

# Learning to Incorporate Texture Saliency Adaptive Attention to Image Cartoonization

Xiang Gao\*, Yuqi Zhang\*, and Yingjie Tian

**Abstract**—Image cartoonization is recently dominated by generative adversarial networks (GANs) from the perspective of unsupervised image-to-image translation, in which an inherent challenge is to precisely capture and sufficiently transfer characteristic cartoon styles (e.g., clear edges, smooth color shading, abstract fine structures, etc.). Existing advanced models try to enhance cartoonization effect by learning to promote edges adversarially, introducing style transfer loss, or learning to align style from multiple representation space. This paper demonstrates that more distinct and vivid cartoonization effect could be easily achieved with only basic adversarial loss. Observing that cartoon style is more evident in cartoon-texture-salient local image regions, we build a region-level adversarial learning branch in parallel with the normal image-level one, which constrains adversarial learning on cartoon-texture-salient local patches for better perceiving and transferring cartoon texture features. To this end, a novel cartoon-texture-saliency-sampler (CTSS) module is proposed to dynamically sample cartoon-texture-salient patches from training data. With extensive experiments, we demonstrate that texture saliency adaptive attention in adversarial learning, as a missing ingredient of related methods in image cartoonization, is of significant importance in facilitating and enhancing image cartoon stylization, especially for high-resolution input pictures.

**Index Terms**—Image Cartoonization, Style Transfer, Generative Adversarial Networks, Image-to-Image Transformation.

## I. INTRODUCTION

IMAGE cartoonization aims at rendering natural images into cartoon styles, which is a challenging computer vision and computer graphics problem. Specially designed algorithms that automatically stylize pictures into cartoons can help relieve cartoon creation artists from laborious manual work, and also have practical values in digital entertainment, advertising, childhood education, image editing, etc.

Cartoons have unique visual features characterized by clear edges and smooth color shading in non-edge areas. The problem of reproducing cartoon-like effect on real photos is widely explored in early time from the perspective of image abstraction [1], [2], [3], [4]. These methods model cartoon styles with established image processing techniques, which could be summarized into three aspects: (i) image smoothing;

Xiang Gao is with School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China. (e-mail: gaoxiang181@mails.ucas.ac.cn)

Yuqi Zhang is with School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China. (e-mail: zhangyuqi201@mails.ucas.ac.cn)

Yingjie Tian is with Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences, Beijing 100190, China; and also with Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China. (e-mail: tyj@ucas.ac.cn)

\* Equal contribution



Fig. 1. Example evaluation results of our method in transforming real-world scenes into cartoon styles (**better to zoom in**).

(ii) color quantization; (iii) edges enhancement. Though successful in mimicking some important cartoon features, they lack data-driven learning ability to capture cartoon styles in more depth.

With the thriving of deep neural networks, recent image cartoonization methods resort to learning-based framework, typically generative adversarial network (GAN) [5], to automatically learn and transfer high-level cartoon styles from real cartoon images. It could be formulated as an unsupervised image-to-image translation problem where the objective is to learn a content-preserving image translation mapping  $\mathcal{X} \rightarrow \mathcal{Y}$  from a source domain  $\mathcal{X}$  of natural images to a target domain  $\mathcal{Y}$  of cartoon images. A general framework is to align style distribution of generated images to that of target-domain real cartoons through adversarial learning, and meanwhile, constrain perceptual content consistency between input photos and generated results to avoid content mismatch. However, it is difficult to produce results with sufficiently salient cartoon features, for which some advanced methods make further progresses to enhance cartoonization effect based on the general framework.

CartoonGAN [6] proposes an edge-promoting adversarial loss to highlight typical cartoon feature of edge clearness. This loss function enforces the discriminator to distinguish real cartoon images from not only the synthesized images but also the edge-smoothed cartoon images, such that the generator can be guided to produce clearer edges to fool the discriminator. However, it requires a separate preparation stage to collect an edge-smoothed cartoon image set before training, losing the elegance of end-to-end learning. Furthermore, the

edges and contours of the generated results are still not very distinct. AnimeGAN [7] introduces Gram loss [8], a classic texture-descriptor-based style loss widely used in style transfer literature, to GAN framework to enhance learning cartoon texture pattern. Nevertheless, its effect in strengthening cartoon texture transfer is still less noticeable. More recently, a white-box image cartoonization framework [9] is proposed. It decomposes images into multiple representations and learns to align styles in the manifold of each representation. This method proposes an adaptive coloring algorithm that generates image color segmentation maps to mimic sparse color blocks of cartoon images, which brings visually appealing celluloid cartoon style. However, the cartoon abstraction and vividness of the generated results are still less prominent, especially for high-resolution input images.

The above-mentioned models resort to different methods to make up for the limitation of the basic adversarial loss in fully transferring cartoon styles. However, we argue that the weak stylization problem is not from the capacity of adversarial loss itself, but due to non-global distribution of salient cartoon texture features. For example, the clear edges are usually distributed in local areas rather than the entire image, and the pixel proportion of clear edges is also very small. Therefore, the feature of edge clearness could be easily overwhelmed by more obvious global features such as color shading smoothness, when trained with basic adversarial loss over the scale of entire image. This enlightens us to attend to cartoon-texture-salient local regions for better perceiving and transferring cartoon texture features. Similar idea of paying attention to local distribution of image patches is also widely explored in many other computer vision fields, such as image classification [10], [11], [12], image restoration [13], [14], [15], object detection [16], [17], [18], etc.

In this paper, we propose a compact and efficient adversarial learning framework with an image-level discriminator examining global cartoon styles like smooth shading and the overall tone, as well as a patch-level discriminator focusing on learning local cartoon texture pattern, i.e., the unique distribution of high- and low-frequency pixels around clear edges. To enhance transfer of cartoon texture pattern, we adaptively constrain patch-level adversarial learning on cartoon-texture-salient local image regions, for which a novel cartoon-texture-saliency-sampler (CTSS) module is proposed to dynamically extract image patches with most salient cartoon texture pattern from each mini-batch of training images. By incorporating such texture saliency adaptive attention to adversarial learning, the typical cartoon texture features are more sufficiently transferred, yielding more abstract and vivid cartoonization effect. Example results of our model are presented in Fig. 1. Our method bypasses separate edge-smoothing data preparation stage, use of additional style losses, and complicated representation extraction processes, while producing more prominent cartoon effect, especially for large input pictures. The effectiveness of our model is fully demonstrated with extensive experiments evaluated on different cartoon datasets.

## II. RELATED WORK

### A. Style Transfer

Neural style transfer (NST) was first proposed as an online optimization-based algorithm that iteratively transfers image styles by minimizing Gram loss [8]. Afterwards, it was transformed into an offline generative model to cater for real-time applications by training a feedforward network [19], [20]. Later on, efforts had been made to extend fast NST from a single style to multiple styles [21], [22], or even arbitrary styles [23], [24]. Beside the Gram loss, various style losses were successively proposed, such as MMD loss [25], CNNMRF loss [26], contextual loss [27], and Relaxed EMD loss [28]. These loss functions suit for transferring low-level texture features from a single image. By contrast, cross-domain style transfer methods use adversarial loss to automatically learn high-level styles from a collection of stylistically similar images. By means of GANs, style transfer problem is enriched with more applications, such as font style transfer [29], [30], painting style transfer [31], [32], makeup style transfer [33], etc.

### B. Image-to-Image Translation

Image-to-image translation refers to transformation of images from a source domain to a target domain, which could be divided into supervised and unsupervised situations according to whether paired training data of two domains are available. For supervised problem, Pix2Pix [34] combines conditional GAN with image-level sparse  $L_1$  regularization, which generalizes well to many applications such as image super-resolution [35], image denoising [36], semantic image synthesis [37], etc. For the latter situation, CycleGAN [38] and UNIT [39] are typical models that realize unsupervised image translation based on cycle-consistency constraint and shared-latent-space assumption, respectively. Afterwards, methods like StarGAN [40] and AttGAN [41] extend translation from two domains to multiple domains, by combining conditional GAN with auxiliary domain classifier. Besides, the problem is also explored in the field of multimodal translation [42], [43] and few-shot learning [44].

## III. METHOD

Cartoon images have smooth shading and vivid colors. In addition to these global features that represent overall pixel distribution, the most salient feature of cartoon images is their unique texture pattern that represents local pixel distribution. Specifically, the high-frequency pixels concentrate on edges, while the low-frequency pixels are smoothly distributed beside edges. This distinct separation of high- and low-frequency pixels clearly differs from natural images where the high- and low-frequency elements are much interweaved. However, as Fig. 3 shows, such cartoon texture pattern manifests obviously only in partial image regions with clear edges, which means that the latent cartoon texture pattern is more visually perceptible from the view of edge-distinct local regions than from the view of entire image. Therefore, we append a patch-level learning branch that adaptively applies adversarial learning on edge-distinct local regions to enhance capturing cartoon texture pattern.

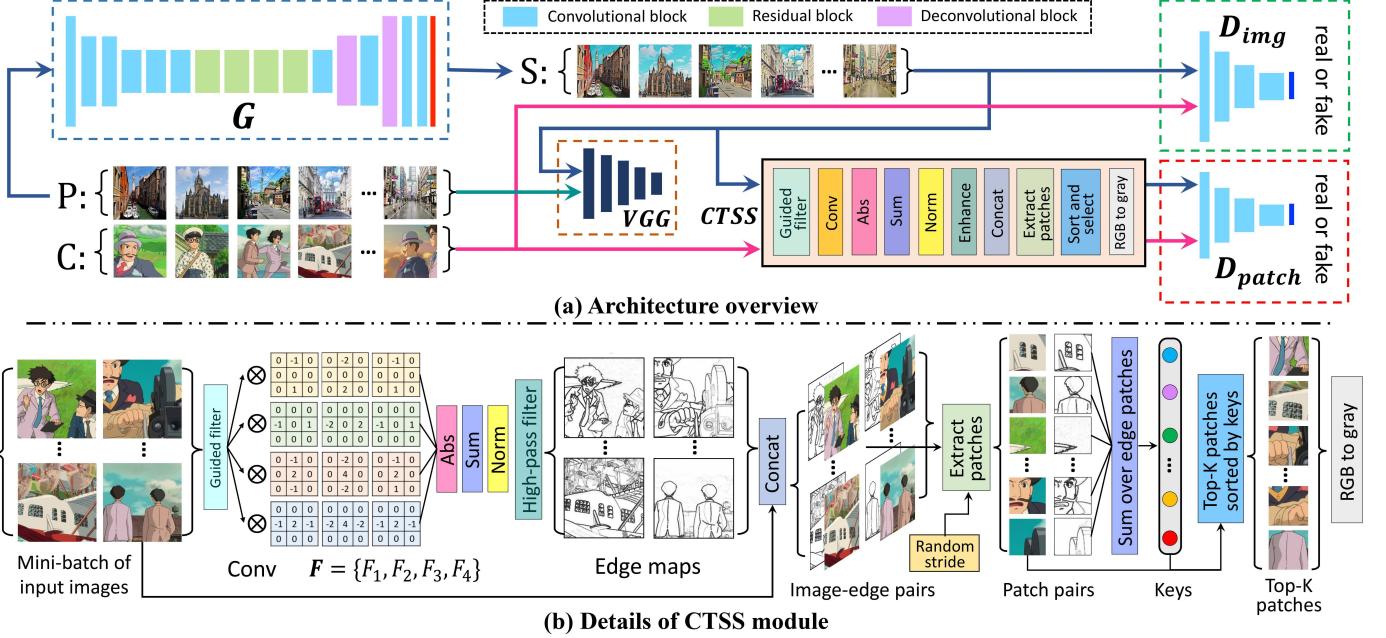


Fig. 2. The overall architecture of our model, as well as details of our proposed cartoon-texture-saliency-sampler (CTSS) module which adaptively extracts local image patches with most salient cartoon texture pattern from each mini-batch of input images.

### A. Global and local adversarial learning

Let  $\mathcal{P}$  denote the domain of real-world photos,  $\mathcal{C}$  be the domain of cartoon images,  $\mathcal{S}$  be the domain of synthesized results. The overall architecture of our adversarial learning framework is illustrated in Fig. 2 (a). In training phase, a mini-batch of natural photos  $P = \{p_i\}_{i=1}^N \in \mathcal{P}$ , and a mini-batch of cartoon images  $C = \{c_i\}_{i=1}^N \in \mathcal{C}$  are sampled at each iteration, where  $N$  is the batch size. A generator  $G$  translates  $P$  into a mini-batch of synthesized cartoon images  $S = G(P) = \{s_i\}_{i=1}^N \in \mathcal{S}$ , which are differentiated from real cartoon images  $C$  by an image-level discriminator  $D_{img}$ . This forms the image-level adversarial learning branch for learning global holistic cartoon styles.

To better seize cartoon texture pattern that is more perceptible at edge-distinct local regions, we append a patch-level adversarial learning branch as a supplement to the image-level one. In this branch, a cartoon-texture-saliency-sampler (CTSS) module is proposed to constrain adversarial learning only on cartoon-texture-distinct local regions. As shown in Fig. 2 (a), the CTSS module takes cartoon mini-batch  $C$  and the synthesized mini-batch  $S$  as input, and outputs top- $K$  edge-distinct local patches  $C_{patch} = \{c_p^i\}_{i=1}^K$  and  $S_{patch} = \{s_p^i\}_{i=1}^K$  from  $C$  and  $S$  respectively. A patch-level discriminator  $D_{patch}$  is built to distinguish  $S_{patch}$  from  $C_{patch}$ , forming the patch-level adversarial learning that reinforces transfer of cartoon texture pattern.

### B. Cartoon-texture-saliency-sampler

Since the unique cartoon texture pattern is more visually prominent at edge-distinct local regions, our CTSS module adaptively samples local image patches with most distinct edges from each input mini-batch of images, the implementation details are illustrated in Fig. 2 (b). Taking an input mini-

batch of cartoon images  $C = \{c_i\}_{i=1}^N$  as example, CTSS starts with a guided-filter [45] sub-module  $\mathcal{F}_{gf}$  for edge-preserving image smoothing, it uses each input image  $c_i$  itself as guide map, returns the smoothed image  $\tilde{c}_i$  with many noise elements removed:

$$\tilde{c}_i = \mathcal{F}_{gf}(c_i, c_i), i = 1, \dots, N. \quad (1)$$

Then, a convolutional layer is applied to extract coarse edge maps  $E = \{e_i\}_{i=1}^N$ , where  $e_i$  is the edge map of  $c_i$ . The convolutional layer has a constant kernel  $\mathbf{F}$  composed of four filters  $\{F_1, F_2, F_3, F_4\}$  as shown in Fig. 2 (b). The designed kernel  $\mathbf{F}$  is specially suitable for cartoon edge extraction, and is essentially an improved Sobel operator. The coarse edge map is obtained by summing over the absolute value of the convolution result with each filter of  $\mathbf{F}$ , followed by Min-Max normalization to rescale to  $[0 - 1]$ :

$$e_i = \text{Norm}_{min\_max}(\sum_{k=1}^4 |\tilde{c}_i \otimes F_k|), i = 1, \dots, N, \quad (2)$$

where  $\otimes$  denotes convolution. Eq. 2 can be efficiently implemented with a single-layer convolution with kernel  $\mathbf{F}$  followed by channel-wise manipulations. Based on coarse edge maps  $E$ , the refined edge maps  $\tilde{E} = \{\tilde{e}_i\}_{i=1}^N$  are obtained by applying a high-pass filter  $h(\cdot)$  that enhances high-frequency pixels and suppresses low-frequency ones:

$$\tilde{e}_i = h(e_i) = 1 - 1/(1 + (e_i/d)^n), i = 1, \dots, N, \quad (3)$$

where  $d$  and  $n$  are hyperparameters that determine threshold and sharpness of the high-pass filter  $h(\cdot)$  respectively. Visualization of the final refined edge maps  $\tilde{E}$  is shown in Fig. 4. The refined edge maps are used to adaptively guide attention to edge-distinct local image regions and extract corresponding image patches:

$$\{c_p^i, e_p^i\}_{i=1}^M = \text{ExtractPatches}(C \oplus \tilde{E}, l, s), \quad (4)$$

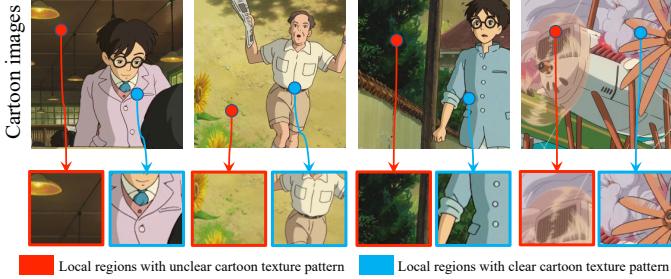


Fig. 3. The typical cartoon texture pattern manifests clearly only in partial image regions with distinct edges.



Fig. 4. Visualization of the refined edge maps  $\tilde{E}$  produced during the forward pass of our CTSS module.

where  $\oplus$  denotes channel-wise concat operation,  $l$  and  $s$  are respectively patch size and stride of the sliding-window-alike patch extraction process.  $c_p^i$  and  $e_p^i$  are the  $i$ th extracted patch of cartoon images  $C$  and edge maps  $\tilde{E}$  respectively, they are paired and correspond to the same image location.  $M$  is the total number of the extracted patches from a mini-batch of training images, i.e.,  $M = N(\lfloor \frac{H-l}{s} \rfloor + 1)(\lfloor \frac{W-l}{s} \rfloor + 1)$ , where  $H$  and  $W$  are height and width of training images, and  $N$  is the batch size. The extracted  $M$  image patches are sorted by the edge intensity of their paired edge patches, where the edge intensity is quantified as the pixel summation over an edge patch. After sorting, top- $K$  image patches  $C_{patch}$  with most distinct edges are sampled:

$$t_i = \sum_{m,n} (e_p^i)_{m,n}, i = 1, 2, \dots, M, \quad (5)$$

$$C_{patch} = \{c_p^{a_j}, j = 1, \dots, K | t_{a_1} \geq t_{a_2} \geq \dots \geq t_{a_M}\}, \quad (6)$$

where  $m$  and  $n$  index pixel coordinate of each local patches,  $\{a_1, a_2, \dots, a_M\}$  is a permutation of  $\{1, 2, \dots, M\}$ . The sampled top- $K$  image patches  $C_{patch}$  contain most evident cartoon texture pattern, they serve as training data of the patch-level adversarial learning branch to promote learning cartoon texture feature. Considering that local patches can not reflect overall color distribution of images, we finally convert the sampled  $C_{patch}$  to grayscale for purpose of learning color-invariant local cartoon texture pattern. Since we use gradient based filtering method to detect edges, some local patches with no clear edges but a lot of noise can still have large accumulated gradients and thus be sampled out for patch-level training. Consequently, we apply guided filtering (Eq. 1) for edge-preserving image denoising before edge extraction (Eq. 2), this guarantees that image patches are sampled out due to clear edges instead of large noises.

It is worth mentioning that our method is more suitable than using pre-trained deep model to detect edges. Firstly,

well-trained deep edge detection models tend to make high-confidence prediction to any edge pixels, the generated edge maps are less able to reflect edge intensity difference, and thus can not locate true edge-salient local regions. Secondly, our filtering based method uses only single-layer convolution to produce edge maps, which is much faster than forward propagation through pre-trained deep models.

### C. Objective functions

The training of our model comprises five loss functions, they are content loss, global adversarial loss, local adversarial loss, color reconstruction loss, and total variation loss.

**Content loss** is used to guarantee content consistency between input photos and cartoonized results, which is realized by matching feature maps at the  $l$ th layer of the pre-trained VGG19 [46] network:

$$L_{con} = \mathbb{E}_{p_i \sim \mathcal{P}} [\|VGG_l(p_i) - VGG_l(G(p_i))\|_1], \quad (7)$$

where the  $l$ th layer is ‘‘conv4-4’’ in VGG19.

**Global adversarial loss** aims to capture global cartoon style through image-level adversarial learning branch. We employ LSGAN [47] loss for better stability:

$$L_{adv\_global} = L_{adv\_global}^D + L_{adv\_global}^G, \quad (8)$$

$$\begin{aligned} L_{adv\_global}^D = & \mathbb{E}_{c_i \sim \mathcal{C}} [(D_{img}(c_i) - 1)^2] + \\ & \mathbb{E}_{p_i \sim \mathcal{P}} [(D_{img}(G(p_i)))^2], \end{aligned} \quad (9)$$

$$L_{adv\_global}^G = \mathbb{E}_{p_i \sim \mathcal{P}} [(D_{img}(G(p_i)) - 1)^2]. \quad (10)$$

**Local adversarial loss** aims at learning local cartoon texture pattern through patch-level adversarial learning branch:

$$L_{adv\_local} = L_{adv\_local}^D + L_{adv\_local}^G, \quad (11)$$

$$\begin{aligned} L_{adv\_local}^D = & \mathbb{E}_{C_{patch}} \left[ \frac{1}{K} \sum_{i=1}^K (D_{patch}(c_p^i) - 1)^2 \right] \\ & + \mathbb{E}_{S_{patch}} \left[ \frac{1}{K} \sum_{i=1}^K (D_{patch}(s_p^i))^2 \right], \end{aligned} \quad (12)$$

$$L_{adv\_local}^G = \mathbb{E}_{S_{patch}} \left[ \frac{1}{K} \sum_{i=1}^K (D_{patch}(s_p^i) - 1)^2 \right], \quad (13)$$

where  $C_{patch} = \{c_p^i\}_{i=1}^K$  are extracted top- $K$  edge-distinct patches from  $C = \{c_i\}_{i=1}^N$ ,  $S_{patch} = \{s_p^i\}_{i=1}^K$  are top- $K$  edge-distinct patches from  $S = \{s_i\}_{i=1}^N = \{G(p_i)\}_{i=1}^N$ .

**Color reconstruction loss** is used to retain color information after cartoonization. Following [7], we convert image from RGB to YUV format, and apply  $L_1$  loss to Y channel and Huber Loss to U and V channels:

$$\begin{aligned} L_{col} = & \mathbb{E}_{p_i \sim \mathcal{P}} [\|Y(G(p_i)) - Y(p_i)\|_1 + \\ & \|U(G(p_i)) - U(p_i)\|_H + \|V(G(p_i)) - V(p_i)\|_H], \end{aligned} \quad (14)$$

where  $Y(\cdot)$ ,  $U(\cdot)$ ,  $V(\cdot)$  represent the three channels of an image in YUV format, and  $H$  denotes Huber Loss.

**Total variation loss** is used to reduce noises and artifacts of the generated results:

$$\begin{aligned} L_{tv} = & \mathbb{E}_{s_i \sim \mathcal{S}} \left[ \frac{1}{H(W-1)} \sum_{r=1}^H \sum_{c=1}^{W-1} (s_{i_{r,c+1}} - s_{i_{r,c}})^2 \right. \\ & \left. + \frac{1}{(H-1)W} \sum_{r=1}^{H-1} \sum_{c=1}^W (s_{i_{r+1,c}} - s_{i_{r,c}})^2 \right], \end{aligned} \quad (15)$$

where  $H$  and  $W$  are height and width of generated images. The total loss function can be decomposed into a generator part and a discriminator part:

$$L_{gen} = \lambda_{global} L_{adv\_global}^G + \lambda_{local} L_{adv\_local}^G + \lambda_{con} L_{con} + \lambda_{col} L_{col} + \lambda_{tv} L_{tv}, \quad (16)$$

$$L_{dis} = \lambda_{global} L_{adv\_global}^D + \lambda_{local} L_{adv\_local}^D, \quad (17)$$

where  $L_{gen}$  is minimized to optimize the generator  $G$ ,  $L_{dis}$  is minimized to jointly optimize the two discriminators  $D_{img}$  and  $D_{patch}$ .  $L_{gen}$  and  $L_{dis}$  are minimized alternately to form the adversarial training framework.

## IV. EXPERIMENTS

### A. Datasets

Our model can be easily trained with unpaired data. The source-domain real-scene photos comprise 6656 images for training, 790 images for quantitative testing, and 300 high-resolution images for qualitative evaluation. The training and quantitative testing sets are borrowed from the training and testing sets of CycleGAN [38] respectively, where the spatial size of all images are fixed at  $256 \times 256$  for fair comparison with baseline models [6], [7], [9]. We additionally collect 300 high-resolution pictures with width ranging from 960 to 3000 pixels to qualitatively evaluate models' cartoonization effect on practical large input pictures. For target-domain training data, we prepare three cartoon datasets of different styles. They are respectively consisted of cartoon frames cropped from "The Wind Rises", "Dragon Ball", and "Crayon Shin-chan". Each cartoon dataset has 2000 cartoon images, they are rescaled to  $256 \times 256$  to be in accordance with the size of source domain training images.

### B. Parameter settings

The training batch size is  $N=8$ , the number of local patches extracted from each mini-batch is  $K=32$ . We set  $d=0.2$ ,  $n=2$  in Eq. 3, which empirically has good effect in refining edge maps. For patch extraction, we set patch size  $l=96$ , and randomly sample moving stride  $s$  in Eq. 4 from a uniform distribution  $U(48, 72)$  at each training iteration, as a kind of patch-level data augmentation strategy.

### C. Training and inference details

At training phase, we set the weights of component loss functions to be  $\lambda_{global}=\lambda_{local}=300$ ,  $\lambda_{con}=1.5$ ,  $\lambda_{col}=15$ ,  $\lambda_{tv}=1$ . We train our model for 80 epochs. In the first 10 epochs, we pre-train the generator by minimizing  $L_{con}$  (Eq. 7) with an initial learning rate of  $2 \times 10^{-4}$ . In the remaining 70 epochs, we alternately minimize  $L_{gen}$  (Eq. 16) and  $L_{dis}$  (Eq. 17) to optimize the generator part and the discriminator part respectively, where the initial learning rate for both two parts are  $2 \times 10^{-5}$ . We use Adam optimizer with  $\beta_1=0.5$ ,  $\beta_2=0.999$ . At inference time, we dynamically increase image natural saturation of the generated results by 40% to produce more vivid colors.

### D. Qualitative and quantitative results

Some qualitative results of our model trained over different cartoon datasets and evaluated over high-resolution input pictures are shown in Fig. 5, our method reproduces vivid cartoon effects on high-resolution real-world-scene photos. For qualitative method comparison, we divide related methods into general image stylization or abstraction methods including neural style transfer (NST) [8], flow-based image abstraction (FBIA) [4], and CycleGAN [38], as well as advanced image cartoonization methods including CartoonGAN [6], AnimeGAN [7], and WhiteBox [9]. Results of these models trained over "The Wind Rises" dataset are shown in Fig. 6 and Fig. 7. NST globally and randomly transfers low-level texture features, the produced results suffer from unappealing artifacts and structure distortions. FBIA abstracts images with learning-free image processing techniques, the generated results fail to present the distribution of target cartoon style. Results of CycleGAN do not manifest local cartoon texture features such as clear edges.

For all the advanced image cartoonization methods to be compared with, we keep the same training epochs as our method. As Fig. 7 demonstrates, the stylization degree of CartoonGAN and AnimeGAN is relatively weak, the typical local cartoon texture pattern is not evident in generated results. Note that both these two methods prepare an edge-smoothed cartoon dataset to explicitly learn to promote edges adversarially, the generated edges are still not as clear as ours. Results of White-Box do not exhibit much cartoon abstraction and vividness for high-resolution input images. All these advanced methods are not able to fully capture the inherent cartoon texture features. By contrast, our results present prominent cartoon texture pattern, and sufficient abstraction and vividness. Apart from landscape pictures, results of our method evaluated on "Crayon Shin-chan" dataset in more scenarios are displayed in Fig. 9.

We quantitatively evaluate model performance by measuring the FID between the collection of generated images and the collection of real cartoon images. Results of FID evaluated on the source-domain test set are reported in Tab. I. Our method achieves much lower FID than related GAN-based methods, indicating that the style distribution of our results is much closer to that of target-domain real cartoon images. Besides, we investigate the influence of  $K$  (i.e., the total number of local patches sampled from each mini-batch of input images) to model performance. Results reported in Tab. II show that both too small and too large value of  $K$  may degrade model performance. This is because sampling less patches leads to training insufficiency of patch-level adversarial learning, while oversampling patches results in many low-quality patches with less salient cartoon textures, which weakens our model's ability to capture and highlight cartoon texture features.

### E. Ablation study

Ablation studies of the components of our model are qualitatively and quantitatively shown in Fig. 8 and Tab. III respectively. Removing patch-level adversarial learning branch leads to results with weak cartoon styles and large FID, reflecting that image-level adversarial loss alone is not sufficient

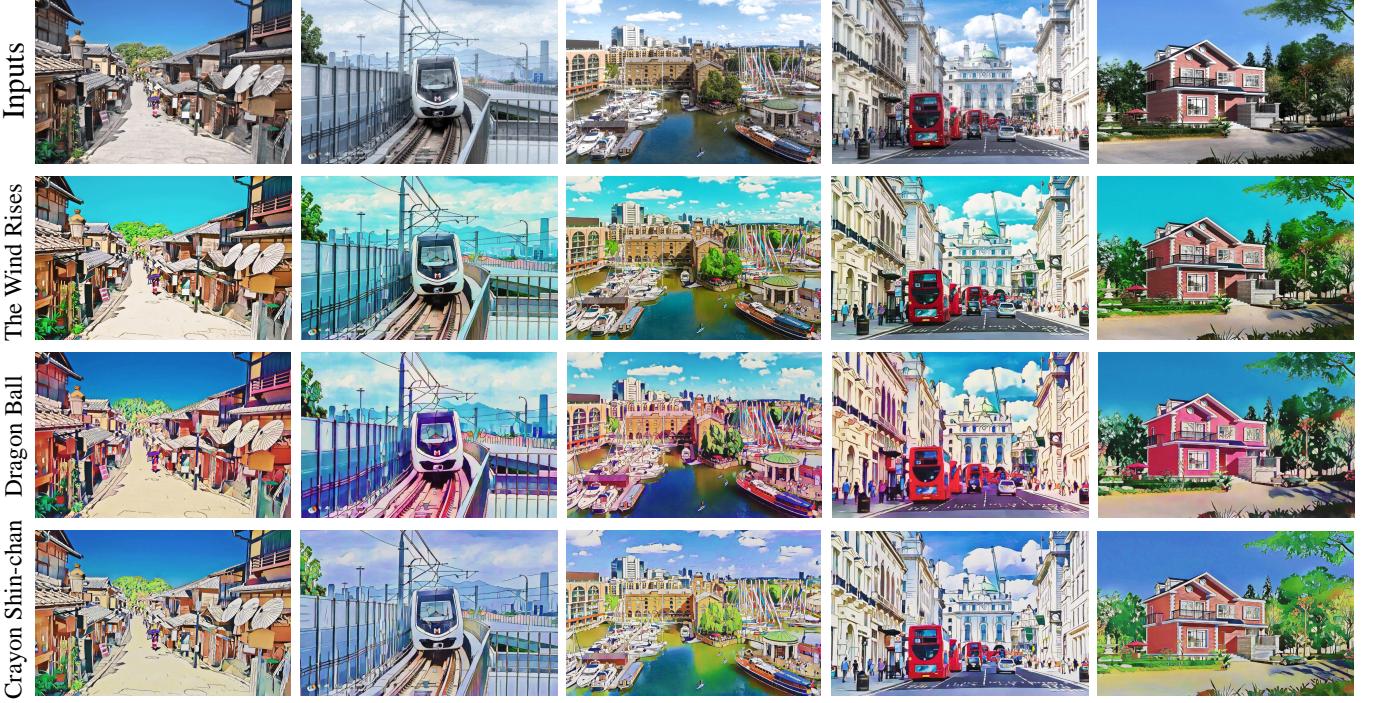


Fig. 5. Example image cartoonization results tested over high-resolution real-world-scene input images. Results are evaluated on our model trained over different cartoon datasets, including “The Wind Rises” (the second row), “Dragon Ball” (the third row), and “Crayon Shin-chan” (the bottom row). **Please zoom in for better resolution.**

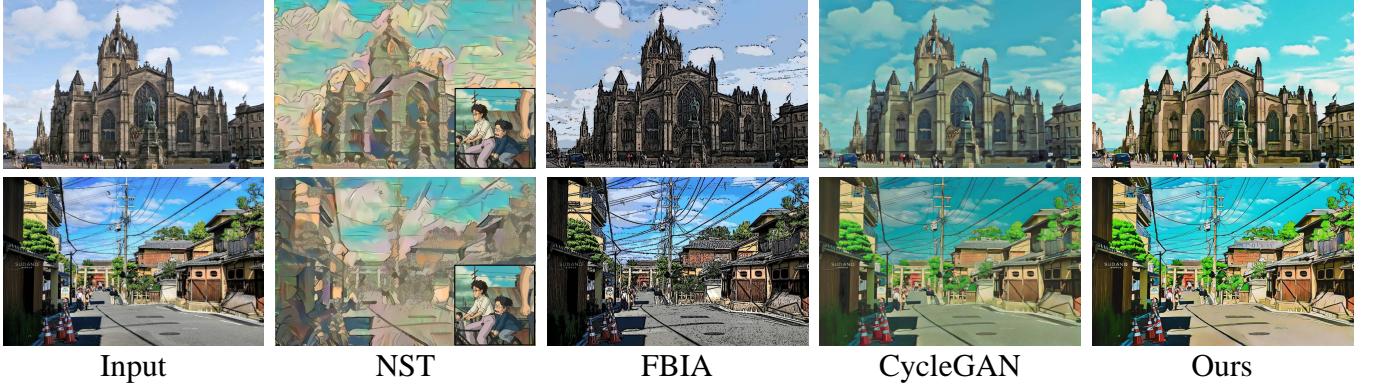


Fig. 6. Comparison of our approach with general image stylization or abstraction methods tested over high-resolution real-world-scene input images (**better to zoom in**). Results are evaluated on “The Wind Rises” dataset.

TABLE I  
COMPARISON WITH RELATED METHODS IN FID ↓.

Models	FID		
	TWR	DB	CSC
CycleGAN	146.45	141.33	142.86
CartoonGAN	143.96	145.58	147.36
AnimeGAN	136.12	134.94	138.83
WhiteBox	132.67	137.45	140.79
Ours	<b>115.25</b>	<b>112.97</b>	<b>124.66</b>

TWR, DB, CSC respectively denote “The Wind Rises”, “Dragon Ball”, and “Crayon Shin-chan” dataset.

to transfer salient cartoon texture features. When removing the image-level learning branch, we observe that the model with

TABLE II  
INFLUENCE OF  $K$  TO FID ↓.

Dataset	FID			
	$K=16$	$K=32$	$K=64$	$K=128$
TWR	115.86	<b>115.25</b>	116.47	126.78
DB	113.53	<b>112.97</b>	113.37	125.64
CSC	125.45	<b>124.66</b>	124.94	130.33

TWR, DB, and CSC denote “The Wind Rises”, “Dragon Ball”, and “Crayon Shin-chan” cartoon dataset respectively.

only patch-level branch suffers from training instability. The adversarial learning fails to converge and the corresponding results do not exhibit any cartoon styles. This indicates the fundamental role of image-level branch in maintaining the



Fig. 7. Comparison of our approach with related advanced image cartoonization methods tested over high-resolution real-world-scene input images (**better to zoom in for details**). Results are evaluated on “The Wind Rises” dataset. For all related methods, the image natural saturation of the generated results is increased by 40% as well for fair comparison with our method.

TABLE III  
ABLATION STUDY OF OUR MODEL WITH FID ↓.

Models	FID		
	TWR	DB	CSC
W/O patch-level branch	145.25	146.63	149.84
W/O image-level branch	182.78	185.66	190.04
W/O CTSS (all patches)	127.09	125.80	130.90
Ours (full model)	<b>115.25</b>	<b>112.97</b>	<b>124.66</b>

balance of adversarial training. Combining both two branches constitutes our full model that produces visually appealing cartoon effects with prominent cartoon texture pattern and substantially lower FID. Lastly, to verify that the performance gain stems from not only patch-level adversarial learning but also adaptive texture saliency local attention brought by our CTSS module, we remove CTSS module and extract all local patches from each mini-batch of images for patch-level adversarial learning, the corresponding results are noticeably inferior to our model with CTSS module both qualitatively and quantitatively. This indicates that adaptively constrain patch-level adversarial learning on cartoon-texture-salient local regions is indeed contributive to enhancing cartoon style rendering.

Besides, we investigate the effectiveness of style related

TABLE IV  
ABLATION STUDY OF LOSS FUNCTIONS FOR STYLIZATION.

Loss functions for stylization	FID		
	TWR	DB	CSC
$L_{adv\_global}$	145.25	146.63	149.84
$L_{ep\_adv\_global}$	143.40	145.64	147.75
$L_{adv\_global} + L_{gram}$	139.05	137.71	141.66
$L_{ep\_adv\_global} + L_{gram}$	137.43	135.50	140.18
$L_{ep\_adv\_global} + L_{mean\_std}$	135.99	135.24	137.54
$L_{ep\_adv\_global} + L_{gram} + L_{mean\_std}$	135.45	134.60	136.08
$L_{adv\_global} + L_{adv\_local}$ (ours)	<b>115.25</b>	<b>112.97</b>	<b>124.66</b>

$L_{ep\_adv\_global}$  denotes the edge-promoting adversarial loss proposed in CartoonGAN [6].

$L_{gram}$  and  $L_{mean\_std}$  are Gram loss [8] and mean-variance loss [23] respectively.

loss functions in image cartoonization, results are reported in Tab. IV. Compared with the basic image-level adversarial loss  $L_{adv\_global}$  (Eq. 8), the edge-promoting adversarial loss  $L_{ep\_adv\_global}$  used in CartoonGAN [6] brings very limited performance gains at the cost of a prior edge-smoothing data preparation stage. Following the idea of AnimeGAN [7] to introduce style transfer losses into GAN framework, we combine the image-level adversarial loss  $L_{adv\_global}$  or  $L_{ep\_adv\_global}$  with the second-order style transfer loss  $L_{gram}$



Fig. 8. Qualitative ablation study of the components of our framework, including the image-level learning branch, patch-level learning branch, and our CTSS module (**better to zoom in**). Results are evaluated on our model trained over “The Wind Rises” dataset.

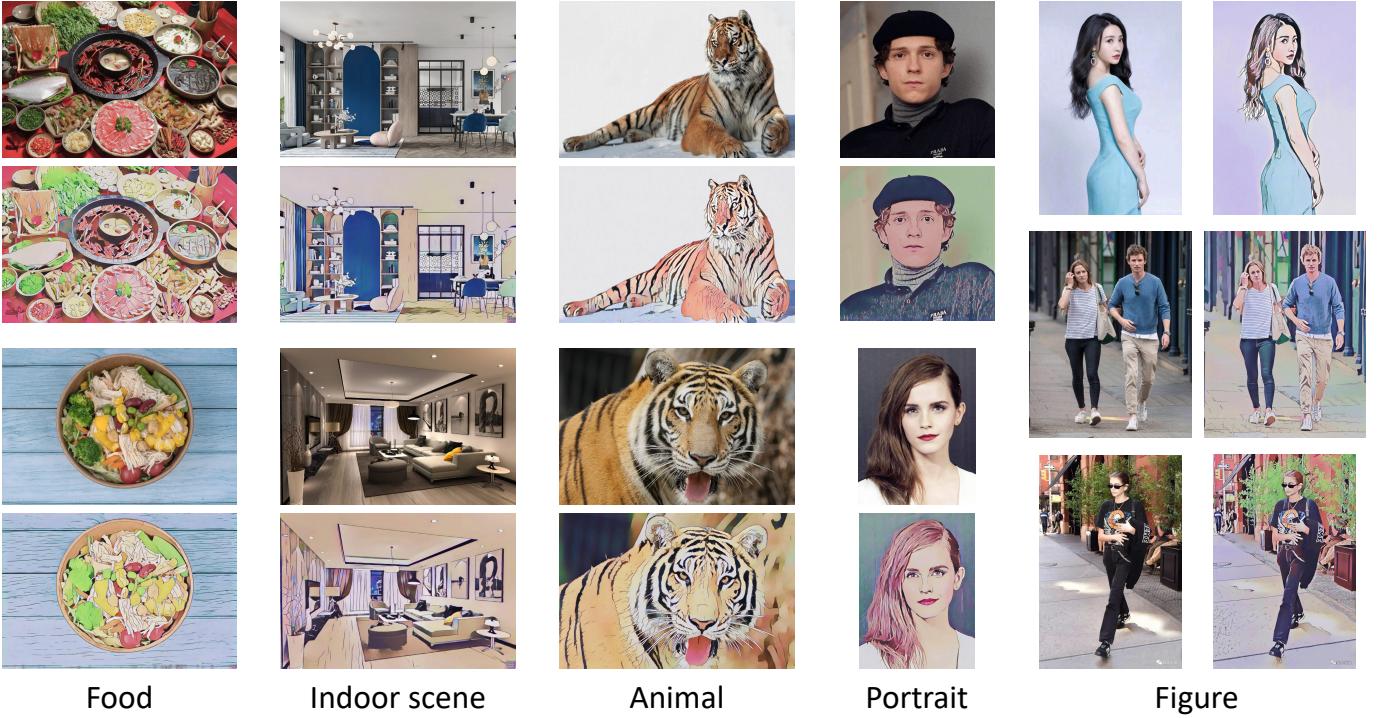


Fig. 9. Qualitative cartoonization results evaluated on “Crayon Shin-chan” dataset in more scenarios including foods, indoor scenes, animals, portraits, and figures. **Better to zoom in for higher resolution.**

[8] or the first-order one  $L_{mean\_std}$  [23]. Results show that introduction of style transfer losses is useful to narrow style distribution gap between generated results and real cartoons. However, these extra style losses transfer global and low-level image styles [48], not able to fully capture local cartoon texture pattern. By contrast, our method improves stylization performance dramatically simply by applying the basic adversarial loss at local views, which bypasses the need of prior data preprocessing stage and additional style transfer losses.

## V. CONCLUSION

This paper proposes a deep generative model for image cartoonization. We supplement the normal image-level adversarial training with a patch-level learning branch, and adaptively constrain patch-level adversarial learning only on cartoon-texture-

salient local regions to enhance capturing cartoon texture pattern. To this end, a novel cartoon-texture-saliency-sampler (CTSS) module is proposed to dynamically sample local image patches containing most prominent cartoon texture features. By incorporating such texture saliency adaptive attention to adversarial learning, our method is able to transfer noticeably more abstract and vivid cartoon styles.

## REFERENCES

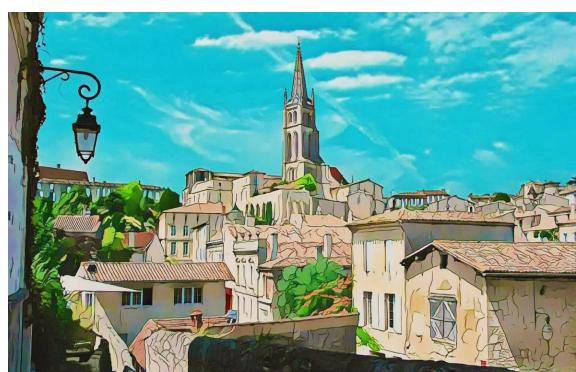
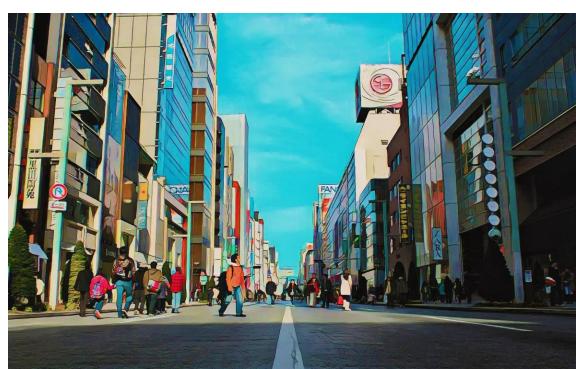
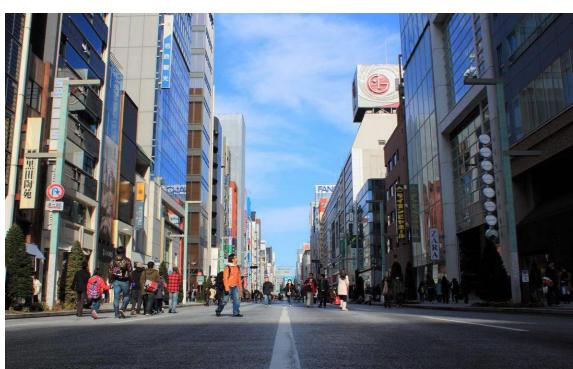
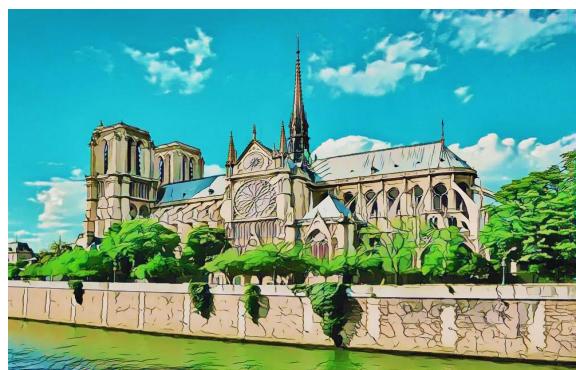
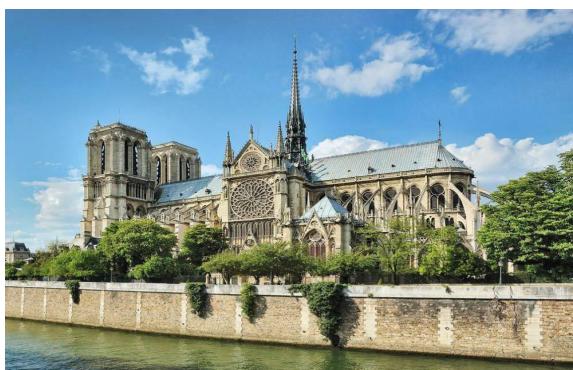
- [1] D. DeCarlo and A. Santella, “Stylization and abstraction of photographs,” *ACM transactions on graphics (TOG)*, vol. 21, no. 3, pp. 769–776, 2002.
- [2] H. Winnemöller, S. C. Olsen, and B. Gooch, “Real-time video abstraction,” *ACM Transactions On Graphics (TOG)*, vol. 25, no. 3, pp. 1221–1226, 2006.
- [3] J. E. Kyprianidis and J. Döllner, “Image abstraction by structure adaptive filtering.” in *TPCG*, 2008, pp. 51–58.

- [4] H. Kang, S. Lee, and C. K. Chui, "Flow-based image abstraction," *IEEE transactions on visualization and computer graphics*, vol. 15, no. 1, pp. 62–76, 2008.
- [5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.
- [6] Y. Chen, Y.-K. Lai, and Y.-J. Liu, "Cartoongan: Generative adversarial networks for photo cartoonization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9465–9474.
- [7] J. Chen, G. Liu, and X. Chen, "Animegan: A novel lightweight gan for photo animation," in *International Symposium on Intelligence Computation and Applications*. Springer, 2019, pp. 242–256.
- [8] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2414–2423.
- [9] X. Wang and J. Yu, "Learning to cartoonize using white-box cartoon representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8090–8099.
- [10] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2424–2433.
- [11] R. Du, D. Chang, A. K. Bhunia, J. Xie, Y.-Z. Song, Z. Ma, and J. Guo, "Fine-grained visual classification via progressive multi-granularity training of jigsaw patches," in *European Conference on Computer Vision*, 2020.
- [12] X. Gao, Y. Tian, and Z. Qi, "Multi-view feature augmentation with adaptive class activation mapping," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 2021, pp. 678–684.
- [13] M. Pesavento, M. Volino, and A. Hilton, "Attention-based multi-reference learning for image super-resolution," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14 677–14 686, 2021.
- [14] V. Papyan and M. Elad, "Multi-scale patch-based image restoration," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 249–261, 2016.
- [15] X. Yan, W. Zhao, K. Yuan, R. Zhang, Z. Li, and S. Cui, "Towards content-independent multi-reference super-resolution: Adaptive pattern matching and feature aggregation," in *Computer Vision – ECCV 2020*. Springer International Publishing, 2020, pp. 52–68.
- [16] Z. Chen, Y. Zhu, C. Zhao, G. Hu, W. Zeng, J. Wang, and M. Tang, "Dpt: Deformable patch-based transformer for visual recognition," in *Proceedings of the ACM International Conference on Multimedia*, 2021.
- [17] C. Guo, B. Fan, J. Gu, Q. Zhang, S. Xiang, V. Prinet, and C. Pan, "Progressive sparse local attention for video object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3908–3917.
- [18] Y. Chai, "Patchwork: A patch-wise attention network for efficient object detection and segmentation in video streams," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, nov 2019, pp. 3414–3423.
- [19] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.
- [20] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images," in *ICML*, vol. 1, no. 2, 2016, p. 4.
- [21] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "Stylebank: An explicit representation for neural image style transfer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1897–1906.
- [22] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," *arXiv preprint arXiv:1610.07629*, 2016.
- [23] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510.
- [24] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," in *NIPS*, 2017.
- [25] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 2230–2236.
- [26] C. Li and M. Wand, "Combining markov random fields and convolutional neural networks for image synthesis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2479–2486.
- [27] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 768–783.
- [28] N. Kolkin, J. Salavon, and G. Shakhnarovich, "Style transfer by relaxed optimal transport and self-similarity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 051–10 060.
- [29] Y. Tian, "zi2zi: Master chinese calligraphy with conditional adversarial networks, 2017," *Retrieved Jun*, vol. 3, p. 2017, 2017.
- [30] Y. Jiang, Z. Lian, Y. Tang, and J. Xiao, "Dcfont: an end-to-end deep chinese font generation system," in *SIGGRAPH Asia 2017 Technical Briefs*, 2017, pp. 1–4.
- [31] B. He, F. Gao, D. Ma, B. Shi, and L.-Y. Duan, "Chigan: A generative adversarial network for chinese ink wash painting style transfer," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 1172–1180.
- [32] X. Gao, Y. Tian, and Z. Qi, "Rpd-gan: Learning to draw realistic paintings with generative adversarial network," *IEEE Transactions on Image Processing*, vol. 29, pp. 8706–8720, 2020.
- [33] H. Chang, J. Lu, F. Yu, and A. Finkelstein, "Pairedcyclegan: Asymmetric style transfer for applying and removing makeup," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 40–48.
- [34] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.
- [35] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.
- [36] A. Alsaifi, R. Rustagi, M. M. Thomas, A. G. Forbes *et al.*, "Image denoising using a generative adversarial network," in *2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT)*. IEEE, 2019, pp. 126–132.
- [37] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.
- [38] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.
- [39] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in neural information processing systems*, 2017, pp. 700–708.
- [40] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797.
- [41] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "AttnGAN: Facial attribute editing by only changing what you want," *IEEE transactions on image processing*, vol. 28, no. 11, pp. 5464–5478, 2019.
- [42] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 172–189.
- [43] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 35–51.
- [44] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, and J. Kautz, "Few-shot unsupervised image-to-image translation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 551–10 560.
- [45] K. He, J. Sun, and X. Tang, "Guided image filtering," in *European conference on computer vision*. Springer, 2010, pp. 1–14.
- [46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [47] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.
- [48] S. Li, X. Xu, L. Nie, and T.-S. Chua, "Laplacian-steered neural style transfer," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1716–1724.

<b>Notations</b>					
LN: layer normalization		IRelu: leaky Relu with $\alpha = 0.2$			
Conv_n(N)k(K)s(S): Convolutional layer with N filters, K×K kernel size, and stride S					
Upsample_n(N): Upsampling module consisting of following layers: nearest neighbor upsampling with factor 2 → Conv_n(N)k(3)s(1) → LN → IRelu					
DConv_n(N)k(K)s(S): Depthwise convolutional layer with N filters, K×K kernel size, and stride S					
<b>Generator (G)</b>		<b>Image-level discriminator (<math>D_{img}</math>)</b>			
Layers	Shape	Layers	Shape		
Input	256x256x3	Input	256x256x3		
Conv_n(32)k(7)s(1), LN, IRelu	256x256x32	Conv_n(32)k(3)s(1), LN, IRelu	256x256x32		
Conv_n(64)k(3)s(2), LN, IRelu	128x128x64	Conv_n(64)k(3)s(2), LN, IRelu	128x128x64		
Conv_n(64)k(3)s(1), LN, IRelu	128x128x64	Conv_n(128)k(3)s(2), LN, IRelu	64x64x128		
Conv_n(128)k(3)s(2), LN, IRelu	64x64x128	Conv_n(256)k(3)s(2), LN, IRelu	32x32x256		
Conv_n(256)k(3)s(1), LN, IRelu	64x64x256	Conv_n(1)k(3)s(1)	32x32x1		
ResBlock	64x64x256	<b>Patch-level discriminator (<math>D_{patch}</math>)</b>			
ResBlock	64x64x256	Layers	Shape		
ResBlock	64x64x256	Input	96x96x1		
ResBlock	64x64x256	Conv_n(16)k(3)s(1), LN, IRelu	96x96x16		
Conv_n(128)k(3)s(1), LN, IRelu	64x64x128	Conv_n(32)k(3)s(2), LN, IRelu	48x48x32		
Upsample_n(128)	128x128x128	Conv_n(64)k(3)s(2), LN, IRelu	24x24x64		
Conv_n(128)k(3)s(1), LN, IRelu	128x128x128	Conv_n(128)k(3)s(2), LN, IRelu	12x12x128		
Upsample_n(64)	256x256x64	Conv_n(1)k(3)s(1)	12x12x1		
Conv_n(64)k(3)s(1), LN, IRelu	256x256x64	<b>ResBlock</b>			
Conv_n(64)k(3)s(1), LN, IRelu	256x256x64	Layers	Shape		
Conv_n(32)k(7)s(1), LN, IRelu	256x256x32	Input	64x64x256		
Conv_n(3)k(1)s(1), Tanh	256x256x3	Conv_n(128)k(1)s(1), LN, IRelu	64x64x128		
		DConv_n(128)k(3)s(1), LN, IRelu	64x64x128		
		Conv_n(256)k(1)s(1), LN	64x64x256		
		Add input	64x64x256		

Fig. 10. Network details of the generator  $G$ , the image-level discriminator  $D_{img}$ , and the patch-level discriminator  $D_{patch}$ .

## More cartoonization results on “The Wind Rises” dataset



Input Photo

Cartoonized Results

## More cartoonization results on “The Wind Rises” dataset



Input Photo

Cartoonized Results

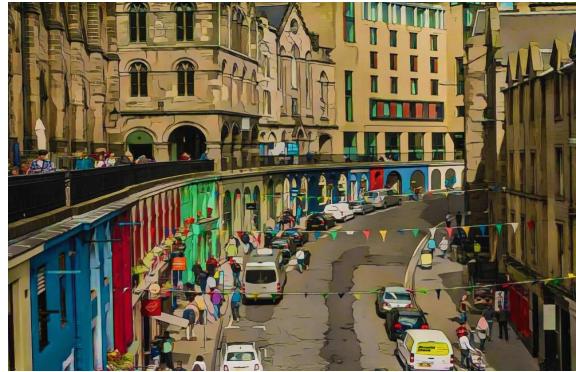
## More cartoonization results on “The Wind Rises” dataset



Input Photo

Cartoonized Results

## More cartoonization results on “The Wind Rises” dataset



Input Photo

Cartoonized Results

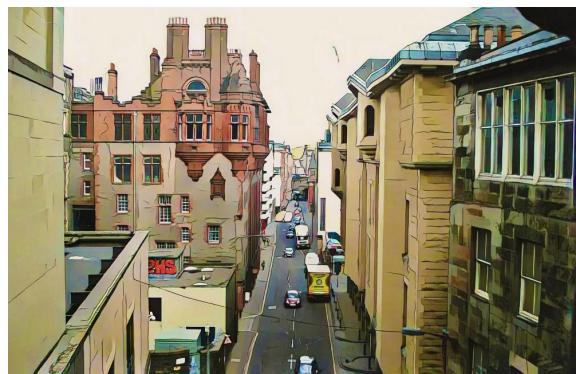
## More cartoonization results on “The Wind Rises” dataset



Input Photo

Cartoonized Results

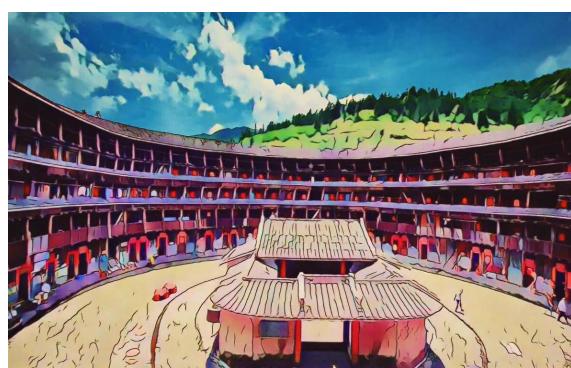
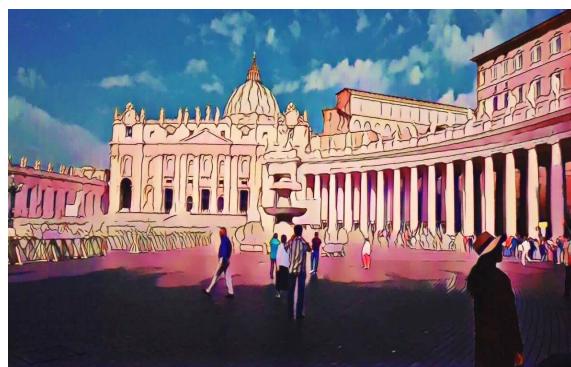
## More cartoonization results on “The Wind Rises” dataset



Input Photo

Cartoonized Results

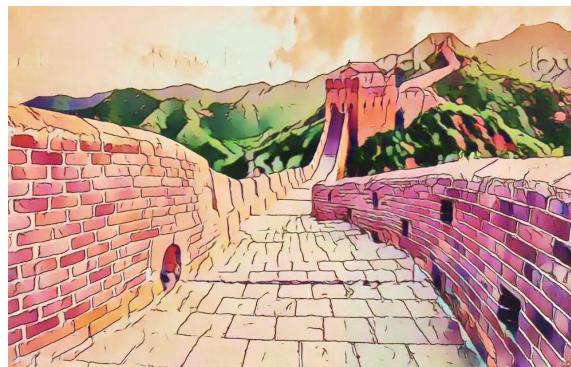
## More cartoonization results on “Dragon Ball” dataset



Input Photo

Cartoonized Results

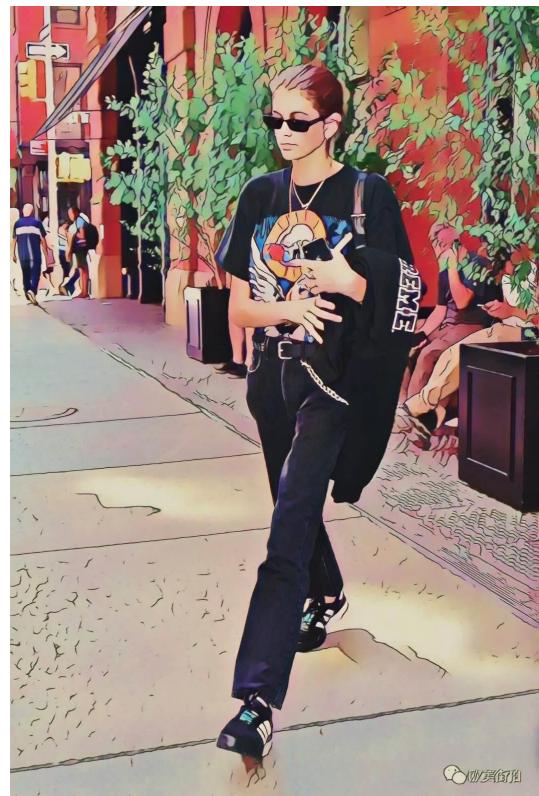
## More cartoonization results on “Dragon Ball” dataset



Input Photo

Cartoonized Results

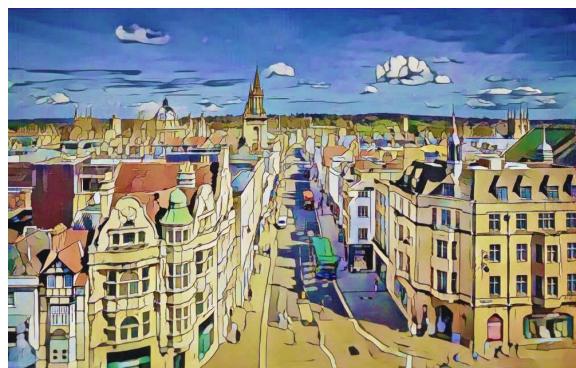
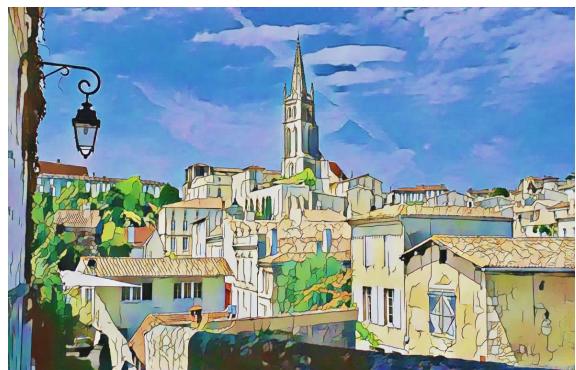
## More cartoonization results on “Dragon Ball” dataset



Input Photo

Cartoonized Results

## More cartoonization results on “Crayon Shin-chan” dataset



Input Photo

Cartoonized Results

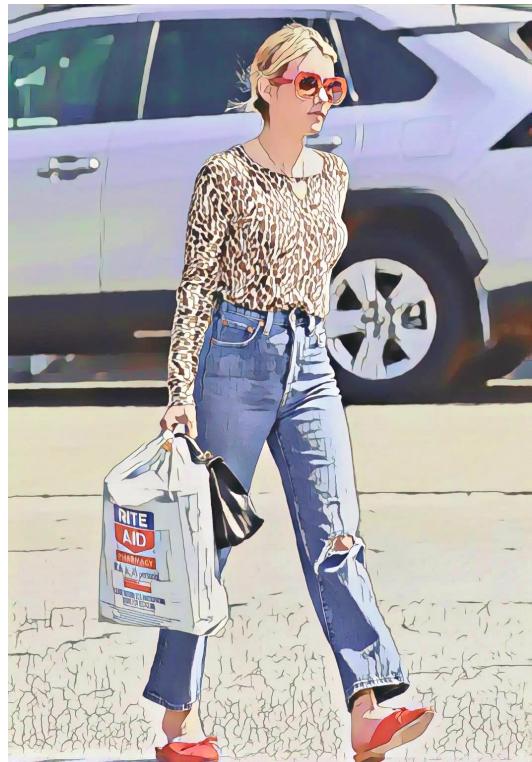
## More cartoonization results on “Crayon Shin-chan” dataset



Input Photo

Cartoonized Results

## More cartoonization results on “Crayon Shin-chan” dataset



Input Photo

Cartoonized Results