





Quality-Aware Unpaired Image-to-Image Translation

Lei Chen , Le Wu , Zhenzhen Hu , and Meng Wang , *Senior Member, IEEE*

Abstract—Generative adversarial networks (GANs) have been widely used for the image-to-image translation task. While these models rely heavily on the labeled image pairs, recently some GAN variants have been proposed to tackle the unpaired image translation task. These models exploited supervision at the domain level with a reconstruction process for unpaired image translation. On the other hand, parallel works have shown that leveraging perceptual loss functions based on high-level deep features could enhance the generated image quality. Nevertheless, as these GAN-based models either depended on the pretrained deep network structure or relied on the labeled image pairs, they could not be directly applied to the unpaired image translation task. Moreover, despite the improvement of the introduced perceptual losses from deep neural networks, few researchers have explored the possibility of improving the generated image quality from classical image quality measures. To tackle the above two challenges, in this paper, we propose a unified quality-aware GAN-based framework for unpaired image-to-image translation, where a quality-aware loss is explicitly incorporated by comparing each source image and the reconstructed image at the domain level. Specifically, we design two detailed implementations of the quality loss. The first method is based on a classical image quality assessment measure by defining a classical quality-aware loss to ensure similar quality score between an original image and the reconstructed image at the domain level. The second method proposes an adaptive deep network based loss that compares the high level content structure between each original image and its reconstructed image from the generator. Finally, extensive experimental results on many real-world datasets clearly show the quality improvement of our proposed framework, and the superiority of leveraging classical image quality measures for unpaired image translation compared to the deep network based model.

Index Terms—Image generation, image processing, image quality, neural networks.

I. INTRODUCTION

MANY real-world computer vision tasks, such as image segmentation, stylization and abstraction, could be treated as an image-to-image translation problem. This problem involves transforming an image from a source domain (e.g., photo) to imitate the image in the target domain (e.g., sketch) [1]–[4]. Since the seminal work of GANs by Goodfellow

et al. at 2014 [5], GAN and their variants provide state-of-the-art solutions to the image-to-image translation task. Given an image pair with a source image and its corresponding image in the target domain, GANs learn an adversarial loss function that tries to maximize the discriminator to correctly classify if the generated image is real or fake in the target scene, and simultaneously trains a generative model that tries to fake the discriminator.

Generally, these classical GAN-based models rely on labeled image pairs from the source domain and the target domain for image translation. Nevertheless, in the real-world, it is relatively easy to collect different images in the source domain and target domain separately, while acquiring such side-by-side matching pairs is time and labor consuming [6], [7]. E.g., we could collect a set of images in the summer scene and a set of images in the winter scene. However, it is nontrivial to get a side-by-side pair of an image taken in the summer next to the same matching image that is taken in the winter. Thus, some unpaired GAN variants have been proposed to translate an image from one domain to the remaining domain without any paired images, including DualGAN [1], CycleGAN [3], and DiscoGAN [4]. Roughly, all these models shared a similar idea. While a normal GAN has only one generator and one discriminator that is trained on the labeled image pairs, these unpaired GANs exploit supervision at the domain set level. Specifically, for image translation from a domain U to a domain V , there are two GANs with one primal GAN with generator G_U learns to translate an image from domain U to that in domain V , and a corresponding GAN with generator G_V that learns to invert the translated image in V to the original domain U with an additional reconstruction loss.

In these GAN variants for (unpaired) image translation task, the performance relies heavily on the designed optimization loss function. All these GAN-based objective function includes an adversarial loss that alternates between identifying and faking. Besides, to ensure the similarity between an original image and its generated version, some (unpaired) GAN variants have introduced the pixel level loss in the modeling process [1], [3], [4], [8], e.g., the error between them with L1 or L2 distance loss [8]. Nevertheless, even though the pixel level matching consistency is high, people are not satisfied with the generated images as human visual system usually focus on the higher level abstractions of images for perceptual quality evaluation [9], [10]. Thus, some works focused on leveraging perceptual loss functions based on high level features of deep networks to enhance the generated image quality [11], [12]. E.g., Johnson *et al.* introduced perceptual loss functions that depended on high level image features extracted from a pretrained deep neural network [13]. The newly added perceptual loss functions could generate quality-enhanced results for some specific image translation tasks. However, these pretrained deep networks are not optimized

Manuscript received September 27, 2018; revised January 15, 2019; accepted March 8, 2019. Date of publication March 25, 2019; date of current version September 23, 2019. This work was supported in part by the National Natural Science Foundation for Distinguished Young Scholars of China under Grant 61725203, in part by the National Key Research and Development Program of China under Grant 2017YFB0803301, and in part by the National Natural Science Foundation of China under Grants 61602147, 61732008, 61802104. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zhu Li.

The authors are with the School of Computer and Information, Hefei University of Technology, Hefei 230009, China (e-mail: chenlei.hfut@gmail.com; lewu.ustc@gmail.com; huzhen.ice@gmail.com; eric.mengwang@gmail.com).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TMM.2019.2907052

1520-9210 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

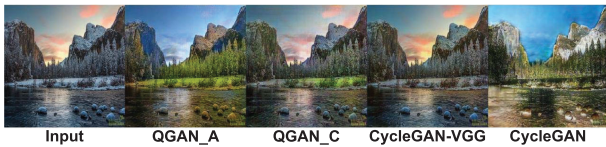


Fig. 1. An example of the winter-to-summer translation task. Given an input from the winter style, it is desired that the generated output images are visually qualified from the human perspective. Here, CycleGAN [3] is a state-of-the-art unpaired image-to-image translation model. CycleGAN-VGG combines CycleGAN and perceptual loss functions based on the pretrained VGG [9]. QGAN_A and QGAN_C are our proposed two quality-aware models.

for the image-to-image translation task, not to mention the situation when the target image domain never appears in the pretrained network. For example, as shown in Fig. 1, the fifth column is the results of an unpaired image translation model of CycleGAN [3]. We also show the results that combined CycleGAN [3] and the pretrained VGG for perceptual loss control in the fourth column. It is clearly observed that some patches in the fourth and fifth column are not summer and contain blurry structure. To tackle the suboptimal performance caused by the pretrained network, Wang *et al.* designed a perceptual adversarial loss that undergoes an adversarial training process for image translation [12]. In the proposed model, the adversarial discriminator evaluates the perceptual loss between the generated image and the ground truth image with the discriminator network, achieving impressive results for image translation.

In this paper, we study the problem of generating high quality images for unpaired image-to-image translation. A natural idea is to combine the recent progress of unpaired image translation and the quality enhancement techniques for paired image translation. Some researchers proposed to introduce a content-based loss in the optimization function, where the content-based loss is constructed from pretrained deep networks between a source image and the corresponding generated image [9], [13]–[15]. These models could generate images with better quality than previous image translation models. However, as the pretrained network parameters are not designed for image translation, simply transferring these parameters would lead to suboptimal quality-aware unpaired image translation task. Therefore, how to adaptively design the content-based quality loss to the unpaired image translation task to further enhance image quality still remains open. Moreover, image quality assessment is a classical topic that has been well studied in the community [16]–[18]. As most image translation models relied on the deep features to model the perceptual losses for enhancing image quality, the question of whether the classical image quality measures could help to improve the quality of generated images is still under explored. To tackle the above two challenges, in this paper, we propose a unified quality-aware GAN-based framework for unpaired image-to-image translation. We extend the adversarial loss functions in unpaired image translation models and put emphasis on how to design quality-aware loss functions that could be applied to the image translation task without image pair information. Instead of comparing the quality between u and the output of generator G_U as $G_U(u)$, we model the

quality loss terms between u and the reconstructed images of u (i.e., $\hat{u} = G_V(G_U(u))$). Thus, the designed quality-aware loss functions are freed from the image pair constraints. Specifically, we introduce two detailed models to implement the quality loss. We first borrow the ideas of these classical quality assessment methods and propose a method that defines the quality loss between each original image and the reconstructed image with a classical image quality assessment measure. We then propose an adaptive content loss that combines the visual content structure loss from GANs for quality-aware unpaired image-to-image translation. The content loss adaptively captures the high level perceptual quality between the original images and the reconstructed images in the generators.

In summary, our paper makes the following contributions:

- 1) We point out that current image-to-image translation tasks relied on the labeled image pairs for designing image quality based losses. To this end, we propose a unified quality-aware unpaired GAN-based image translation framework, which relies on the quality losses between u and $\hat{u} = G_V(G_U(u))$.
- 2) Under the proposed quality-aware framework, we design two detailed model implementations of the quality loss. The first proposed model introduces a classical quality assessment loss, and the second model combines a high level adaptive visual content structure loss in addition to the adversarial loss in GAN for modeling human perceptual quality evaluation.
- 3) We perform extensive experimental results on four real-world datasets. Extensive experimental results from both the quality assessment measures and the human opinion scores show that our proposed models improve the quality of the generated images. Also, we observe the superiority of the classical quality assessment loss compared to the high level content-based loss.

II. RELATED WORK

The idea of image-to-image translation goes back at least to Gatys *et al.*'s neural algorithm of artistic style [19], which designed a neural algorithm to separate content and style and then recombined the two parts. Since the seminal work of GAN by Goodfellow *et al.* at 2014 [5], the recent image-to-image translation task has been tackled under the GAN framework. GANs learn an adversarial optimization function that maximizes the discriminator to correctly classify if the output image is real or fake, and simultaneously a generative model that minimizes the loss. Different GAN variants for the image-to-image task varied in the specific implementations of the discriminator and the generator. E.g., Isola *et al.* proposed a conditional adversarial network with generic loss function [8] and Mao *et al.* presented a GAN variant that adopted the least square based loss function for the discriminator [20]. While these models relied on image pairs for the translation task, recently some unpaired image-to-image models have been proposed [1], [3], [4]. These models shared a similar idea by learning a primal GAN from a source domain to a target domain, and a dual process that transformed the generated images from the target domain to the source domain.

Then, a pixel level reconstruction loss, such as mean squared error (MSE) [21], [22], is introduced between the original images and the reconstructed images. These models advanced previous works by loosening the inputs to unlabeled training data.

In fact, the performance of the image translation task relies heavily on the designed optimization function, which can effectively drive the network's learning, leading to a large impact on the performance of this task. In the real world, the human visual system is quite subjective and human usually focus on the higher level abstractions for perceptual quality evaluation [9], [18], [23]. Nevertheless, the pixel-wise loss functions suffered from the limitation of poorly reflecting the human visual experience, and thus typically induce blurry parts. Luckily, Convolutional Neural Networks (CNN) have shown promising performance to automatically extract high level content structure information of images [24]–[27]. Many works empirically validated that the higher layers of the CNN network capture the perceptual abstractions of images. Thus, many image processing related tasks, such as image resolution [9], [28], style transfer [13], [29] and unsupervised depth and motion estimation [30]–[33], are proposed to leverage the feature maps in CNNs as perceptual quality measures, and incorporated the perceptual quality into the optimization function to generate quality-aware images [9], [13]–[15]. Usually, the perceptual loss function [9] modeled the image features from a pretrained VGG network [34]. As these models do not need any labeled information for quality evaluation, they could be applied to the unpaired image translation task. Without confusion, in the following of this paper, we use “CycleGAN-VGG” to denote the unpaired image translation task that combined the loss of CycleGAN and the perceptual loss function from pretrained VGG network. However, since these pretrained networks are tailored to a specific dataset, it is suboptimal to directly transfer them to the image translation task. For example, as shown in Fig 1, CycleGAN-VGG fails for the unpaired winter-to-summer translation as the pretrained network could not effectively capture the perceptual network configurations for winter and summer domain. To tackle the limitations of the pretrained network, recently researchers proposed a perceptual adversarial loss that undergoes an adversarial training process between the generator and the discriminator [35]. The proposed model introduced an adaptive perceptual loss that automatically discovered the discrepancy between the generated image and the ground truth with higher layer based abstractions from deep networks. To tackle the case in unpaired image translation, the authors use a small paragraph to illustrate how to extend this method to unpaired image translation. To avoid using the ground truth of labeled image pairs, instead of comparing the generated image and its ground truth, they proposed to calculate the discrepancy of mean features of these two domains. This proposed method showed better performance for paired image translation. However, as it is not dedicated for unpaired image translation, simply comparing the mean discrepancy between two domains would discard the characteristics of the currently generated image, leading to unsatisfactory performance. In summary, our work borrows the ideas of these previous works, and we advance in the following two aspects. Firstly, we would like to explore how to effectively improve the perceptual quality

of the generated images in unpaired image translation task. Secondly, despite the breakthroughs of leveraging perceptual losses derived from deep neural networks, to the best of our knowledge, we are one of the first few attempts that explore the possibility of combining the classical quality assessment measures for unpaired image translation.

Our work is also closely related to the Image Quality Assessment (IQA) measures. This direction aims to use computational models to measure the image quality that is consistent with human subjective evaluations. Generally speaking, current IQA techniques mainly follow two directions: the blind reference (BR) and the full reference (FR). BR evaluates image quality without any reference, and this kind of models usually designed some features for quality modeling [36]–[39]. However, as the overall quality of each domain varies [16], [38], [40], [41], these models either needed human labeled image quality values or were only applicable to a specific domain for quality evaluation. Different from BR, FR usually evaluates the visual quality of an image by comparing a generated image with the original image in the image-to-image translation task [16], [18]. Instead by comparing the pixel level similarity such as peak signal-to-noise ratio (PSNR) and the mean-squared error (MSE) that directly operate on the pixel level of images, the FR methods show great success by designing the specific subjective features to simulate human visual evaluation. E.g., SSIM [16] proposed a complementary method for structural similarity. Based on the physiological and psychophysical evidences, FSIM [18] emphasized the human visual system to understand the image based on the Fourier low frequency features of images. As the human visual system is adapted to structural information of images, GMSD [17] is proposed to use the gradient similarity based method to measure image quality efficiently. With the development of deep neural networks, recently some methods of IQA have made preliminary attempts for automatically capturing image quality related features from deep neural networks [42]–[45]. But all of these methods directly or indirectly required examples and corresponding human opinion scores, which usually are expensive. In this work, we would like to borrow a classical FR method for designing a quality-aware unpaired image translation model.

III. PRELIMINARIES

In this section, we introduce the key ideas of several recent state-of-the-art GAN-based unpaired image-to-image translation models, including DualGAN [1], CycleGAN [3], and DiscoGAN [4]. Since all these models share a similar idea, for ease of explanation, we take the CycleGAN [3] as an example to show the key ideas of the GAN-based unpaired image-to-image translation models.

CycleGAN learns to translate an image u from a source domain U ($u \in U$) to a target domain V with a generator G_U in the absence of paired examples, such that $G_U(u)$ is indistinguishable from the distribution of images in V . As this process is highly under constrained, CycleGAN couples this process with an inverse mapping $G_V : V \rightarrow U$. Correspondingly, a cycle consistent loss, i.e., u and $\hat{u} = G_V(G_U(u))$ is introduced for each

image in the source domain. Besides, there are two adversarial discriminators: D_U and D_V . D_U aims to distinguish between images in U and the translated images with generator G_V . Similarly, D_V is a discriminator that distinguishes between images in V and the translated images with generator G_U . The closed loop which is made by the cycle consistent loss allows images from either domain to be translated and then reconstructed. Given the above analysis, the objective loss function in CycleGAN contains two terms: the adversarial loss $L_{GAN}(u, v)$ inherited from the GAN-based model that tries to match the distribution of the generated images to the real image distribution in that domain, and the reconstruction loss $L_R(u, v)$ to ensure the learned dual mappings are consistent with the images:

$$L(u, v) = L_{GAN}(u, v) + L_R(u, v), \quad (1)$$

where $u \subseteq U, v \subseteq V$.

In the above equation, for the adversarial loss, we model the objective as:

$$L_{GAN}(u, v) = \log D_V(v) + \log(1 - D_V(G_U(u))) \\ + \log D_U(u) + \log(1 - D_U(G_V(v))) \quad (2)$$

where the first row models the adversarial loss for mapping from U to V , and the second row defines the adversarial loss for mapping from V to U . Specifically, $G_U(u)$ transforms images of domain U to domain V , and D_V aims to classify the translated image $G_U(u)$ and real image v . The adversarial process of G_U aims to minimize the objective against an adversary D_V that tries to maximize it. Similarly, the second row introduces a similar adversarial loss for mapping from V to U .

For the cycle consistent reconstruction, G_U and G_V satisfy backward reconstruction consistency as: $u \rightarrow G_U(u) \rightarrow G_V(G_U(u)) \approx u$ and $v \rightarrow G_V(v) \rightarrow G_U(G_V(v)) \approx v$. Typically, the reconstruction loss is defined as a pixel wise reconstruction error with L1 or L2 loss [8]. Without loss of generality, similar as CycleGAN, we use the L1 reconstruction loss as:

$$L_R(u, v) = \vartheta_u \|u - G_V(G_U(u))\|_1 + \\ \vartheta_v \|v - G_U(G_V(v))\|_1, \quad (3)$$

where ϑ_u , and ϑ_v are typically set to 10.

Given the detailed adversarial loss in Eq. (2) and the reconstruction loss in Eq.(3), CycleGAN aims to solve the following objective function:

$$G_U^*, G_V^*, D_U^*, D_V^* = \arg \min_{G_U, G_V} \max_{D_U, D_V} \mathbb{E}_{u, v \sim p_{data}} L(u, v). \quad (4)$$

IV. THE PROPOSED MODEL

In this section, we propose an overall Quality-aware GAN (QGAN) framework for unpaired image-to-image translation, which is based on CycleGAN. As shown in the overall loss function (Eq.(1)) of CycleGAN, it has a GAN-based loss term and a pixel-to-pixel level reconstruction loss. In fact, as the generated images are finally evaluated by human, it is important to generate visually qualified images. Nevertheless, human

rely on the high level abstractions of images for perceptual quality evaluation, which is neglected in this process. To generate quality-aware image translations, it is important to incorporate the objective of QGAN with human visual quality constraints. Thus, we define an overall loss function of the proposed QGAN framework as:

$$L(u, v) = L_{GAN}(u, v) + L_R(u, v) + L_Q(u, v) \quad (5)$$

where $L_R(u, v)$ and $L_{GAN}(u, v)$ share the similar formulations as the unpaired image translation task as introduced before. $L_Q(u, v)$ directly measures the quality of reconstructed images in U (i.e, $G_V(G_U(u))$) and real images in U (i.e, $u \in U$), and the reconstructed images in V (i.e, $G_U(G_V(v))$) and real images in V (i.e, $v \in V$). By comparing the quality between each image and its reconstructed version, the proposed QGAN framework could be generally applied to unpaired image-to-image translation without any paired images. In the following of this section, we provide two detailed implementations of the QGAN framework, i.e., two methods of the quality-aware loss $L_Q(u, v)$. We present the overall ideas of our proposed two models, as well as the CycleGAN model in Fig. 2. Specifically, we would first show how to implement a detailed quality-aware loss based on the classical image quality assessment measures (middle part of the Fig. 2). Besides, as CNNs show a huge success for capturing the higher content structure information, instead of applying pretrained deep networks for quality assessment [9], [13]–[15], we would also like to explore whether it is possible to adaptively model the high level quality loss (right part of the Fig. 2). Next, we introduce the implementations of these two quality-aware losses in detail.

A. QGAN_A: Classical IQA Loss

In Section II, we introduce some classical methods for IQA. As mentioned before, as most BR based IQA models either needed additional labeled image quality values or were only applicable to a specific domain, these BR based quality measures are not suitable for modeling the quality loss in the QGAN framework. Therefore, we plan to adopt FR based IQA models for quality loss modeling. In fact, there are various computational FR models for IQA by measuring the image quality consistently with human subjective evaluations, such as SSIM [16], FSIM [18] and GMSD [17]. Since the focus of this paper is not to design more sophisticated IQA measures, we choose FSIM method [18] for modeling the quality loss $L_Q(u, v)$ as it is profound with physiological and psychophysical evidences, and showed great success for modeling IQA. We call this proposed model as QGAN_A (Assessment).

Specifically, as well researched by physiologists and neuroscientists, visually discernable features coincide with those points that Fourier waves at different frequencies have congruent phases. Thus, by transforming images into a frequency domain, FSIM selects two low-level features based on the phase consistency and gradient magnitude. Additionally, the color characteristics are added to establish the FSIM. For more details of FSIM, please refer to Zhang *et al.* [18]. Thus, we build the $L_Q(u, v)$

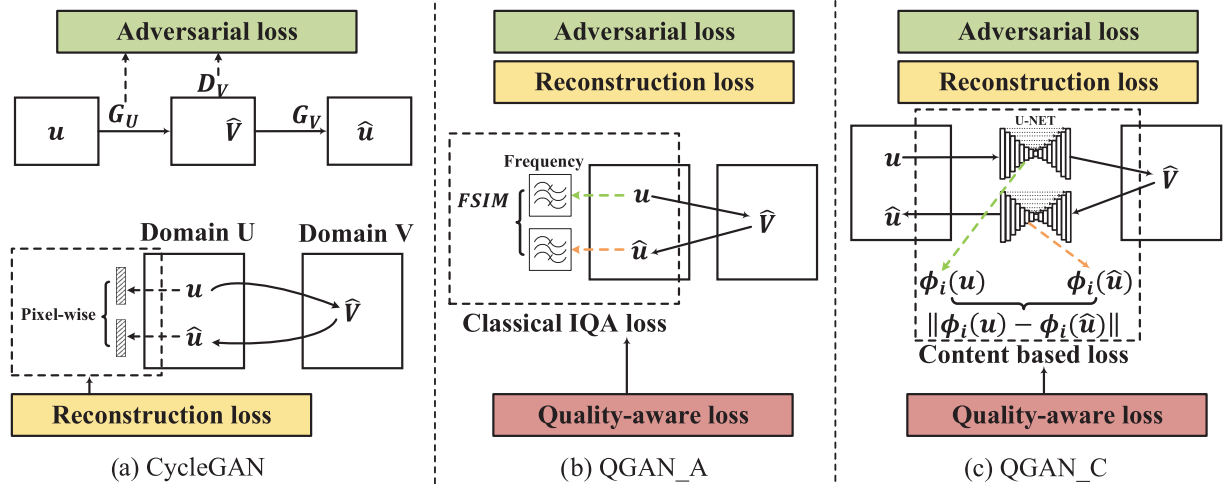


Fig. 2. Our proposed Quality-aware GAN-based models for unpaired image-to-image translation. The leftmost part shows the CycleGAN [3], and the remaining two parts shows our proposed two detailed models under the QGAN framework. In the middle part, our proposed QGAN_A model designs a quality loss term that is based on a classical image quality assessment model of FSIM. The QGAN_C model designs an adaptive content loss that captures the high level perceptual structure of images from the GANs without any labeled image pairs.

based on FSIM as:

$$L_Q(u, v) = \alpha_u [1 - FSIM(u, G_V(G_U(u)))]_1 + \alpha_v [1 - FSIM(v, G_U(G_V(v)))]_1, \quad (6)$$

where α_u and α_v are regularization parameters. $FSIM(x, y)$ is the quality similarity score calculated using the FSIM [18] algorithm. Then, by minimizing the loss function $L_Q(u, v)$, we encourage the generators (G_U, G_V) to generate images such that each input image and its reconstructed image have similar quality scores. In contrast, if the generators (G_U, G_V) could not accomplish the translation task well, the $FSIM$ value between an input image and its reconstructed image is closer to 0. Then, the loss function $L_Q(u, v)$ becomes larger. Therefore, based on FSIM loss, the quality of the generated image can be constantly utilized to induce a positive effect.

B. QGAN_C: Adaptive Content-Based Loss

In this subsection, we introduce how to design an adaptive perceptual quality-aware loss for unpaired image translation. We call the proposed QGAN framework QGAN_C (Content) as it models the higher level content of deep networks.

In fact, some previous works have shown that high-quality images can be generated by defining the high level perceptual losses based on high level features from pretrained neural networks [9], [13]. The intuition is that, the pixel-based loss functions focus on low level image information, which optimize the pixel-wise average between the ground truth images and the generated images and then lead to over-smooth results. Instead, the perceptual losses optimize the high level content losses from pretrained VGG networks, encouraging the ground truth images and the generated images are similar in the VGG feature space.

Despite the visually superior performance, we argue that these previous works have some limitations. Specifically, these pretrained VGG networks are capable of extracting high level features that are well trained for specific classification tasks. The

GAN-based approach also naturally achieves the high level feature learning of images with an adversarial learning process. In fact, as the high level features extracted from pretrained VGG networks are optimized for the image classification task, they are inferior when transferred to image translation task. Thus, we argue that it is better to design an adaptive content-based loss from the high level features from the GAN-based approach that is tailored to the image translation task. Next, we would detail how to design the adaptive content-based loss directly from the GAN framework.

Specifically, instead on a pretrained VGG network in related works [9], [13], we define the content-based loss as:

$$L_Q(u, v) = \beta_u \|\phi_i(u) - \phi_i(G_V(G_U(u)))\|_1 + \beta_v \|\phi_i(v) - \phi_i(G_U(G_V(v)))\|_1, \quad (7)$$

where β_u and β_v are regularization parameters, ϕ_i indicates the feature map located in the i -th convolutional (after activation) layer of the generator. In practice, we use the popular ‘‘U-NET’’ as the generator [8], which is shown in the rightmost part of Fig 2. As the parameters of the ‘‘U-NET’’ change in the adversarial training process, for any image u , its content features $\phi_i(u)$ adaptively updates during the training process of image translation task.

Please note that, recently researchers also proposed a Perceptual Adversarial Network (PAN) for enhancing image-to-image translation quality on the labeled image pairs under the GAN framework [35]. In PAN, besides the generative adversarial loss widely used in GANs, a perceptual adversarial loss (*PA loss*) is introduced to undergo an adversarial training process between the image generation network and the hidden layers of the discriminative network. In designing the *PA loss*, it is required to compare the quality of each generated image and its corresponding ground truth. PAN also demonstrates the possibility to be extended for the unpaired image translation task, which is achieved

by calculating the difference of the mean features on two domains. Nevertheless, the uniqueness of each individual image is neglected and smoothed by the average operation. Therefore, PAN could not well tackle the unpaired image translation. In contrast, we focus on the unpaired image translation task and propose an adaptive content-based loss, which considers each individual image by measuring image quality between each input and its corresponding reconstructed image. In summary, our proposed model is more robust for unpaired image-to-image translation.

V. MODEL TRAINING

Our proposed QGAN framework (Eq.(5)) with two detailed quality loss implementations, i.e., QGAN_A and QGAN_C, could be trained under a unified optimization framework. We show the training process of QGAN in Alg. 1. It includes training the discriminators D_A and D_B (from 4th to 9th line) and training the generators G_A and G_B (from 11th to 14th line). In practice, we set the number of iterations (i.e., n) to be 2–5 [1], [46], [47]. We use the RMSProp [48] solver as it performs well on highly non stationary problems. We initialize the learning rate for RMSProp as 0.00005. The clip parameter c clips value to a specified range, and c is set in [0.01, 0.1] [46]. For all experiments, the values of α_u and α_v are set as 15 times as large as that of ϑ_u and ϑ_v , β_u and β_v are set to 20 times as large as that of ϑ_u and ϑ_v . The feature map in Eq.(7) is chosen as $i = 6$, i.e., we choose $\phi_i(u)$ as the sixth convolutional (after activation) layer of the generator for the content-based loss. As the original loss of L_{GAN} is unstable during the training process, similar as many works [1], [3], [46], we train G_U and G_V to minimize $L_g(u, v) = (D_U(G_V(v)) - 1)^2 + (D_V(G_U(u) - 1))^2$, and train D_U and D_V using $L_d(u, v) = (1 - D_U(u))^2 + D_U(G_V(v))^2 + (1 - D_V(v))^2 + D_V(G_U(u))^2$. Specifically, Alg. 1 shows the training procedure for optimizing the proposed loss function.

For the generators (G_U and G_V) and discriminators (D_U and D_V), we use the architectures that are widely adopted in image-to-image translation. Specifically, we adopt the “U-NET” structure [8] for our generative networks as this structure is successfully applied in many image generation tasks [1], [3], [8]. The “U-NET” architecture contains two stride-2 convolutions, several residual blocks and two fractionally strided convolutions. This network contains 16 layers with skip connection between each layer i and layer $16 - i$, where $0 < i < 9$. Thus, the “U-NET” allows information to short across the network, and models more high level structure information in one layer. For the discriminator, we use the same CNN architecture as in [1], [3], [8], where each discriminator contains four layers.

VI. EXPERIMENTS

In this section, we conduct extensive experimental results to show the quality improvement of our proposed two methods (i.e., QGAN_A and QGAN_C) on unpaired image-to-image translation. We perform experiments on four datasets that are widely

Algorithm 1: The Algorithm of QGAN Framework

Input: real data U , real data V , batch size m , the number of discriminator iterations per generator iteration n , generator parameters Θ_U and Θ_V , discriminator parameters Ω_U and Ω_V , clipping parameter c .

- 1: Randomly initialize $\Theta_i, \Omega_i, i \in \{U, V\}$
- 2: **repeat**
- 3: **for** $t = 1$ to n **do**
- 4: get mini-batch $\{u^{(i)}\}_{i=1}^m$ from the real data U
- 5: get mini-batch $\{v^{(i)}\}_{i=1}^m$ from the real data V
- 6: $d \leftarrow \frac{1}{m} \sum_{i=1}^m L_d(u^{(i)}, v^{(i)})$
- 7: update $\Omega_U, \Omega_V \leftarrow$ RMSProp optimizer d
- 8: clip($\Omega_U, -c, c$) {Clip the weight of D_U }
- 9: clip($\Omega_V, -c, c$) {Clip the weight of D_V }
- 10: **end for**
- 11: get mini-batch $\{u^{(i)}\}_{i=1}^m$ from the real data U
- 12: get mini-batch $\{v^{(i)}\}_{i=1}^m$ from the real data V
- 13: $g \leftarrow \frac{1}{m} \sum_{i=1}^m [L_G(u^{(i)}, v^{(i)}) + L_R(u^{(i)}, v^{(i)}) + L_Q(u^{(i)}, v^{(i)})]$
- 14: update $\Theta_U, \Theta_V \leftarrow$ RMSProp optimizer g
- 15: **until** convergence

TABLE I
DATASET DESCRIPTION

Dataset	Num. of training images	Num. of test images
PHOTO-SKETCH	1990	388
OIL-CHINESE	2354	94
SUMMER-WINTER	1924	476
LABEL-FACADE	800	200

used for image translation: PHOTO-SKETCH [1], LABEL-FACADE [1], OIL-CHINESE [1], SUMMER-WINTER [3]. Table I shows the details of all these datasets, where the training and testing images are automatically divided in these datasets. The number of training images contains the images from both domains for training. Specifically, the PHOTO-SKETCH and the LABEL-FACADE datasets include the paired images between the corresponding two domains. In the model training process, we omit the paired correspondence on these two datasets, and we use the ground truth (GT) of the pair relationships for better visual evaluation.

For fair comparisons, firstly we choose CycleGAN [3] as a baseline as it is a state-of-the-art unpaired image translation model, and it shares similar ideas with many unpaired image-to-image translation models, including DualGAN [1] and DiscoGAN [4]. In the experimental setup process, we use the same settings as CycleGAN. Besides, as the pretrained VGG based perceptual loss functions have been widely used for enhancing the quality of the image translation tasks [9], [13], we combine the metrics of the pretrained VGG based perceptual loss function with the optimization function of CycleGAN as a baseline. We call this baseline as CycleGAN-VGG. The comparison between QGAN_C and CycleGAN-VGG would show whether the adaptive content loss in QGAN_C is superior than the pretrained perceptual loss from the VGG.

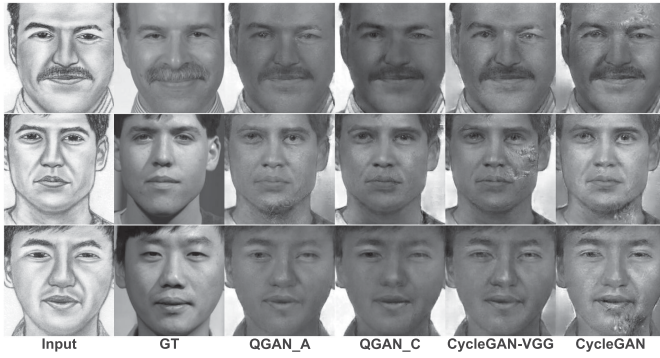


Fig. 3. Results of sketch→photo translation.

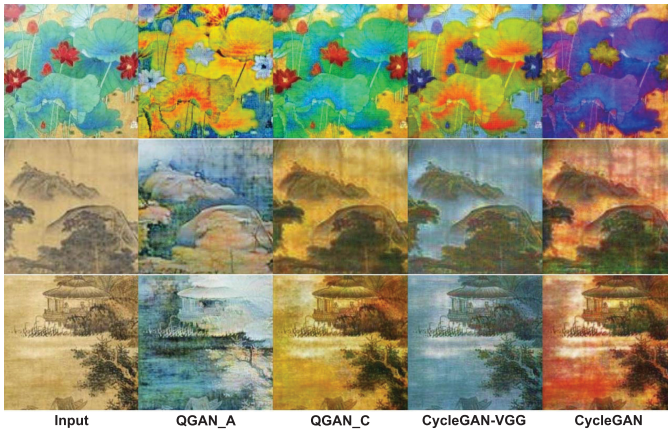


Fig. 4. Results of Chinese→oil translation.

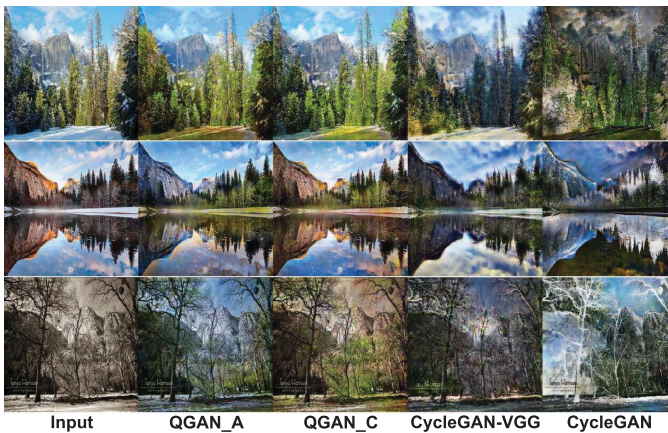


Fig. 5. Results of winter→summer translation.

A. Experimental Results

We perform extensive experimental results on four widely used tasks: sketch→photo (Fig. 3), Chinese→oil (Fig. 4), summer↔winter (Fig. 1, 5, 6), and label↔facade (Fig. 7). In all these tasks, the *Input* shows the input images from the source domain, and *QGAN_A* and *QGAN_C* are our proposed two methods. For better comparison, we also show the ground truth if the dataset contains the ground truth of the paired images. As can be seen from these figures, in almost all tasks, compared to CycleGAN and CycleGAN-VGG, our proposed two models

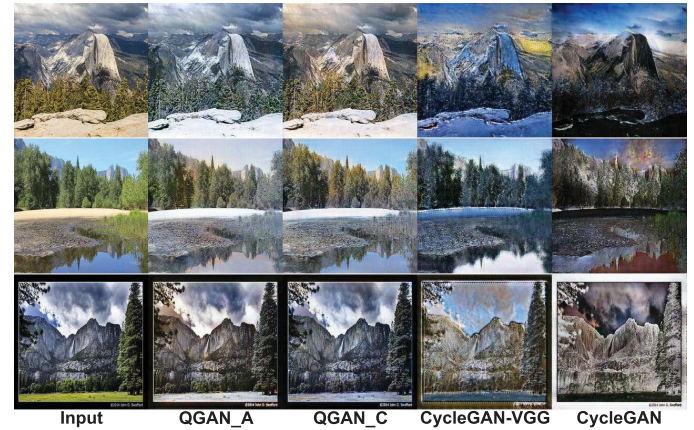


Fig. 6. Results of summer→winter translation.

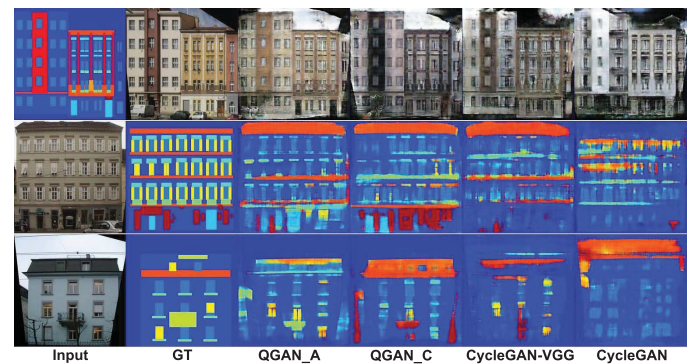


Fig. 7. Results of label↔facade translation.

generate quality-enhanced images with less blurry parts. E.g., in Fig. 3, the results of QGAN are less blurry with more details, while there exists blurry parts on the forehead that appeared in CycleGAN and CycleGAN-VGG.

We observe a typical phenomenon in Fig. 5 and Fig. 6, which performs image translation in domains of winter and summer. A key characteristic of the domain translation task is that snow would fall on objects in winter, while it disappears in summer. E.g., as shown in the upper-right corner in the last column of Fig. 6, the result of CycleGAN shows colorful patches. To analyze this phenomenon, we find that as the snow falls on the objects in an image taken in winter, some parts in this image contain colorful patches (e.g., the green mountain with white snow). The generator in CycleGAN would indecisively pick a color (e.g., the color of the snow or the color of the mountain) as this decision would not have a large impact on the pixel-wise based reconstruction loss. Luckily, our proposed two models could automatically distinguish the varying frequency of these colorful patches. Besides, we observe that snow is also not well captured in CycleGAN-VGG. We guess a possible reason is that CycleGAN-VGG relied on the pretrained network parameters from the ImageNet dataset, which is designed on the classification task with specific categories. As snow usually appears in limited images with small regions, it is also not well captured by the VGG network. In our proposed two models, the adaptive quality-aware loss is optimized for the current image

TABLE II

QUALITY SCORE EVALUATED ON SKETCH→PHOTO AND LABEL→FACADE WITH DIFFERENT METHODS. WE CALCULATE EACH MEASURE BETWEEN THE GENERATED IMAGE AND ITS CORRESPONDING GROUND TRUTH (GT)

Input and Method		QGAN_A	QGAN_C	CycleGAN-VGG	CycleGAN
GT and sketch→photo	SSIM	0.5683	0.5681	0.5412	0.5256
	FSIM	0.7749	0.7746	0.7658	0.7616
	GMSD	0.2128	0.2145	0.2148	0.2146
GT and label→facade	SSIM	0.1686	0.1233	0.0710	0.0545
	FSIM	0.6058	0.588	0.5897	0.5719
	GMSD	0.2816	0.2855	0.2849	0.2901

translation domain, in order to avoid the deficiencies of using pretrained network based perceptual loss. To better show the effectiveness of using the adaptive content features (*QGAN_C*) compared to the pretrained network, we present the results of the label↔facade task as an example. The pretrained ImageNet dataset does not contain the label category. As shown in Fig. 7, it is obvious that *QGAN_C* generate more details of edge. E.g., as shown in the last column, the outline of the label in *CycleGAN-VGG* disappears. In contrast, *QGAN_C* generates more detailed label information compared to the baselines.

When comparing our proposed two quality losses, the results of *QGAN_C* look similar to the results of *QGAN_A*, but *QGAN_A* sometimes produces images with better quality that meet human perception, such as in Fig. 3. We leave the quantitative quality comparisons of these two proposed models in the next part. To analyze the main reason, the adaptive content loss in *QGAN_C* relies on the training of the generator and it is more likely to be affected by the instability of the training process. In contrast, the classical IQA based loss function in *QGAN_A* is stable and independent of the process of GAN-based training, and it is easier in model tuning process.

B. Quantitative Evaluation

To better compare the quality of the generated images from different models, in this part, we conduct quantitative evaluation under different quality metrics. Specifically, we adopt three commonly used full reference image quality assessment methods: SSIM [16], FSIM [18] and GMSD [17]. These three methods measure the image quality that reflect the human visual experience from various aspects. The larger values of SSIM and FSIM denote better quality, and the smaller scores of GMSD denote better quality. Since all these measures rely on the paired image information, for the two tasks of sketch→photo and label→facade with ground truth of paired information, we can calculate the quality measure with the generated image and the ground truth in the test evaluation process. With the unpaired images in the remaining two datasets, we could not calculate these measures. It can be seen from Table II that our proposed methods perform consistently better than *CycleGAN-VGG* and *CycleGAN* under the three measures. E.g., *QGAN_A* improves over *CycleGAN-VGG* about 5% and more than 8% improvement over *CycleGAN* under the SSIM measure in sketch→photo task. Since these metrics evaluate human visual experience from various perspectives, we could empirically conclude the effectiveness of quality enhancement of our proposed models. Last but

TABLE III

MOS SCORE FOR DIFFERENT METHODS ON ALL TASKS

Tasks	QGAN_A	QGAN_C	CycleGAN-VGG	CycleGAN
sketch→photo	2.71	2.63	2.65	2.55
label→facade	3.26	3.03	2.91	2.61
Chinese→oil	3.13	3.73	2.95	2.79
winter→summer	3.91	3.76	2.56	2.06

not least, by comparing the results of *QGAN_A* and *QGAN_C*, we observe that for each assessment measure, the results of *QGAN_A* always outperform *QGAN_C*.

C. MOS Testing

Assessing the quality of generated image is an open question. Though we have conducted various measures for image quality assessment in previous part, human visual experience is the golden standard for assessing the quality of generated images. Thus, to better compare our proposed models with the baselines, we conduct a Mean Opinion Score (MOS) testing by human evaluation. This MOS testing avoids the shortcomings of each quality evaluation metric and gives the overall perceptual experience. To realize this, we design a MOS system that asks each rater to give a numerical indication of the perceived quality of each generated image from each method. Specifically, we ask 24 raters to assign a score from 1–5, where 1 denotes the lowest perceived quality, and 5 is the highest perceived quality. In the system design process, the images generated by different methods are listed randomly. Also, we randomly repeat some images in the system to see whether the rater gives the same rating to the same image that appears in different orders. We remove the raters that give different ratings to the same image. We evaluate on the four tasks: sketch→photo, Chinese→oil, label→facade, and winter→summer. Since manual scoring is time-consuming and expensive, researchers often run small datasets with a random selection to approximate human perception, we randomly select a third of all test image of four tasks. Thus, each rater rated four tasks with 772 images. The final MOS scores of all methods are summarized in Table III. As can be seen from this table, the MOS testing results show that *QGAN* framework outperforms *CycleGAN* and *CycleGAN-VGG* on all tasks. Generally, *CycleGAN-VGG* shows better results compared to *CycleGAN* by adding the pretrained perceptual loss, and our proposed *QGAN_C* further improves *CycleGAN-VGG* with the adaptive perceptual loss. When comparing the results of all models, generally our proposed *QGAN_A* shows the best performance, followed by our proposed *QGAN_C* model. However, we observe that for the Chinese→oil translation task, *QGAN_C* shows better results than *QGAN_A* with MOS testing. Meanwhile, in Table II, the quantitative results show that *QGAN_A* performs better than *QGAN_C* on four quality measures. This inconsistency between the quality measures and the MOS score also observed by previous works [9]. We guess a possible reason is that human visual evaluation is quite subjective, and each quality measure could only partially reflect human visual experience. Nevertheless, as our proposed two models consistently outperform the baselines to a large margin, we could conclude the effectiveness of our proposed two models for quality-enhanced

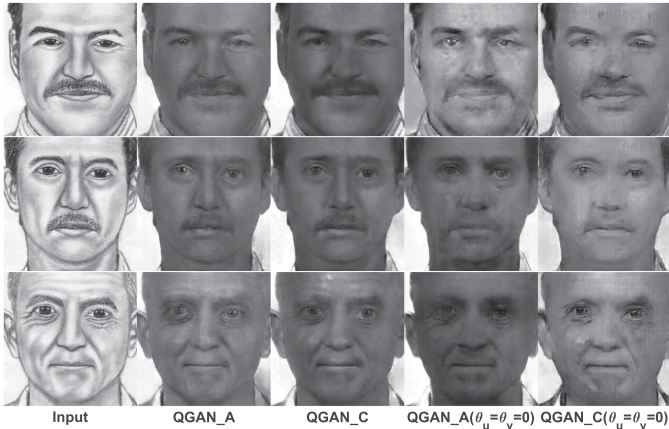


Fig. 8. Variants of our proposed methods for the task of sketch→photo. Specifically, the last two columns are two simplifications of our proposed models without the reconstruction loss.

unpaired image translation. Also, our proposed QGAN_A that relies on the classical image quality measures outperforms QGAN_C in most situations.

D. Analysis of the Objective Function

In our proposed QGAN framework, in addition to the adversarial loss, there are two important regularization terms: a reconstruction loss $L_R(u, v)$ and a quality loss $L_Q(u, v)$. In this part, we would demonstrate the effectiveness of the proposed framework from these two aspects.

Impact on the reconstruction loss. To study the effect of the reconstruction loss in QGAN framework, we design a simplified model that discards the reconstruction loss in QGAN framework and redefine the objective function as:

$$L(u, v) = L_{GAN}(u, v) + L_Q(u, v). \quad (8)$$

In other words, the regularization terms in Eq.(3) of the reconstruction loss are set as: $\theta_u = \theta_v = 0$. We show some qualitative examples in Fig. 8 under this setting, where the last two columns are the generated images without the reconstruction loss in our proposed two models, respectively. We observe that the generated images are blacker in color with more noisy points as compared to the QGAN framework that considers the reconstruction loss. We guess a possible reason is that QGAN_A relies on the FSIM based quality loss that emphasizes on the low-level features in frequency. Thus, only relying on quality assessment based loss without any reconstruction loss makes it unstable in the training process, leading to a smooth distribution that prefers black images. QGAN_C($\theta_u = \theta_v = 0$) relies on the higher level content structure based similarity without any pixel level constraint, thus also degrades quality of the generated images. Therefore, we conclude that all terms are critical in the model training process.

Impact on the choice of the classical image quality measures. Image quality assessment models could be classified into FR and BR measures. In our proposed QGAN_A model, we use the FR based quality measure for modeling the quality loss. Thus, it is natural to ask, is it possible to define image quality

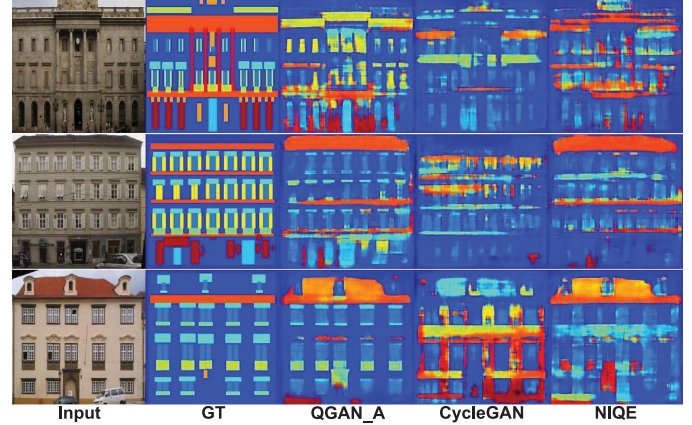


Fig. 9. Variants of our proposed methods for the task of facade→label.

loss based on the BR methods? To answer this question, in this part we evaluate the effects of using a blind reference IQA measure based loss in the proposed QGAN framework. Without loss of generality, we select NIQE [38] as a typical blind-reference method to evaluate the image quality. NIQE uses measurable deviations from statistical regularities observed in natural images without labeling on human-rated distorted images. Hence, it is suitable for many image-to-image translation tasks without any human labeling effort. A smaller score of NIQE indicates better perceptual quality, so we change the classical IQA loss function in Eq.(6) and redefine it as:

$$L_Q(u, v) = \vartheta_u NIQE(G_V(G_U(u))) + \vartheta_v NIQE(G_U(G_V(v))), \quad (9)$$

where $NIQE(x)$ calculates the no-reference image quality score for image x using the NIQE method.

In Fig. 9, we show the facade→label task of the QGAN results with the above defined BR quality loss, where the last column shows the results of using NIQE based loss in the QGAN framework. It is visually obvious that the NIQE based quality loss does not improve the quality of generated images compared to CycleGAN. In fact, despite NIQE can compute the quality for an image, NIQE is based on a certain priori knowledge about natural images. For different tasks, the target images are possibly not consistent with the certain priori knowledge. If we blindly minimize the score of NIQE, the generated images are hard to meet human visual experience. For example, in label→facade task [1], the NIQE score of a label image is about 6, while the NIQE score of a facade image is about 9. Please note that, besides the NIQE measure, nearly all blind reference based models need to rely on the prior assumption of the image domains. Thus, we conclude that FR based measures are more suitable in designing image quality-aware loss evaluation.

E. Discussion of Our Proposed Two Models

In our proposed QGAN framework, there are two kinds of models (QGAN_A and QGAN_C). In this subsection, we compare the experimental results and discuss the strengths and weaknesses of them.

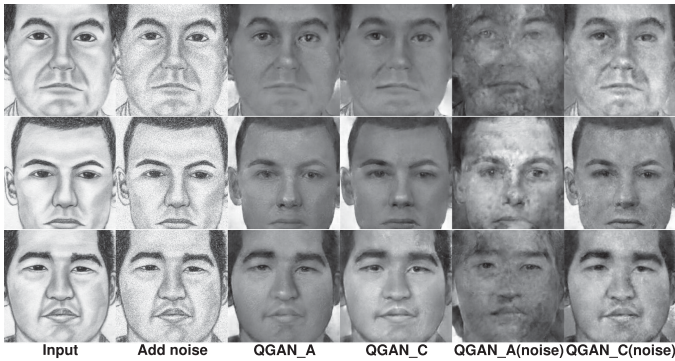


Fig. 10. Typical failure cases of QGAN_A. “Add noise” represents we add Gaussian noise (mean = 0, var = 0.001) to the original images.

QGAN_A relies on a static classical IQA measure, i.e., FSIM, to calculate the quality loss. As FSIM models the phase congruency and other features that are based on physiological and psychophysical evidences from human visual systems, the FSIM based quality loss is intuitive to understand. Besides, this classical IQA based measure does not rely on any intermediate architecture of the generator. Therefore, QGAN_A does not introduce additional parameters in the model training process and is empirically easier to train compared to QGAN_C. However, there are also limitations of QGAN_A. On one hand, as there is no universal golden-standard metric for image quality evaluation, the performance is limited by the specific IQA based quality measure. E.g., the phase congruency (PC) feature in FSIM is sensitive to image noise [49]. On the other hand, as the calculation of FSIM is computationally expensive, the runtime of QGAN_A is longer than QGAN_C.

Different from QGAN_A, QGAN_C does not introduce any prior knowledge of IQA and relies on an adaptive content-based loss. Specifically, QGAN_C uses the features extracted by the generators, so it can deal with any dataset. Furthermore, the “U-NET” architecture of generators can eliminate noisy effects from inputs. By comparing the feature maps extracted from this architecture, denoising effect is automatically achieved for some noisy input images. However, as QGAN_C is an adaptive algorithm, and the proposed adaptive content loss relies on the intermediate feature extractor of the generator, this proposed model introduces additional parameters, i.e., feature map ϕ_i after the i -th convolutional layer. In practice, we find this parameter adds additional tuning complexity considering the instability in the GAN training process.

For most experimental cases, the results of QGAN_A outperform QGAN_C. However, as discussed before, the QGAN_A would fail if the training data contains noise. To empirically validate this, we perform an experiment as shown in Fig. 10. In this figure, we add Gaussian noise (mean = 0, var = 0.001) to inputs for the sketch \rightarrow photo translation task. The last two columns are the generated images with the noisy inputs. We observe QGAN_A performs much worse than QGAN_C. The reason is that: the FSIM based loss contains the phase congruency measure, which is sensitive to noises [49]. In contrast, the denoising effect is automatically achieved by the “U-NET” architecture of generators in QGAN_C.

VII. CONCLUSION

In this paper, we revisited the problem of unpaired image-to-image translation, and designed a unified QGAN framework for quality-aware unpaired image translation. In the QGAN framework, a quality-aware loss term is explicitly incorporated in the optimization function. Specifically, we designed two detailed implementations of the quality loss, i.e., QGAN_A and QGAN_C, that considered the classical quality assessment model and the adaptive high level content structure information from deep network. Extensive quantitative comparisons against prior models, as well as a mean opinion score test clearly showed the superior quality of our proposed framework and the two detailed implementations. In the future, we would like to apply our proposed framework to image applications that rely on image quality, e.g., image super resolution.

REFERENCES

- [1] Z. Yi, H. Zhang, P. Tan, and M. Gong, “Dualgan: Unsupervised dual learning for image-to-image translation,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2868–2876.
- [2] A. Elgammal, B. Liu, M. Elhoseiny, and M. Mazzone, “Can: Creative adversarial networks, generating ‘art’ by learning about styles and deviating from style norms,” in *Proc. IEEE Int. Conf. Comput. Commun.*, 2017, pp. 96–103.
- [3] 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.*, 2017, pp. 2242–2251.
- [4] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, “Learning to discover cross-domain relations with generative adversarial networks,” in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1857–1865.
- [5] I. Goodfellow *et al.*, “Generative adversarial nets,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
- [6] D. He *et al.*, “Dual learning for machine translation,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 820–828.
- [7] M. Wang, R. Hong, X.-T. Yuan, S. Yan, and T.-S. Chua, “Movie2comics: Towards a lively video content presentation,” *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 858–870, Jun. 2012.
- [8] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5967–5976.
- [9] C. Ledig *et al.*, “Photo-realistic single image super-resolution using a generative adversarial network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4681–4690.
- [10] Z. Chen, T. Jiang, and Y. Tian, “Quality assessment for comparing image enhancement algorithms,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3003–3010.
- [11] X. Wang and A. Gupta, “Generative image modeling using style and structure adversarial networks,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 318–335.
- [12] T.-C. Wang *et al.*, “High-resolution image synthesis and semantic manipulation with conditional GANs,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8798–8807.
- [13] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
- [14] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.
- [15] G. Guo, H. Wang, C. Shen, Y. Yan, and H.-Y. M. Liao, “Automatic image cropping for visual aesthetic enhancement using deep neural networks and cascaded regression,” *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2073–2085, Aug. 2018.
- [16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
- [17] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, “Gradient magnitude similarity deviation: A highly efficient perceptual image quality index,” *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014.

- [18] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.
- [19] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2414–2423.
- [20] X. Mao *et al.*, "Least squares generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2794–2802.
- [21] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.
- [22] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 184–199.
- [23] M. Morrone and D. Burr, "Feature detection in human vision: A phase-dependent energy model," *Roy. Soc. Lond. B, Biol. Sci.*, vol. 235, no. 1280, pp. 221–245, 1988.
- [24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [25] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, 2015.
- [26] C. Luo, B. Ni, S. Yan, and M. Wang, "Image classification by selective regularized subspace learning," *IEEE Trans. Multimedia*, vol. 18, no. 1, pp. 40–50, Jan. 2016.
- [27] K. Zhang *et al.*, "Image-enhanced multi-level sentence representation net for natural language inference," in *Proc. IEEE Int. Conf. Data Mining*, 2018, pp. 747–756.
- [28] D. Liu *et al.*, "Robust single image super-resolution via deep networks with sparse prior," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3194–3207, Jul. 2016.
- [29] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "CamStyle: A novel data augmentation method for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1176–1190, Mar. 2019.
- [30] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1983–1992.
- [31] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 270–279.
- [32] R. Li, S. Wang, Z. Long, and D. Gu, "UnDeepVo: Monocular visual odometry through unsupervised deep learning," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 7286–7291.
- [33] P. P. Srinivasan, T. Wang, A. Sreelal, R. Ramamoorthi, and R. Ng, "Learning to synthesize a 4D RGBD light field from a single image," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2262–2270.
- [34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.
- [35] C. Wang, C. Xu, C. Wang, and D. Tao, "Perceptual adversarial networks for image-to-image transformation," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4066–4079, Aug. 2018.
- [36] M. A. Saad, A. C. Bovik, and C. Charrier, "A DCT statistics-based blind image quality index," *IEEE Signal Process. Lett.*, vol. 17, no. 6, pp. 583–586, Jun. 2010.
- [37] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 50–63, Jan. 2015.
- [38] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.
- [39] Q. Wu *et al.*, "Blind image quality assessment based on rank-order regularized regression," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2490–2504, Nov. 2017.
- [40] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.
- [41] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rating image aesthetics using deep learning," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2021–2034, Nov. 2015.
- [42] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1733–1740.
- [43] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, "A deep neural network for image quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 3773–3777.
- [44] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1275–1286, Jun. 2015.
- [45] H. R. Sheikh, A. C. Bovik, and L. Cormack, "No-reference quality assessment using natural scene statistics: JPEG2000," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1918–1927, Nov. 2005.
- [46] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, vol. 70, pp. 214–223.
- [47] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5769–5779.
- [48] G. Hinton, N. Srivastava, and K. Swersky, "Rmsprop: Divide the gradient by a running average of its recent magnitude," *Neural Netw. Mach. Learn., Coursera Lecture 6.5*, 2012.
- [49] P. Kovesi, "Phase congruency: A low-level image invariant," *Psychol. Res.*, vol. 64, no. 2, pp. 136–148, 2000.



Lei Chen received the B.S. degree from Anhui University, Hefei, China, in 2016. He is currently working toward the M.S. degree with the Hefei University of Technology, Hefei, China. His research interests include multimedia analysis and data mining.



Le Wu received the Ph.D. degree from the University of Science and Technology of China, Hefei, China. She is currently an Assistant Professor with the Hefei University of Technology, Hefei, China. She has authored or coauthored more than 30 papers in referred journals and conferences. Her research interests include data mining, recommender systems, and social network analysis. She is the recipient of the Best of SDM 2015 Award, and the Distinguished Dissertation Award from China Association for Artificial Intelligence 2017.



Zhenzhen Hu received the Ph.D. degree from the Hefei University of Technology (HFUT), Hefei, China, in 2014, under the supervision of Prof. Jianguo Jiang and Prof. Richang Hong. She is currently an Associate Professor with the School of Computer and Information, HFUT. She was a Research Fellow with Nanyang Technological University, Singapore, directed by Prof. Yonggang Wen. Her research interests include cross-media computing and computer vision.



Meng Wang received the B.E. degree and the Ph.D. degree in the special class for the gifted young from the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively. He is currently a Professor with the Hefei University of Technology, Hefei, China. He has authored more than 200 book chapters, journals, and conference papers in his research topics, which include multimedia content analysis, computer vision, and pattern recognition. He is the recipient of the ACM Special Interest Group on Multimedia Rising Star Award 2014. He is an Associate Editor for the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON MULTIMEDIA, and IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.