. During training, we randomly cropped 256×256 patches and then resize them to 128×128. For testing, the resolution of all test images is 512×512. The weight parameter $\mathcal{L}_{cc}$ was

(a) Input (b) CycleGAN [31] (c) Exposure [8] (d) EnlightenGAN [11]

(e) DPE [2] (f) Ours (FiveK) (g) Ours (Flickr) (h) Expert-retouched

Figure 6: The visual quality comparison results between the proposed method against state-of-the-art methods.

Table 1: Quantitative comparison results between our proposed method and state-of-the-art methods on MIT-Adobe FiveK Dataset [1].

Method	PSNR	SSIM	NIMA
Input	17.21	0.8012	4.41
CycleGAN [31]	20.53	0.7786	4.26
Exposure [8]	19.59	0.8361	4.55
UIE [14]	21.80	0.8708	4.53
EnlightenGAN [11]	16.83	0.7513	4.18
DPE [2]	22.28	0.8542	4.39
Ours	**22.59**	**0.8782**	**4.58**

empirically set to 2 and 10 for MIT-Adobe FIveK Dataset and Flickr Dataset, respectively.

4.3 Quantitative Comparison

We evaluated the performance of the proposed model against four state-of-the-art image enhancement methods learned without paired data: Deep Photo Enhancer (DPE) [2], EnlightenGAN [11], Exposure [8], and UIE [14]. We also compared it with the classical unpaired image-to-image translation model CycleGAN [31]. The DPE is based on CycleGAN model that trained with several improvements to stabilize the adversarial training process, while EnlightenGAN is a unidirectional GAN model that performs unpaired image enhancement. The Exposure and UIE are both unpaired image enhancement method based on differential filters and RL. In order to make a fair comparison, all codes of the comparison models are download from the website provided by the corresponding authors. All models under comparison are trained on the MIT Adobe FiveK dataset and our proposed model also trained on the Flickr dataset.

Table 1 tabulates the PSNR, SSIM, and NIMA comparison results of the proposed method against the state-of-the-art models on MIT Adobe FiveK dataset. The first-ranked performance for each evaluation metric is highlighted in bold black. For the Flickr dataset we collected, we only performed visual quality comparison and user study due to the lack of ground truth. It can be seen from Table 1 that on the MIT Adobe FiveK dataset, our proposed method has achieved the best PSNR, SSIM, and NIMA among all these comparison methods. Considering that DPE which ranks second in all metrics, although is the improved version of the CycleGAN from three aspects, it is still inferior to our proposed QAGAN model. We attribute this to the proposed QAM such that the generator can effectively learn domain-relevant features directly from both two domains. In addition, it can be observed that for some images the performances of CycleGAN and EnlightenGAN are even worse than the input images. This is mainly due to the CycleGAN introduces blocking

Table 2: The results of pairwise comparison in user study. EG denotes EnlightenGAN.

	Input	CycleGAN [31]	Exposure [2]	EG [11]	DPE [8]	Ours (FiveK)	Ours (Flickr)	Total
Input	-	328	125	145	46	26	14	684
CycleGAN	632	-	157	224	91	73	46	1223
Exposure	835	803	-	588	254	227	176	2883
EG	815	736	372	-	179	152	121	2375
DPE	914	869	706	781	-	362	267	3899
Ours (FiveK)	934	887	733	808	598	-	389	4349
Ours (Flickr)	946	914	784	839	693	571	-	4747

artifacts, and EnlightenGAN significantly over-adjusts the contrast of images.

4.4 Qualitative Comparison

We also conducted extensive visual quality comparisons between our proposed method and the state-of-the-art methods. As shown in Fig. 5 and Fig. 6, one can observe that our proposed model trained on the collect Flickr dataset achieves the best visual experience as it produces better contrast and brightness. In addition, our proposed model trained on MIT-Adobe FiveK dataset can also produce satisfactory results. However, the methods under comparison have more or less failed in terms of contrast, color, or details. CycleGAN generates severe blocking artifacts in smooth regions and produces checkerboard artifacts at the boundary. Exposure tends to generate over-saturated results and make the colors dull. EnlightenGAN is less effective in producing good contrast and usually causes the input images to be over-exposed. DPE can produces the most comparable results to ours, but it is inferior to ours in contrast. Generally speaking, the results generated by our proposed method have comfortable contrast, vivid colors, and clear details, which are more satisfying and superior to the state-of-the-art methods in comparison.

4.5 User Study

It is a challenge to evaluate image quality enhancement from an objective perspective. To this end, we studied how users preference the results of each model. Specifically, we used pairwise comparison on six methods (including two versions of our method) to conduct a user study with 24 participants and 40 images. Each time, we randomly present the results of two different models to the participants and ask them to choose his/her favorite result from the displayed image pair. The results of the pairwise comparison are documented in Table 2, where each value represents the number of times the method in that row outperforms the method in that column. From the results, we can observe that the results generated by the proposed method is more favourite with users, compared with other models (*i.e.*, CycleGAN, Exposure, EnlightenGAN, and DPE) our proposed model has been selected the most times.

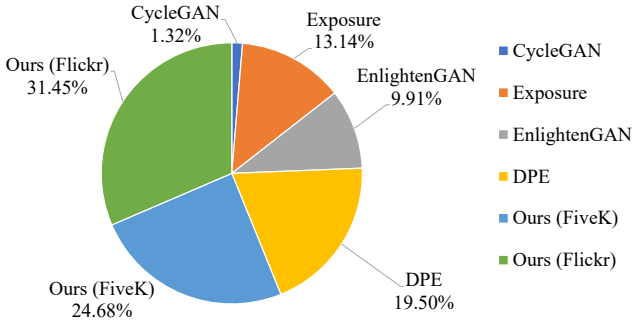

Figure 7: User preference results of different aesthetic quality enhancement algorithms.

This confirms that our proposed method is not only superior to the state-of-the-art methods in the objective aspects of PSNR, SSIM, and NIMA, but is also favored by users in subjective aspect of visual quality.

We also evaluated the overall quality by again randomly selecting 100 test images and 100 enhancement results corresponding to each model. Each time, six results were randomly presented to the participants and asking them to select their favorite one. At last, a total of 2400 votes were collected and visualized in Fig. 7. It can be seen that the results generated by our proposed method are again preferred much more often than those by other methods under comparison.

5 CONCLUSIONS

In this paper, we present a bidirectional *Generative Adversarial Network* (GAN) with *quality attention module* (QAM) to train which is trained on unpaired data to solve the task of image enhancement. We embed the proposed QAM in the generator so that the generator directly learns the domain-relevant quality features from the features of low-quality images. Specifically, the proposed QAM allows the generator to effectively select semantic-relevant characteristics from the spatial-wise and adaptively incorporate style-related attributes from the channel-wise, respectively. The quantitative and qualitative experimental results show that our proposed method is superior to the state-of-the-art methods.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous referees for their insightful comments and suggestions. This work was supported in part by the Hong Kong RGC General Research Funds under Grant 9042322 (CityU 11200116), Grant 9042489 (CityU 11206317), and Grant 9042816 (CityU 11209819), and in part by the Natural Science Foundation of China under Grant 61672443.

REFERENCES

[1] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. 2011. Learning photographic global tonal adjustment with a database of input/output image pairs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 97–104.

[2] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. 2018. Deep photo enhancer: unpaired learning for image enhancement from photographs with GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6306–6314.

[3] Dinu Coltuc, Philippe Bolon, and J-M Chassery. 2006. Exact histogram specification. *IEEE Transactions on Image Processing* 15, 5 (2006), 1143–1152.

[4] Yubin Deng, Chen Change Loy, and Xiaoou Tang. 2018. Aesthetic-driven image enhancement by adversarial learning. In *Proceedings of the ACM International Conference on Multimedia*. 870–878.

[5] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. 2017. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics* 36, 4 (2017), 118.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the Advances in Neural Information Processing Systems*. 2672–2680.

[7] Xiaojie Guo, Yu Li, and Haibin Ling. 2017. LIME: low-light image enhancement via illumination map estimation. *IEEE Transactions Image Processing* 26, 2 (2017), 982–993.

[8] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. 2018. Exposure: A white-box photo post-processing framework. *ACM Transactions on Graphics* 37, 2 (2018), 26.

[9] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*. 1501–1510.

[10] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. 2017. DSLR-quality photos on mobile devices with deep convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 3277–3285.

[11] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. 2019. EnlightenGAN: deep light enhancement without paired supervision. *arXiv preprint arXiv:1906.06972* (2019).

[12] Alexia Jolicoeur-Martineau. 2018. The relativistic discriminator: a key element missing from standard GAN. *arXiv preprint arXiv:1807.00734* (2018).

[13] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[14] Satoshi Kosugi and Toshihiko Yamasaki. 2019. Unpaired image enhancement featuring reinforcement-learning-controlled image editing Software. *arXiv preprint arXiv:1912.07833* (2019).

[15] Mading Li, Jiaying Liu, Wenhan Yang, Xiaoyan Sun, and Zongming Guo. 2018. Structure-revealing low-light image enhancement via robust Retinex model. *IEEE Transactions on Image Processing* 27, 6 (2018), 2828–2841.

[16] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. 2018. Non-local recurrent network for image restoration. In *Proceedings of the Advances in Neural Information Processing Systems*. 1673–1682.

[17] Victor A Mateescu and Ivan V Bajic. 2015. Visual attention retargeting. *IEEE MultiMedia* 23, 1 (2015), 82–91.

[18] Roey Mechrez, Eli Shechtman, and Lihi Zelnik-Manor. 2019. Saliency driven image manipulation. *Machine Vision and Applications* 30, 2 (2019), 189–202.

[19] Tam V Nguyen, Bingbing Ni, Hairong Liu, Wei Xia, Jiebo Luo, Mohan Kankanhalli, and Shuicheng Yan. 2013. Image re-attentionizing. *IEEE Transactions on Multimedia* 15, 8 (2013), 1910–1919.

[20] Wenqi Ren, Sifei Liu, Lin Ma, Qianqian Xu, Xiangyu Xu, Xiaochun Cao, Junping Du, and Ming-Hsuan Yang. 2019. Low-light image enhancement via a deep hybrid network. *IEEE Transactions on Image Processing* 28, 9 (2019), 4364–4375.

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[22] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[23] Hossein Talebi and Peyman Milanfar. 2018. NIMA: Neural image assessment. *IEEE Transactions on Image Processing* 27, 8 (2018), 3998–4011.

[24] Gabriel Thomas, Daniel Flores-Tapia, and Stephen Pistorius. 2011. Histogram specification: a fast and flexible method to process digital images. *IEEE Transactions on Instrumentation and Measurement* 60, 5 (2011), 1565–1578.

[25] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. 2019. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6849–6857.

[26] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7794–7803.

[27] Lai-Kuan Wong and Kok-Lim Low. 2011. Saliency retargeting: an approach to enhance image aesthetics. In *IEEE Workshop on Applications of Computer Vision*. IEEE, 73–80.

[28] Zhicheng Yan, Hao Zhang, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. 2016. Automatic photo adjustment using deep neural networks. *ACM Transactions on Graphics* 35, 2 (2016), 11.

[29] Wei Ye and Kai-Kuang Ma. 2018. Blurriness-guided unsharp masking. *IEEE Transactions on Image Processing* 27, 9 (2018), 4465–4477.

[30] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2018. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318* (2018).

[31] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 2223–2232.