SDP-GAN: Saliency Detail Preservation Generative Adversarial Networks for High Perceptual Quality Style Transfer

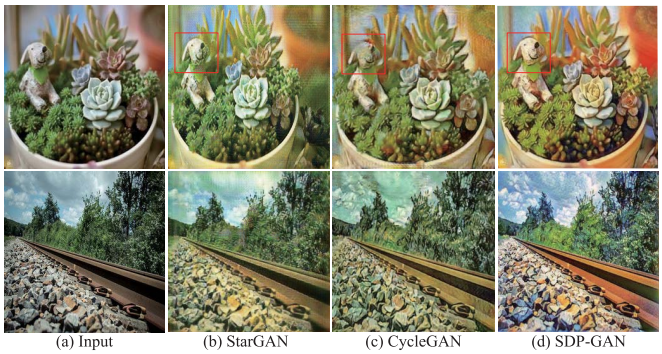
Ru Li, *Student Member, IEEE*, Chi-Hao Wu, Shuaicheng Liu, *Member, IEEE*, Jue Wang, *Senior Member, IEEE*, Guangfu Wang, Guanghui Liu, *Senior Member, IEEE*, and Bing Zeng, *Fellow, IEEE*

***Abstract*— The paper proposes a solution to effectively handle salient regions for style transfer between unpaired datasets. Recently, Generative Adversarial Networks (GAN) have demonstrated their potentials of translating images from source domain X to target domain Y in the absence of paired examples. However, such a translation cannot guarantee to generate high perceptual quality results. Existing style transfer methods work well with relatively uniform content, they often fail to capture geometric or structural patterns that always belong to salient regions. Detail losses in structured regions and undesired artifacts in smooth regions are unavoidable even if each individual region is correctly transferred into the target style. In this paper, we propose SDP-GAN, a GAN-based network for solving such problems while generating enjoyable style transfer results. We introduce a saliency network, which is trained with the generator simultaneously. The saliency network has two functions: (1) providing constraints for content loss to increase punishment for salient regions, and (2) supplying saliency features to generator to produce coherent results. Moreover, two novel losses are proposed to optimize the generator and saliency networks. The proposed method preserves the details on important salient regions and improves the total image perceptual quality. Qualitative and quantitative comparisons against several leading prior methods demonstrates the superiority of our method.**

***Index Terms*— Generative adversarial network, style transfer, detail preservation.**

I. Introduction

THE task of image-to-image translation is to capture special characteristics for one image collection and figure out how these characteristics could be translated to target image collection, e.g., images to semantic labels. Many researches have produced attractive transfer results, where image pairs are required [1]–[6]. However, obtaining paired training data is difficult. Some style transfer works avoid the need for such paired datasets by introducing the unpaired image-to-image translation using generative adversarial networks [7], [8].



(a) Input (b) StarGAN (c) CycleGAN (d) SDP-GAN

Fig. 1. The top result of CycleGAN loses some information in the doll's head region, and the results of StarGAN cannot produce desired style transformation, while our method solves these problems well.

Although large improvements have been achieved with learning-based stylization over unpaired datasets, state-of-the-art (SOTA) methods often fail to produce satisfying visual results. These methods aim to transfer the holistic style from the source domain X to a new domain $\hat{Y}$ that has identical distribution to the target domain Y. However, such a translation does not guarantee a high perceptual quality style transfer for the whole image, even if each region in an individual x is correctly transferred into the target style. For example, Fig. 1 (b) and (c) show some style transfer results with a uniform style across the whole image. However, more favorable style transfer results can be obtained when image content is better preserved in certain regions, especially salient regions, as shown in the red box in Fig. 1 (d).

Saliency detection aims to locate important regions or objects in images [9], which is analogous to the selective process in the human visual system. It provides benefits for a broad range of applications such as segmentation [10] and classification [11] by various means. In this work, we believe that saliency information is also significant for style transfer.

Manuscript received December 30, 2019; revised June 9, 2020 and September 5, 2020; accepted October 21, 2020. Date of publication November 16, 2020; date of current version November 20, 2020. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61872067, Grant 61720106004, and Grant 62071097, in part by the "111" Projects under Grant B17008, and in part by the Sichuan Science and Technology Program under Grant 2019YFH0016. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Marta Mrak. *(Corresponding authors: Guanghui Liu; Shuaicheng Liu.)*

Ru Li, Shuaicheng Liu, Guanghui Liu, and Bing Zeng are with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: guanghuiliu@uestc.edu.cn; liushuaicheng@uestc.edu.cn).

Chi-Hao Wu, Jue Wang, and Guangfu Wang are with Megvii Technology, Chengdu 610000, China.

Digital Object Identifier 10.1109/TIP.2020.3036754

While conventional style transfer methods [7], [8] excel at generating images with relatively uniform content (e.g. sky and landscape photos, which are distinguished more by texture than by geometry), they often fail to capture geometric or structural patterns that always belong to salient regions.

According to our observation, traditional approaches often produce two types of problems in salient regions: loss of details in structured regions, and excess of details in smooth regions. One possible explanation for this is that previous models rely heavily on convolution to model the dependencies across different image regions. To achieve better perceptual quality, the transformation should preserve more edge details in structured regions and reduce unwanted artifacts in smooth regions, while the overall style is maintained. For learning-based methods, increasing the size of the convolution kernels can improve the representational capacity of the network but doing so also loses the computational and statistical efficiency. To solve such problems, Liu *et al.* have proposed to improve style transfer results by preserving salient information of content images, which simply adds a localization network to calculate the region loss. However, it tends to generate discontinuous salient regions and background if it cannot keep the balance between style loss and region loss [12]. We propose a GAN-based method to preserve saliency detail information while correctly mapping unpaired images from source domain to target domain, named SDP-GAN. The SDP-GAN introduces an extra saliency network that concurrently predicts the saliency map, which helps the calculation of newly proposed objective functions. The encoder part of the saliency branch is also concatenated into the generator network to yield a smooth overall style transformation, with content details properly preserved in salient regions.

We trained and evaluated SDP-GAN on a variety of style transfer tasks, including different artistic styles and cartoon styles. Fig. 1 shows the comparisons with two classic GAN-based style transfer methods in terms of detail preservation and style transformation for Van Gogh style. The top result of CycleGAN apparently loses some information in the doll's head region. This artifact appears not only in CycleGAN, but in other latest style transfer methods [8], [13], [14] as well. Our top result demonstrates that the stylization quality in salient regions is important and should be handled properly. For the comparisons with StarGAN, the proposed method shows a more attractive style transformation than StarGAN. In addition to visual comparisons, we also organized quantitative comparisons and a user study against several GAN-based style transfer methods to validate the superior performance of SDP-GAN. In quantitative comparisons, the proposed SDP-GAN achieves the best Inception score (IS) [15] and Fréchet Inception distance (FID) [16] score. As for the user study, our results are favored by a majority of viewers (over 50%). Overall, the main contributions are:

- We propose a GAN-based approach that effectively preserves details of salient areas and learns the stylization using unpaired image sets by adding a saliency network. The sub-network provides saliency features to help the generator and generates saliency maps to constrain content loss simultaneously.
- We introduce two novel loss functions in our architecture. a) We introduce a saliency-constrained content loss defined as ℓ_1 sparse regularization, which applies saliency maps to constrain conventional content loss. b) The saliency content loss is proposed to further maintain the content information, which minimizes the difference between saliency results generated from input and from stylized image.
- We provide qualitative and quantitative comparisons with several SOTA GAN-based methods on artistic stylization using SDP-GAN, showing its superiority in image quality over existing methods.

II. Related Work

A. Style Transfer

Many non-photorealistic rending (NPR) methods have been developed since the mid-1990*s*, and nowadays NPR is a firmly established field in computer graphics [17]. There are some works proposed to mimic specific artistic styles [18]. However, these methods use low-level image features and often fail to capture image structures effectively, such as making object boundaries clear. Inspired by the development of convolution neural networks (CNN), Gatys *et al.* first applied CNN feature activations to recombine the content of a given photo and the style of famous artworks [19]. Their subsequent work [13] used the feature maps of a pre-trained VGG network to optimize the content, and captured texture information of the style image using the Gram matrix [20]. However, the algorithm of Gatys *et al.* does not perform well in preserving the coherence of fine structures and details. Also, it generally fails for photorealistic synthesis, due to the limitations of Gram-based style representation. To solve the problem, some derived Gram-based representation methods are put forward to encode style information [21]–[23]. Then, some efficient methods are proposed to optimize a generative model offline and produce the stylized image with a single forward pass [24]–[26]. Johnson *et al.* aimed to pre-train a feed-forward style-specific network and produced a stylized result with the forward pass at the testing stage [4]. The design basically follows Gatys *et al.* [13], which suffers from the same aforementioned issues. Recently, some structure-preserving style transfer methods are proposed [12], [27], [28]. Liu *et al.* focused on salient regions in style transfer with the help of region loss, computed by introducing a localization network [12]. Cheng *et al.* used the depth and edge information to construct a structure representation to solve the disruption of content [27]. Liu *et al.* applied saliency map to fuse the real image and stylized image, generating an output with stylized salient regions and real background regions [28]. All these methods, however, use a single style image and generate results whose style heavily depends on the chosen style image.

For arbitrary style transfer, a few methods [29]–[31] holistically adjusted the content features to match the statistics of the style features. Isola *et al.* [5] and Li *et al.* [32] applied generative adversarial learning to achieve the transformation between two domains. Zhu *et al.* further introduced

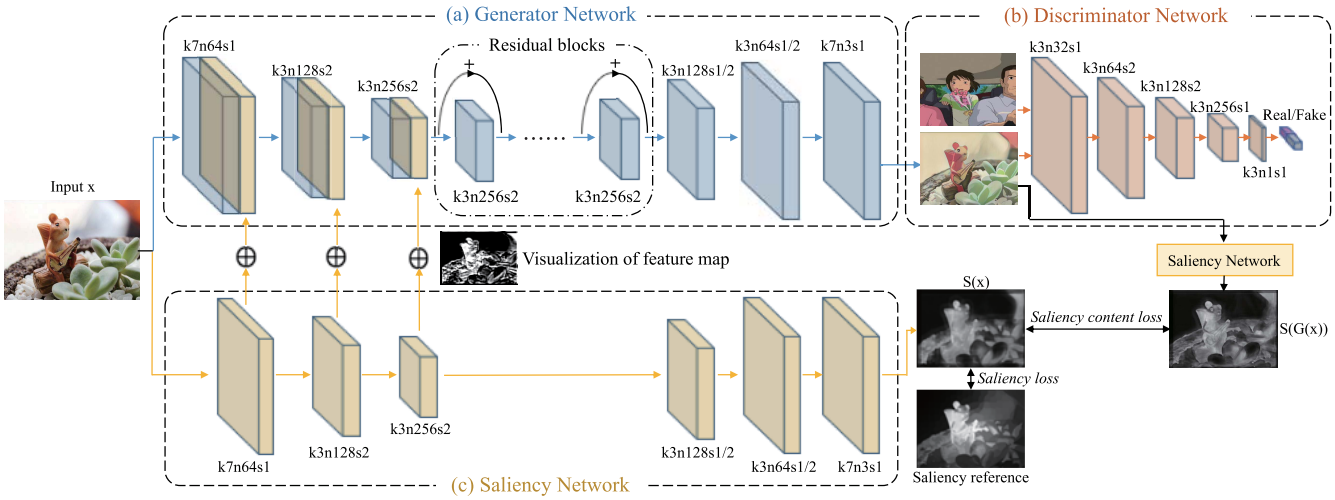


Fig. 2. The proposed model seeks to generate results with proper saliency content information, while still transforming image style comparable to the target dataset. The pipeline exhibits the architecture of the proposed SDP-GAN, in which k represents the kernel size, n is the number of feature maps and s is the stride in each convolutional layer. The pipeline consists of two streams operating different functions. **The top stream** processes the inputs through down-convolution layers, residual blocks and up-convolution layers so as to generate results with target style. In turn, **the bottom stream** produces a saliency map to constrain content loss and provides saliency features to the generator simultaneously. The saliency network is optimized by minimizing the saliency loss between the saliency output and saliency reference. The saliency content loss computes the difference between the saliency results of input image $S(x)$ and generated image $S(G(x))$.

CycleGAN to train unpaired image datasets [7]. Some methods [14], [33], [34] aimed at obtaining one single model to transfer multiple artistic styles. Chen *et al.* introduced CartoonGAN to generate sharp edges by slightly modifying adversarial loss [8]. However, these methods cannot perform well when encountering structured regions with many details or smooth regions with unneglectable continuity. The methods [35], [36] concentrated on attention-based style transfer techniques, which lead to some improvement, but still unsatisfactory for generating high-quality results. The saliency information plays an important role in leading to a better perspective experience. To solve such problems that appeared in conventional style transfer methods, we introduce a supplementary saliency network and two corresponding losses to improve the total performance.

B. Saliency Detection

Saliency detection has received great interest for many years. Itti *et al.* first proposed a saliency model that simulates the visual search of humans [37]. The method predicts saliency maps considering low-level features at multiple scales. Some subsequent works introduced bottom-up saliency models based not only on low but mid and high-level image features [38], [39]. Eleonora *et al.*'s ensembles of Deep Networks (eDN) was the first attempt at predicting saliency with a network [40]. Then, a number of deep learning solutions, including generative adversarial networks, have been put forward to improve the detection performance [41]–[45]. Saliency detection is intrinsic to various tasks such as image caption generation [46], image segmentation [47] and style transfer [48]. Zhang *et al.* [49] applied some handcrafted saliency maps as initial reference maps first and then optimized them to obtain accurate saliency results. We apply a similar idea as [49] to get the saliency reference.

III. SDP-GAN

We propose a GAN-based framework to achieve unpaired image transformation from a source domain X to a target domain Y. The framework aims to generate artistic images with more content details in the salient regions while still maintaining the desirable target style. Like common GAN frameworks, the generator G learns the mapping function between different domains, while the discriminator D aims to optimize G by distinguishing source domain images from generated ones. We further build a saliency sub-network S to simultaneously obtain a saliency map used to constrain the generator network and our new objective function. To better demonstrate the effectiveness of the framework, we adopt a wide diversity of real photos $\{x_i\}_{i=1,\ldots,N} \in X$ as our source domain data, and a collection of artistic images $\{y_j\}_{j=1,\ldots,M} \in Y$, as the target domain data. The saliency images $\{z_k\}_{k=1,\ldots,N}$, $z_k \in Z$ form another data domain, and the data distributions of these three domains are denoted as $x \sim p_{\text{data}}(x)$, $y \sim p_{\text{data}}(y)$ and $z \sim p_{\text{data}}(z)$, respectively.

The proposed SDP-GAN is capable of achieving enjoyable style transformation with necessary detail preservation. We illustrate our network architecture in Section III-A, and the objective function is presented in Section III-B.

A. Network Architecture

We present the generator network G, discriminator network D and the saliency network S in Fig. 2. They are simultaneously optimized during the training process. Specifically, G transfers the input image into a specific artistic style and S is used to generate a saliency map for using in the loss function. The addition of the saliency network S as well as extra losses enable the generator to not only learn the mapping but keep the details and continuity within desired regions.

The generator network is constructed with the sequence of a 7×7 convolutional layer, followed by two down-convolution blocks with stride 2, eight residual blocks [50], two transposed convolutional blocks with stride 1/2 for upsampling, and another 7×7 convolutional layer (Fig. 2 (a)). The saliency network has a similar structure with G in down-convolution and up-convolution blocks (Fig. 2 (c)), but it does not include extra residual blocks as presented in G. We found the simpler design provides relatively accurate saliency results and improves efficiency.

Although adding the saliency branch is effective for preserving content details in salient regions, purely using the saliency map to constrain the loss function often causes the loss of artistic style at salient regions and produces discontinuity at non-salient regions. To alleviate such problems, we concatenate the features of the first few convolutional blocks in the saliency network into the corresponding layers in the generator network. Adding saliency features to down-sample convolutional layers help to fuse image features and saliency features better. We just connect the saliency features to image features at down-sample convolutional layers because the quality of generated images will be influenced if the feature connection is also applied in up-sample convolutional layers. The encoder tends to extract image features and decoder is designed to implement different tasks. The up-sample layers of generator and saliency network play absolutely different roles. The modification offers additional saliency information into the encoding procedure, empowering the decoder of the generator to produce stylized images that have fluent transitions from region to region.

For the discriminator network D, we apply PatchGANs [5], [32] to classify each image patch into a real or fake one. Overlapped patches with size 70×70 are cropped from generated or real artistic images for training. D begins with flat layers (a 3×3 convolutional layer with stride 1 and a Leaky ReLU (LReLU) layer), followed by two convolutional blocks with stride 2. Then, a feature reconstruction block with stride 1 and a 3×3 convolutional layer are applied to obtain the classification results (Fig. 2 (b)). Such a simple patch-level discriminator uses fewer parameters and can work on arbitrarily-size images.

B. Loss Function

Based on style transfer properties, we design our objective function to include the following four losses: (1) the adversarial loss $L_{\mathrm{GAN}}(G, D)$, which drives the generator network to achieve the desired manifold transformation; (2) the image content loss $L_{con}(G, D, S)$, which preserves the image content during stylization; (3) the saliency content loss $L_{con\_s}(G, D, S)$, which minimizes the difference between saliency results generated from input and from stylized image in order to further retain the content consistency; (4) the saliency loss $L_{sali}(S)$, which optimizes the saliency network to obtain relatively accurate saliency maps. The full objective function is:

$$\begin{aligned} L(G, D, S) = {} & L_{\mathrm{GAN}}(G, D) + w_{con} L_{con}(G, D, S) \\ & + w_{con\_s} L_{con\_s}(G, D, S) + w_{sali} L_{sali}(S), \end{aligned} \quad (1)$$

where w controls the relative importance of these losses. w_{con} has great influence on the balance of style transformation and content preservation. A larger w_{con} produces images with more content information from the input, generating images that are not stylized enough. However, a small w_{con} learns the stylization excessively so that the semantic content information cannot be preserved well. To strike a balance, we set w_{con} to be 0.25 at the initial stage to correctly preserve content information with moderate style transformation. Then, w_{con} is gradually decreased to obtain better style transformation results after training is stable and the semantic content information has been properly reconstructed at the initial stage. The influence of different w_{con} is illustrated in Section V-D.

w_{con_s} is also important for content preservation. Sometimes certain areas lose their saliency after style transformation. A larger w_{con_s} enforces these areas to preserve more original content to result in a similar saliency result with input. w_{sali} controls the accuracy of the saliency map. While saliency map is only used as a guidance in SDP-GAN, there is no need to train a highly accurate saliency map. Thus, a relatively lower w_{sali} is allowed. Empirically, we set $w_{con\_s} = 1.5$ and $w_{sali} = 1$ in our implementation. The proposed method aims to solve:

$$G^*, D^*, S^* = \arg\min_{G,S} \max_{D} L(G, D, S), \quad (2)$$

1) Adversarial Loss: As in classic style transformation GAN networks, the adversarial loss is used to constrain the results of G to look like target domain images. At the same time, D aims to distinguish whether a given image belongs to the synthesized or the real target dataset. However, simply applying the common adversarial loss is not sufficient for preserving clear edge information. Inspired by CartoonGAN [8], the proposed method also confuses D with a blur dataset to push generator to produce images with sharp edges. Specifically, from the target dataset $\{y_j\}_{j=1,\ldots,M} \in Y$, we generate the same amount of blur images $\{c_j\}_{j=1,\ldots,M} \in C$ by removing clear edges in Y. In more detail, for each y_j, we apply the following three steps: (1) detecting edge pixels using a standard Canny edge detector, (2) dilating the edge regions, and (3) applying a Gaussian smoothing in the dilated edge regions. That is to say, the discriminator tries to correctly classify an image into three categories: the generated images $G(x)$, the artistic images y, and the blurred artistic images c, as formulated in Eq. 3.

$$\begin{aligned} L_{\mathrm{GAN}}(G, D) = {} & E_{y\sim p_{\mathrm{data}}(y)}[\log D(y)] \\ & + E_{c\sim p_{\mathrm{data}}(c)}[\log(1 - D(c))] \\ & + E_{x\sim p_{\mathrm{data}}(x)}[\log(1 - D(G(x))], \end{aligned} \quad (3)$$

2) Image Content Loss: Although the adversarial loss is effective for generating stylized images, it does not guarantee that the translated images preserve the content information. It is essential to include a content loss to ensure the output artistic images retain the semantic content. Similar to [4], we define the image content loss function that measures high-level semantic differences using features from the VGG network [51] pretrained on ImageNet [52]. Different from [4], we propose a mask-based content loss, which uses the saliency

mask $S'(x)_m$ to restrict the content loss in salient regions, as formulated in Eq. 4.

$$S'(x)_m = \begin{cases} 1 & S'(x) < 1 \\ S'(x) & 1 < S'(x) < 2, \end{cases} \tag{4}$$

We first define $S(x)$ as the output of the saliency network, and it is normalized to $S'(x)$ with pixel values ranging from 0∼2. The saliency mask $S'(x)_m$ is designed to have larger values in salient regions with $S'(x) > 1$. The saliency-guide image content loss makes the proposed method flexible for translating images with and without salient objects. For images with obvious salient objects in foreground and background, different values in $S'(x)_m$ offer different penalty for L_{con}. For images without salient regions, their saliency maps tend to be uniform. $S'(x)_m$ is assigned with same values, which has same penalty for x and $G(x)$. To match with the VGG feature dimension, we downsample the saliency mask $S'(x)_m$ and duplicate its channels to get the adjusted saliency map denoted as $S''(x)_m$. The complete image content loss is defined as:

$$L_{con}(G, D, S) = E_{x\sim p_{\text{data}}(x)}[||S''(x)_m \cdot VGG_l(G(x)) - S''(x)_m \cdot VGG_l(x)||_1], \tag{5}$$

where l refers to a specific layer in the VGG network, and we apply the 'conv4_4' layer in our implementation. The loss is amplified in salient regions so that G is capable of keeping more content details in these regions.

3) Saliency Content Loss: Another content loss is proposed to further ensure the maintenance of content information. It calculates the difference between the saliency results of the input image $S(x)$ and that of the generated image $S(G(x))$:

$$L_{con\_s}(G, D, S) = E_{x\sim p_{\text{data}}(x)}[||S(x) - S(\mathrm{G}(x))||_1], \tag{6}$$

The saliency content loss is designed based on the concept that the input and output should have very similar saliency maps. They all go through the saliency network and own similar performance. Therefore, it is easy to find the tiny difference between them, which reflects the information loss caused by the generator. Most of the time it offers similar functionality as the image content loss L_{con}, and does not offer extra benefits. However, it sometimes complements L_{con} and has significant influences on certain cases that have different $S(x)$ and $S(G(x))$. Examples will be presented in Section V-D.

4) Saliency Loss: The saliency loss seeks to drive S to detect correct salient regions. It is defined to minimize the difference between the saliency output and the saliency reference:

$$L_{sali}(S) = E_{x\sim p_{\text{data}}(x), z\sim p_{\text{data}}(z)}[||S(x) - z||_1], \tag{7}$$

The saliency reference is obtained by computing the average of two saliency detection methods: robust background saliency detection (RBD) [53] and minimum barrier salient object detection (MBD) [54]. Computing average results of two methods can combine their advantages and avoid one of them being distorted if encountering extreme examples, which is simple yet effective.

(a) (b)

Fig. 3. Results of initialization phase. (a) An input and its saliency reference. (b) Generated result and rough saliency result after 10 epochs pre-training.

IV. Implementation

A. Data Collection

Many existing style transfer datasets contain real photos and target images with specific painting styles [7]. In this work, we also collect two kind of datasets, one containing diverse realistic photos and the other containing a large variety of images with specific artistic styles. All the training images are scaled to 256 × 256.

1) Realistic Photos: The realistic training dataset includes 8,936 images, among which 6,153 images originate from CycleGAN [7] and 2,783 images are our own collection. The CycleGAN dataset includes many landscape pictures with relatively uniform content, so we gathered images from movies or from the Internet that own clear salient objects.

2) Stylized Images: The proposed method learns to mimic the style of an entire collection of artwork. For example, we can learn to generate photos in the style of Van Gogh or Miyazaki Hayao. In our experiments, the datasets of Van Gogh style (401 images), Ukiyo-e style (1,433 images) and Monet style (1,074 images) originate from CycleGAN [7]. The dataset of Miyazaki Hayao style contains 3,617 images which are derived from the film 'Spirited Away'. 4,992 and 3,572 images from several short cartoon videos are applied for training the Makoto Shinkai and Mamoru Hosoda styles. To obtain cartoon images, we first extracted frames from cartoon films. Duplicated frames are then discarded using SSIM and PSNR similarity [55].

B. Training Details

We implement our SDP-GAN in PyTorch and all the experiments are performed on an NVIDIA Titan Xp GPU. Similar to CartoonGAN [8], an initialization phase is introduced to improve the convergence and avoid training from being trapped in a suboptimal local minimum. We initialize the saliency-guide generator that reconstructs the content of inputs and ignores the style translation. For this purpose, we pre-train both the generator G (to reconstruct the content of input images) and the saliency network S (to generate saliency maps) using merely L_{con} and L_{sali}. An example is presented in Fig. 3, where Fig. 3 (a) includes the original input and its corresponding saliency reference, and Fig. 3 (b) shows the generated image and the generated saliency result after initialization. 10 epochs are trained in the initialization

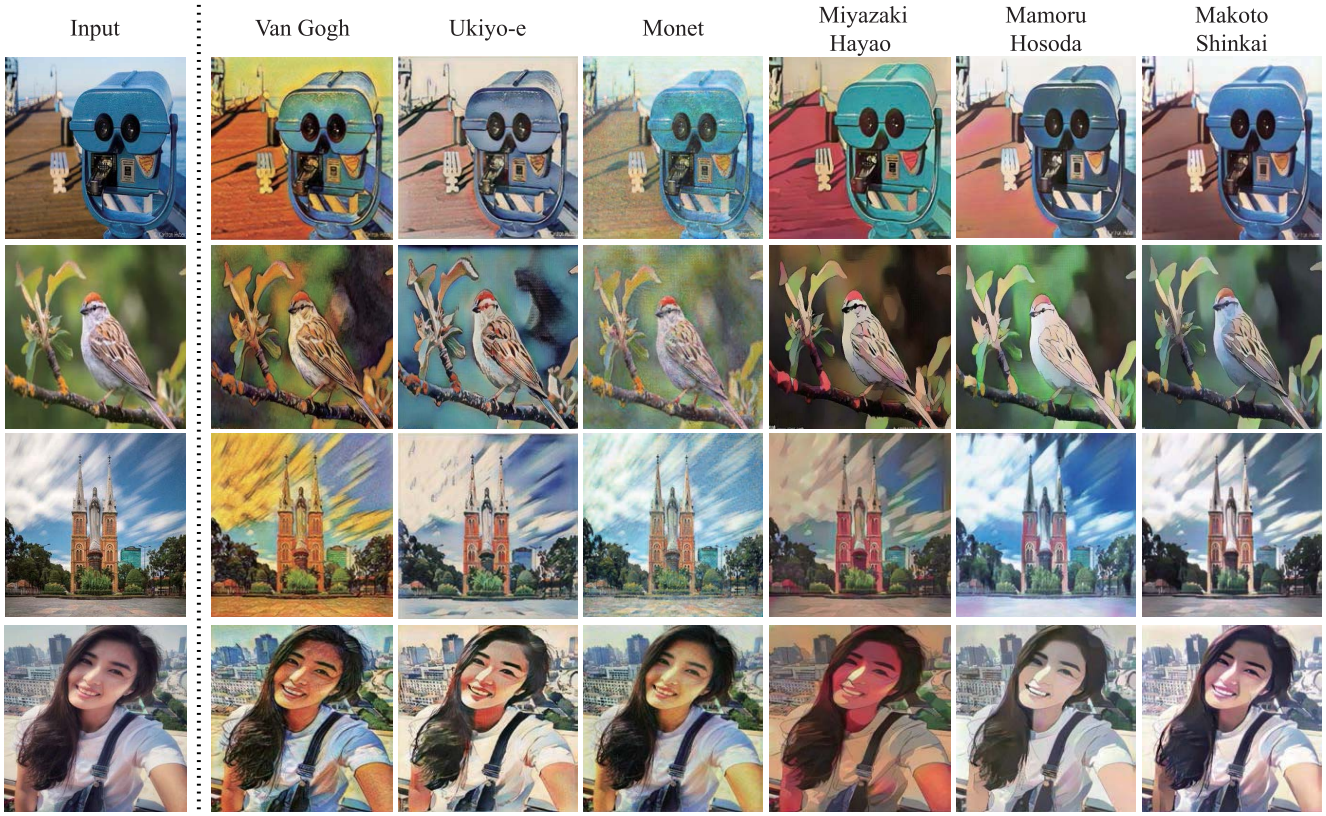


Fig. 4. Examples generated by the proposed method for different artistic styles. Each row shows an input and corresponding six style transfer results. Different styles can be effectively learned by SDP-GAN.

phase, which already gives reasonable reconstruction results and rough saliency maps.

V. Experiments

In this section, we first show some images of different artistic styles generated by SDP-GAN in Fig. 4, among which the proposed method is able to produce high-quality stylization results. Then, we compare our approach against several previous style transfer works, including a classic neural style transfer (NST) method [13], a structure-preserving neural style transfer (SP-NST) method, StarGAN [33], CartoonGAN [8] and a popular GAN-based stylization method CycleGAN [7]. We first compare SDP-GAN with these methods qualitatively. Quantitatively comparison and a user study evaluation are further conducted. Next, we perform the ablation studies to illustrate the importance of different components.

A. Qualitative Comparisons With SOTA Methods

Qualitative comparisons of SDP-GAN and aforementioned methods for two styles are presented in Fig. 5. NST takes one style image and one content image as inputs, and transfers content image into the target style. SP-NST is trained with COCO dataset [56] and one style image. We manually choose a style image which has close content to the input and similar style to our target dataset for NST and SP-NST. For StarGAN, it is trained for 200,000 iterations using the collected datasets. We compare two versions of CycleGAN, i.e., without and with the identity loss $L_{identity}$. The incorporation of this loss tends to produce stylized images with better color and content preservation. Our collected datasets are utilized to train both CartoonGAN, CycleGAN and SDP-GAN with 200 epochs.

Results in Fig. 5 clearly demonstrate that NST, SP-NST, StarGAN, CartoonGAN and CycleGAN cannot work well at salient regions. For NST method, only using a single style image cannot fully capture the style, especially for areas whose content does not appear in the style image (Fig. 5 (b)). SP-NST is also limited by a single style image (Ukiyo-e style in Fig. 5 (c)) and not robust for structure preservation (Van Gogh style in Fig. 5 (c)). StarGAN is designed for facial attribute transfer and facial expression synthesis, which is not sensitive to artistic style transfer tasks. Therefore, it generates results with insufficient stylization (Fig. 5 (d)). CycleGAN can capture the artistic style sometimes. However, it cannot preserve the saliency content well (Fig. 5 (f)). The identity loss is useful to avoid such problem, but the stylization results are still far from satisfactory. The edge information loss in structured regions and undesired artifacts are obvious in Fig. 5 (g). CartoonGAN obtains results with attractive cartoon style but ignores saliency features and causes same artifacts similar to CycleGAN (Fig. 5 (e)). In comparison, by reducing unnecessary artifacts and preserving desired edges in salient regions, the proposed SDP-GAN produces the highest quality results.

Our method has similar properties with CycleGAN and CartoonGAN. However, SDP-GAN takes less training time

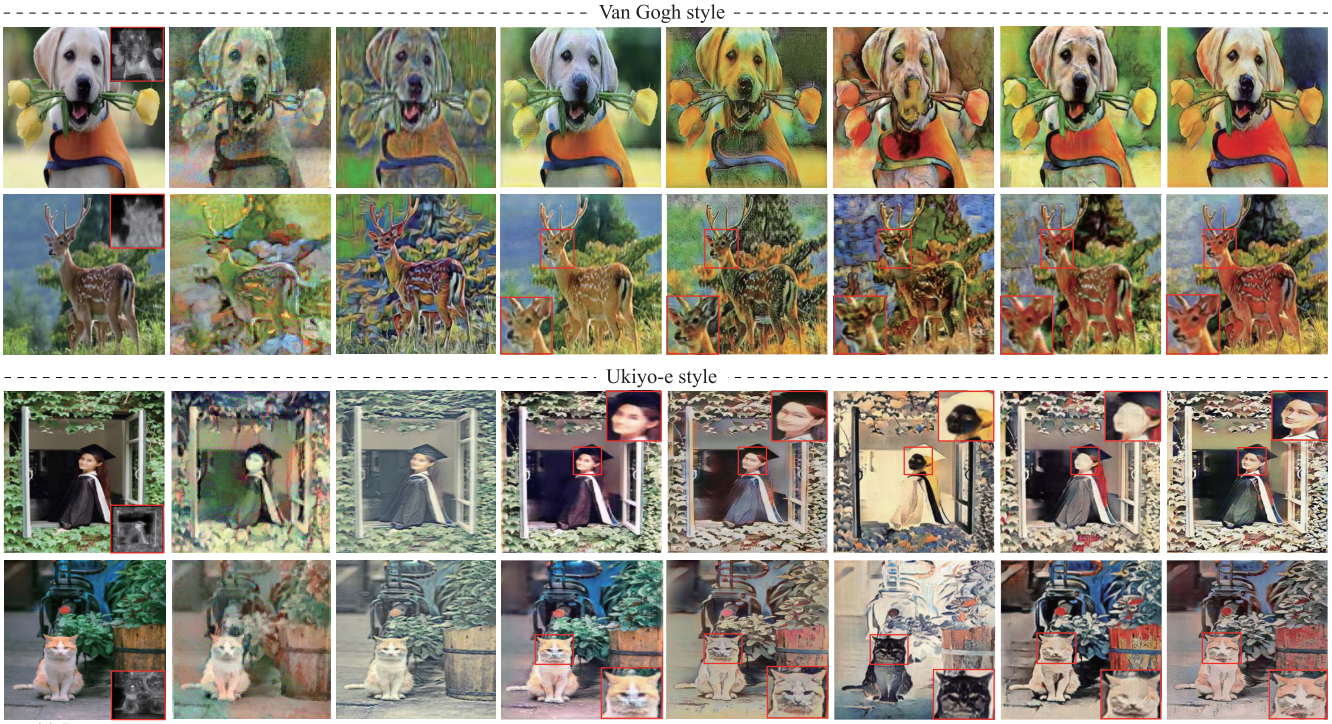


Fig. 5. Comparisons of SDP-GAN with NST [13], SP-NST [27], StarGAN [33], CartoonGAN [8] and CycleGAN [7] for two different styles. Note that StarGAN is mainly designed for facial attribute transfer and facial expression synthesis, which is not sensitive to artistic style transfer tasks. Although it keeps many content information from inputs, it has unqualified performance for style transformation. Our method is more effective for handling salient regions and producing high-quality style transfer results.

than CycleGAN and similar training time as CartoonGAN. For each epoch, CycleGAN and CartoonGAN take about 2350s and 1500s, respectively, whereas our method takes 1660s approximately. CycleGAN spends more time on its bidirectional training. Our saliency network S has about 11% number of parameters compared to generator G, and our method only consumes extra 160s compared with CartoonGAN. Adding a simple network with a little more training time is advisable in exchange for better results. Our method can learn many artistic styles better and handle salient regions more effectively.

B. Quantitative Comparisons With SOTA Methods

We choose the Inception score (IS) [15] and the Fréchet Inception distance (FID) [16] for quantitative evaluation. IS computes the KL divergence between the conditional class distribution and the marginal class distribution. Higher Inception score indicates better image quality. We include the Inception score evaluation because it is widely used [35], [57] and thus makes it possible to compare our results with previous works. However, it is important to understand that the IS has some limitations: the statistics of real world samples are not used and compared to the statistics of synthetic samples [16]. FID is a more principled and comprehensive metric, and has been shown to be more consistent with human evaluation in assessing the realism and variation of the generated samples. FID calculates the Wasserstein-2 distance between the generated images and the real images. To this end, the generated samples are first embedded into a feature space of an Inception-v3 network. Then, taking the embedding layer as a continuous multi-variate Gaussian, the mean and covariance are estimated for both the generated data and the real data. The Fréchet distance between these two Gaussians is then used to quantify the quality of the samples:

$$FID(x, g) = ||\mu_x - \mu_g|| + T_r(\sum_x + \sum_g - 2(\sum_x \sum_g)^{1/2}), \tag{8}$$

where $(\mu_x, \sum_x)$ and $(\mu_g, \sum_g)$ are the mean and covariance of sample embedding for real data distribution and generative model distribution. Lower FID values mean closer distances between synthetic and real data distributions.

Table I lists the IS and FID scores of StarGAN [33], CartoonGAN [8], CycleGAN (with $L_{identity}$) [7] and the proposed SDP-GAN for three artistic styles. We do not quantitatively compare the results of NST, SP-NST and CycleGAN without $L_{identity}$ because their results are obviously inferior to results of aforementioned methods. Compared to other three methods, our SDP-GAN achieves better IS. As for FID, SDP-GAN has great improvement compared with StarGAN and CartoonGAN, and has approximate 5 point improvement compared with CycleGAN.

TABLE I

QUANTITATIVE COMPARISONS OF SDP-GAN WITH OTHER REPRESENTATIVE GAN METHODS IN TERMS OF IS AND FID SCORES

Styles	IS and FID	StarGAN	CartoonGAN	CycleGAN+$L_{identity}$	SDP-GAN
Miyazaki Hayao	IS	5.48±0.52	6.09±0.83	4.37±0.58	**6.38±0.98**
	FID	169.42	159.69	136.82	**133.76**
Van Gogh	IS	4.28±0.53	4.77±0.48	4.58±0.74	**4.86±0.59**
	FID	162.89	135.93	106.94	**101.59**
Ukiyo-e	IS	5.33±0.52	**6.10±0.72**	5.75±0.64	6.07±0.77
	FID	152.46	123.42	107.25	**101.92**
MEAN	IS	5.03±0.52	5.65±0.68	4.9±0.65	**5.77±0.78**
	FID	161.59	139.68	117	**112.42**

TABLE II

USER STUDY RESULTS. THE NUMBERS ARE THE PERFORMANCE OF VOTES OBTAINED BY EACH METHOD

Method	StarGAN	CartoonGAN	CycleGAN+$L_{identity}$	SDP-GAN
Percentage	7%	15.67%	26.67%	**50.66%**

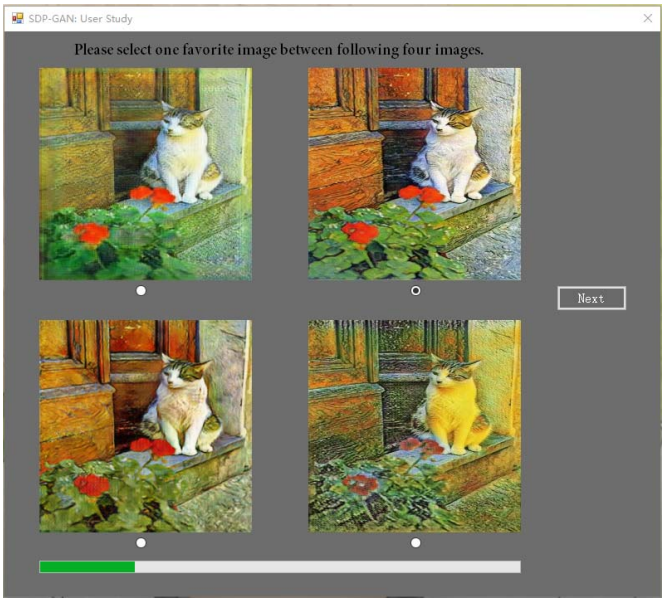


Fig. 6. The screen shot of user study. Four stylized results are shown to the viewers at the same time.

C. User Study

We further conduct a user study to evaluate our method subjectively. We invited 20 users (10 males and 10 females) to evaluate the visual quality of different style transfer methods. In the evaluation, 30 groups of results are randomly selected from the test dataset and every group involves the results of StarGAN, CartoonGAN, CycleGAN (with $L_{identity}$) and the proposed SDP-GAN. Each user is required to discriminate 15 groups of images that are randomly selected from 30 groups of results. The 15 groups of images are shown to each user in a random order and the results in each group are also arranged randomly. During the test, four stylized results are shown to the viewers at the same time on a computer screen with resolution 256×256, as shown in Fig. 6. For each group, the viewers were asked to answer 'Which is the most favorite image between following style transfer results?'. We calculate the number of best scores of 300 groups (20 viewers$\times$15 groups of results) and display the results in Table II. It is obvious that our results are favored by a majority of viewers, which indicates our method achieves better style transformation by preserving the saliency content information.

D. Ablation Studies

1) Influence of Parameters in Loss Function: We first conduct some experiments to illustrate why we set $w_{con} = 0.25$ at the initial stage. Fig. 7 shows the corresponding results when we select different w_{con}. A larger w_{con} produces images that are similar to realistic images because it leads to more content information from the input photos (Fig. 7 (b)-Fig. 7 (d)). On the contrary, a smaller w_{con} learns the stylization excessively so that the semantic content information cannot be preserved well. Moreover, it generates results with unnatural color (Fig. 7 (f)). We set w_{con} to be 0.25 at the initial stage to keep a balance between style transformation and content preservation. It correctly maintains semantic content with enjoyable style transformation (Fig. 7 (e)).

In our implementation, w_{con} is gradually decreased to achieve better style transfer results while the content information has been properly reconstructed at the initial stage. If w_{con} is fixed, the final results would bring more content information from inputs and are not stylized enough to be comparable with target images. We set $w_{con} = 0.25$ at initial stage to preserve content information and introduce a decay factor of w_{con} to learn better style transfer mapping. The variation trend of w_{con} can be described as:

$$w_{con} = w_{con} \times 0.96^{\lfloor N_e/10 \rfloor}, \tag{9}$$

where N_e is the number of epochs in the training process. Fig. 8 shows the comparison results. We found that the results with fixed w_{con} (Fig. 8 (b)) is easier to lose style information than the results with gradually degressive w_{con} (Fig. 8 (c)).

2) Ablation Study of Different Loss Terms: We perform the ablation study on the variants of loss functions to understand how these main modules contribute to final results. Fig. 9 and Fig. 10 display the ablation results of our loss functions, in which all the results are trained with data of Hayao style.

Fig. 7. The effect of different w_{con}. Large w_{con} values generate images that are not stylized enough, while a small w_{con} learns the stylization excessively.

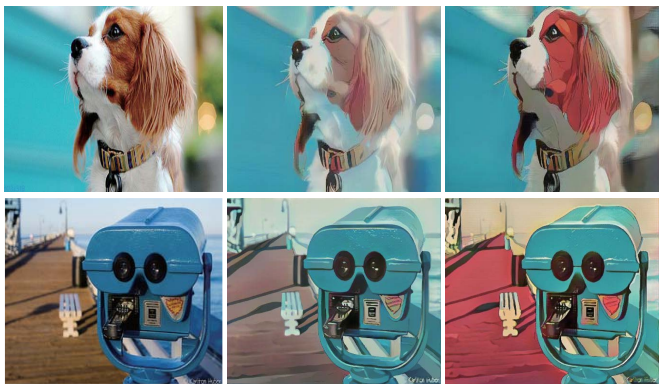


Fig. 8. Results to show the influence of decay factor. (a) Input images. (b) Results with fixed w_{con}. (c) Results with gradually degressive w_{con}. It's obvious that the introduce of decay factor is beneficial for stylization.

These results show that each component plays an important role in our objective function.

As shown in Fig. 9, removing saliency loss substantially degrades the results, so as removing the saliency mask in the image content loss. Fig. 9 (c) shows the results without L_{sali}, which indicates the saliency network and its feature sharing with the generator are removed, and the saliency mask $S''(x)_m$ is simply substituted by the saliency reference. The results in Fig. 9 (c) are apparently inferior to the final results. Fig. 9 (d) displays the results without using the saliency mask $S''(x)_m$, which means the content loss only computes the difference between inputs and outputs expressed as: $E_{x\sim p_{\text{data}}(x)}[||VGG_l(G(x)) - VGG_l(x)||_1]$. The results are also not as good as our final results. We conclude that both L_{sali} and $S''(x)_m$ terms are critical.

The importance of L_{con_s} is demonstrated in Figure 10. Fig. 10 (a) shows an example of an original input and its saliency result. If L_{con_s} is not applied, the translated image loses a lot of details as shown in Fig. 10 (b). For example, the details of the stamen regions are apparently lost, and the boundaries between the stamens and petals are no longer distinguishable compared to the input image. We observe that its saliency map in Fig. 10 (b) is very different from input's saliency map in Fig. 10 (a). By introducing L_{con_s}, we enforce the generator to keep more original content to yield a similar saliency map. In this way, the final result has more details preserved in salient regions as shown in Fig. 10 (c).

3) Ablation Study of Feature Concatenation: To demonstrate the effectiveness of feature concatenation of the proposed method, we conduct experiments that without and with feature concatenation between generator and saliency network, respectively. The results are shown in Fig. 11, in which the result without concatenation (Fig. 11 (b)) is obviously inferior to the result with concatenation (Fig. 11 (c)). The connection helps to maintain the uniformity of salient regions, and therefore obtain better style transfer results.

E. Discussion

The proposed method can translate images with and without salient objects. For images without salient objects, their saliency maps tend to be uniform, which play the same role in the entire image. For images with obvious salient objects, our method can preserve details in salient regions. Notably, the salient objects appear in the foreground or background are determined by the saliency detection methods. Our method concentrates on preserving details for salient regions once the saliency map is given, but cannot control the saliency detection results. Moreover, the proposed SDP-GAN is not suitable

Fig. 9. Ablation experiments to show the importance of L_{sali} and $S''(x)_m$ in the loss functions. (a) Input photos. (b) The corresponding saliency maps. (c) Results of removing the saliency network and L_{sali}. (d) Results of removing $S''(x)_m$ in image content loss. (e) Our results.

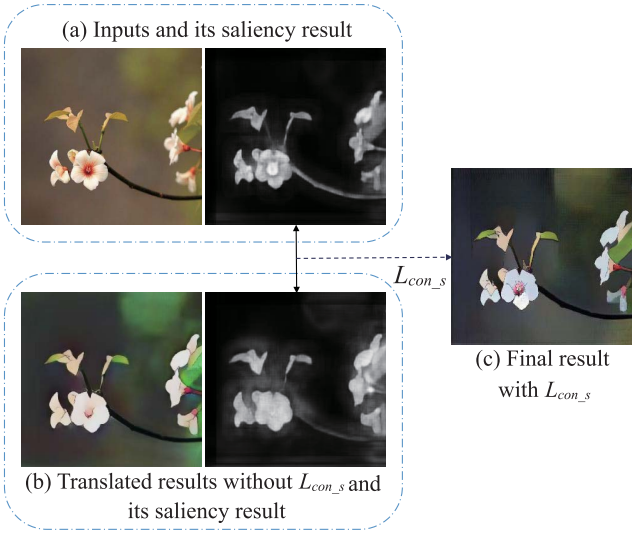


Fig. 10. Results to show the importance of L_{con_s}. (a) Input and saliency result. (b) Translated image without L_{con_s} and corresponding saliency result. (c) Final result with L_{con_s}.

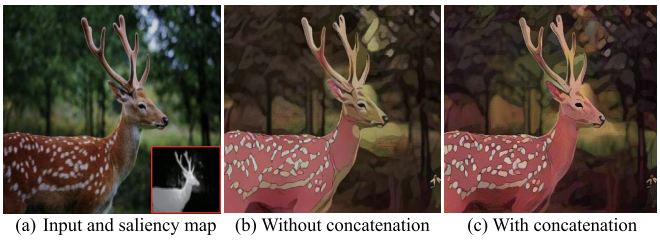


Fig. 11. Results to demonstrate the effectiveness of feature concatenation. (a) Input and saliency result. (b) Generated image without feature concatenation. (c) Generated image with feature concatenation.

for the tasks that the saliency objects are changed, such as dog2cat translation. Adding a content transfer network and preserving details for required salient objects are also valuable. We consider them as future works.

VI. Conclusion

We propose SDP-GAN, a GAN-based method to preserve saliency information while achieving appropriate artistic stylization between unpaired datasets. The architecture contains a generator network, a discriminator network and a saliency network. The saliency network generates saliency maps to restrict the image content loss on one side and provides saliency features to generator on the other side. In addition to the GAN loss, three losses suitable for the task are introduced to improve the performance. Image content loss is designed to ensure the output artistic images retain their semantic content. Saliency content loss plays a supplementary role in image content loss and works with special cases. Saliency loss drives saliency network to detect correct salient regions. SDP-GAN can generate images of higher quality compared to existing methods. We believe the proposed framework can also be generalized to other style transfer tasks.

References

[1] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2650–2658.

[2] X.-C. Liu, M.-M. Cheng, Y.-K. Lai, and P. L. Rosin, "Depth-aware neural style transfer," in *Proc. Symp. Non-Photorealistic Animation Rendering (NPAR)*, 2017, pp. 1–10.

[3] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman, "Controlling perceptual factors in neural style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3730–3738.

[4] 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.

[5] 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. (CVPR)*, Jul. 2017, pp. 1125–1134.

[6] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2337–2346.

[7] 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. (ICCV)*, Oct. 2017, pp. 2223–2232.

[8] Y. Chen, Y.-K. Lai, and Y.-J. Liu, "CartoonGAN: Generative adversarial networks for photo cartoonization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9465–9474.

[9] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1155–1162.

[10] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 20–33, Jan. 2018.

[11] F. Murabito, C. Spampinato, S. Palazzo, D. Giordano, K. Pogorelov, and M. Riegler, "Top-down saliency detection driven by visual classification," *Comput. Vis. Image Understand.*, vol. 172, pp. 67–76, Jul. 2018.

[12] Y. Liu *et al.*, "Image neural style transfer with preserving the salient regions," *IEEE Access*, vol. 7, pp. 40027–40037, 2019.

[13] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

[14] X. Chen, C. Xu, X. Yang, L. Song, and D. Tao, "Gated-GAN: Adversarial gated networks for multi-collection style transfer," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 546–560, Feb. 2019.

[15] T. Salimans, H. Zhang, A. Radford, and D. Metaxas, "Improving GANs using optimal transport," 2018, *arXiv:1803.05573*. [Online]. Available: http://arxiv.org/abs/1803.05573

[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. NIPS*, 2017, pp. 6626–6637.

[17] B. Gooch and A. Gooch, *Non-Photorealistic Rendering*. Boca Raton, FL, USA: CRC Press, 2001.

[18] P. Rosin and J. Collomosse, *Image and Video-Based Artistic Stylisation*, vol. 42. Springer, 2012.

[19] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," 2015, *arXiv:1508.06576*. [Online]. Available: http://arxiv.org/abs/1508.06576

[20] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Proc. NIPS*, 2015, pp. 262–270.

[21] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2230–2236.

[22] E. Risser, P. Wilmot, and C. Barnes, "Stable and controllable neural texture synthesis and style transfer using histogram losses," 2017, *arXiv:1701.08893*. [Online]. Available: http://arxiv.org/abs/1701.08893

[23] S. Li, X. Xu, L. Nie, and T.-S. Chua, "Laplacian-steered neural style transfer," in *Proc. ACM Multimedia Conf. (MM)*, 2017, pp. 1716–1724.

[24] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," 2016, *arXiv:1610.07629*. [Online]. Available: http://arxiv.org/abs/1610.07629

[25] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Diversified texture synthesis with feed-forward networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3920–3928.

[26] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "StyleBank: An explicit representation for neural image style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1897–1906.

[27] M.-M. Cheng, X.-C. Liu, J. Wang, S.-P. Lu, Y.-K. Lai, and P. L. Rosin, "Structure-preserving neural style transfer," *IEEE Trans. Image Process.*, vol. 29, pp. 909–920, 2020.

[28] X. Liu, Z. Liu, X. Zhou, and M. Chen, "Saliency-guided image style transfer," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2019, pp. 66–71.

[29] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1501–1510.

[30] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," in *Proc. NIPS*, 2017, pp. 386–396.

[31] L. Sheng, Z. Lin, J. Shao, and X. Wang, "Avatar-net: Multi-scale zero-shot style transfer by feature decoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8242–8250.

[32] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 702–716.

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

[34] A. Romero, P. Arbelaez, L. Van Gool, and R. Timofte, "SMIT: Stochastic multi-label Image-to-Image translation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–10.

[35] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," 2018, *arXiv:1805.08318*. [Online]. Available: http://arxiv.org/abs/1805.08318

[36] D. Y. Park and K. H. Lee, "Arbitrary style transfer with style-attentional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5880–5888.

[37] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[38] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2106–2113.

[39] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 438–445.

[40] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2798–2805.

[41] M. Kümmerer, L. Theis, and M. Bethge, "Deep gaze I: Boosting saliency prediction with feature maps trained on imagenet," in *Proc. ICLR*, 2014, pp. 1–12.

[42] M. Kümmerer, T. S. Wallis, L. A. Gatys, and M. Bethge, "Understanding low-and high-level contributions to fixation prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 4789–4798.

[43] J. Pan *et al.*, "SalGAN: Visual saliency prediction with generative adversarial networks," 2017, *arXiv:1701.01081*. [Online]. Available: http://arxiv.org/abs/1701.01081

[44] H. Pan and H. Jiang, "Supervised adversarial networks for image saliency detection," 2017, *arXiv:1704.07242*. [Online]. Available: http://arxiv.org/abs/1704.07242

[45] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Task specific visual saliency prediction with memory augmented conditional generative adversarial networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1539–1548.

[46] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, 2015, pp. 2048–2057.

[47] C.-C. Tsai, W. Li, K.-J. Hsu, X. Qian, and Y.-Y. Lin, "Image co-saliency detection and co-segmentation via progressive joint optimization," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 56–71, Jan. 2019.

[48] Y.-L. Chen and C.-T. Hsu, "Towards deep style transfer: A content-aware perspective," in *Proc. BMVC*, 2016.

[49] J. Zhang, T. Zhang, Y. Daf, M. Harandi, and R. Hartley, "Deep unsupervised saliency detection: A multiple noisy labeling perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9029–9038.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

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

[52] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[53] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2814–2821.

[54] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80 FPS," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1404–1412.

[55] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2366–2369.

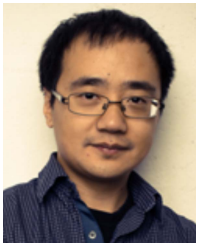
[56] T. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, 2014, pp. 740–755.

[57] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie, "Stacked generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5077–5086.

Ru Li (Student Member, IEEE) received the B.E. degree in electronic information engineering from the China University of Petroleum, Qingdao, China, in 2016. She is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China. From July 2019 to October 2019, she visited the University of Oxford, where she worked on image enhancement. Her research interests include image processing and computer vision.

Chi-hao Wu received the B.Sc. and M.Sc. degrees in computer science from National Taiwan University, Taipei, Taiwan, in 2004 and 2006, respectively, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 2017. Prior to his Ph.D., he has over six years of industrial experience as a Software Engineer in Foxconn, MediaTek, and IMEC-Taiwan, Hsinchu, Taiwan, from 2007 to 2013. In 2018, he joined Megvii, as a Senior Research Scientist. His research interests include computer vision, machine learning, and multimedia signal processing.

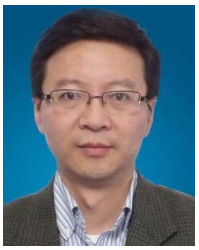
Shuaicheng Liu (Member, IEEE) received the B.E. degree from Sichuan University, Chengdu, China, in 2008, and the M.Sc. and Ph.D. degrees from the National University of Singapore, Singapore, in 2010 and 2014, respectively. In 2014, he joined the University of Electronic Science and Technology of China, Chengdu, where he is currently an Associate Professor with the School of Information and Communication Engineering, Institute of Image Processing. In 2018, he joined Megvii, and is currently the Research Lead at Megvii Research Chengdu. His research interests include computer vision and computer graphics.



Jue Wang (Senior Member, IEEE) received the B.E. and M.Sc. degrees from the Department of Automation, Tsinghua University, Beijing, China, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering from the University of Washington, Seattle, WA, USA, in 2007. He is currently the Senior Director of Megvii. Before that, he has been a Principle Research Scientist at Adobe Research for nine years. His research interests include image and video processing and computational photography. He is a Senior Member of ACM. He received the Microsoft Research Fellowship and the Yang Research Award from the University of Washington in 2006.

Guangfu Wang received the B.E. and M.Sc. degrees from the Department of Automation, University of Electronic Science and Technology of China, Chengdu, China, in 2013 and 2016, respectively. In 2016, he worked at Alibaba Group as an Algorithm Engineer. He is currently a Researcher of Megvii. His research interests include computer vision and deep learning.

Guanghui Liu (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in electronic engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2002 and 2005, respectively. In 2005, he joined Samsung Electronics, Seoul, South Korea, as a Senior Engineer. In 2009, he became an Associate Professor with the School of Electronics Engineering, UESTC, where he has been a Full Professor since 2014. He is with the School of Information and Communication Engineering, UESTC. His general research interests include multimedia, remote sensing, and wireless communication. In these areas, he has authored over ten papers in refereed journals or conferences, and holds more than 60 patents (six U.S. granted patents). He served as the Publication Chair for the IEEE ISPACS-2010 and the IEEE VCIP-2016. In 2015, he received the Natural Science Award and the Science and Technology Progress Award from the Ministry of Education of China.



Bing Zeng (Fellow, IEEE) received the B.E. and M.Sc. degrees in electronic engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 1983 and 1986, respectively, and the Ph.D. degree in electrical engineering from the Tampere University of Technology, Tampere, Finland, in 1991. He worked as a Post-Doctoral Fellow with the University of Toronto from September 1991 to July 1992, and as a Researcher with Concordia University from August 1992 to January 1993. Then, he joined The Hong Kong University of Science and Technology (HKUST). After 20 years of service, he returned to UESTC in Summer 2013, through Chinas 1000-Talent-Scheme. At UESTC, he leads the Institute of Image Processing to work on image and video processing, 3D and multiview video technology, and visual big data. During his tenure with HKUST and UESTC, he has supervised more than 30 master's and Ph.D. students, received over 20 research grants, filed eight international patents, and published more than 250 articles. He was a recipient of the 2nd Class Natural Science Award (the first recipient) from the Ministry of Education of China in 2014, and was elected as a Fellow of the IEEE in 2016, for contributions to image and video coding. He was the General Co-Chair of the IEEE VCIP-2016, Chengdu, in November 2016. He is currently on the Editorial Board of *Journal of Visual Communication and Image Representation* and serves as the General Co-Chair for PCM-2017. He served as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT) for eight years, and received the Best Associate Editor Award in 2011.