
Learning to Incorporate Texture Saliency Adaptive Attention to Image Cartoonization

Xiang Gao^{*12} Yuqi Zhang^{*1} Yingjie Tian¹

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.

1. Introduction

Image cartoonization aims at rendering natural images into cartoon styles, which is a challenging computer vision and

^{*}Equal contribution ¹University of Chinese Academy of Sciences ²Noah’s Ark Lab, Huawei Technologies. Correspondence to: Yingjie Tian <tyj@ucas.ac.cn>.



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

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 (DeCarlo & Santella, 2002; Winnemöller et al., 2006; Kyprianidis & Döllner, 2008; Kang et al., 2008). These methods model cartoon styles with established image processing techniques. 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) (Goodfellow et al., 2014), 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 (Chen et al., 2018) 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 (Chen et al., 2019) introduces Gram loss (Gatys et al., 2016), 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 (Wang & Yu, 2020) 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.

To this end, 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.

2. Related Work

2.1. Style Transfer

Neural style transfer (NST) was first proposed as an online optimization-based algorithm that iteratively transfers image styles by minimizing Gram loss (Gatys et al., 2016). Afterwards, it was transformed into an offline generative model to cater for real-time applications by training a feed-forward network (Johnson et al., 2016; Ulyanov et al., 2016). Later on, efforts had been made to extend fast NST from a single style to multiple styles (Chen et al., 2017; Dumoulin et al., 2016), or even arbitrary styles (Huang & Belongie, 2017; Li et al., 2017b). Beside the Gram loss, various style losses were successively proposed, such as MMD loss (Li et al., 2017c), CNMRF loss (Li & Wand, 2016), contextual loss (Mechrez et al., 2018), and Relaxed EMD loss (Kolkin et al., 2019). 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 (Tian, 2017; Jiang et al., 2017), painting style transfer (He et al., 2018; Gao et al., 2020), makeup style transfer (Chang et al., 2018), etc.

2.2. 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 (Isola et al.,

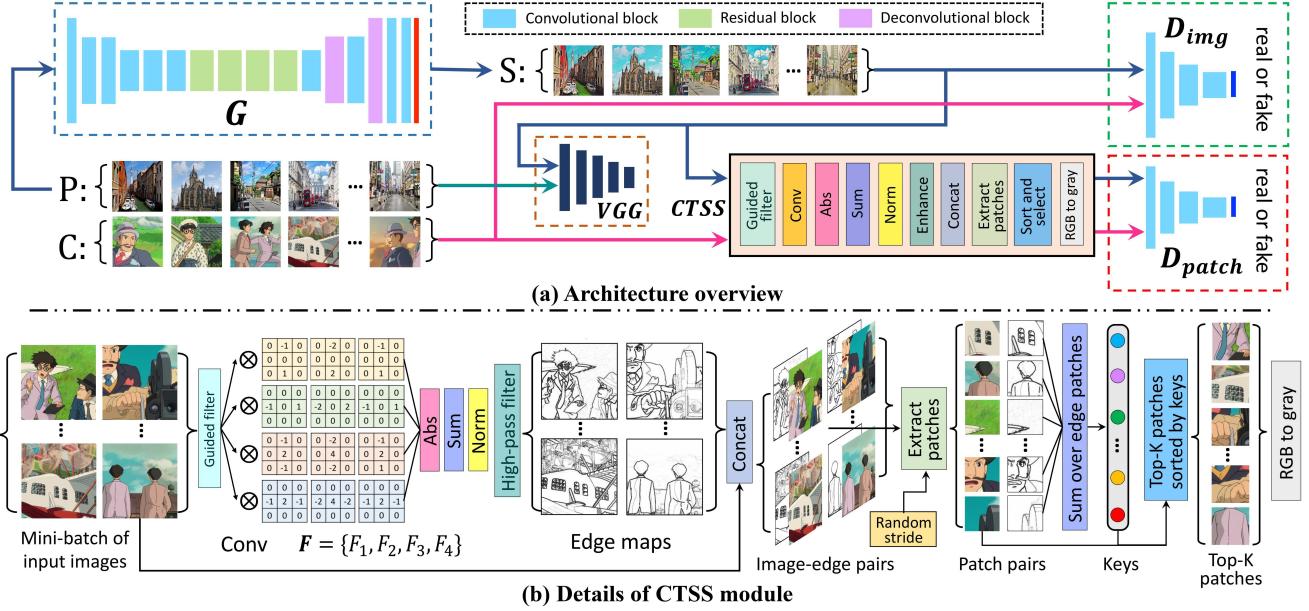


Figure 2. The overall architecture of our model, as well as details of our proposed cartoon-texture-saliency-sampler (CTSS) module which adaptively extracts local image patches with most salient cartoon texture pattern from each mini-batch of input images.

2017) combines conditional GAN with image-level sparse L_1 regularization, which generalizes well to many applications such as image super-resolution (Ledig et al., 2017), image denoising (Alsaifi et al., 2019), semantic image synthesis (Wang et al., 2018), etc. For the latter situation, CycleGAN (Zhu et al., 2017) and UNIT (Liu et al., 2017) are typical models that realize unsupervised image translation based on cycle-consistency constraint and shared-latent-space assumption, respectively. Afterwards, methods like StarGAN (Choi et al., 2018) and AttGAN (He et al., 2019) 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 (Huang et al., 2018; Lee et al., 2018) and few-shot learning (Liu et al., 2019).

3. 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 interwoven. 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.

3.1. 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.

3.2. 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 (He et al., 2010) 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 $[0 - 1]$:

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

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:

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

where \oplus denotes channel-wise concat operation, l and s are respectively patch size and stride of the sliding-window-like 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.

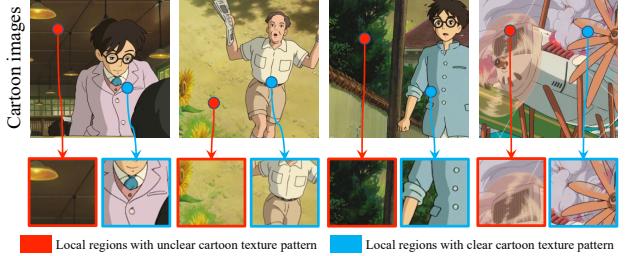


Figure 3. The typical cartoon texture pattern manifests clearly only in partial image regions with distinct edges.



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

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.

3.3. 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 (Simonyan & Zisserman, 2014) 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 (Mao et al., 2017) loss for better stability:

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

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

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

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 (Chen et al., 2019), 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:

$$\begin{aligned} L_{gen} = & \lambda_{global} L_{adv_global}^G + \lambda_{local} L_{adv_local}^G + \\ & \lambda_{con} L_{con} + \lambda_{col} L_{col} + \lambda_{tv} L_{tv}, \end{aligned} \quad (16)$$

$$L_{dis} = \lambda_{global} L_{adv_global}^D + \lambda_{local} L_{adv_local}^D, \quad (17)$$

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.

4. Experiments

4.1. 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 (Zhu et al., 2017) respectively, where the spatial size of all images are fixed at 256×256 for fair comparison with baseline models (Chen et al., 2018; 2019; Wang & Yu, 2020). 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 256×256 to be in accordance with the size of source domain training images.

4.2. 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

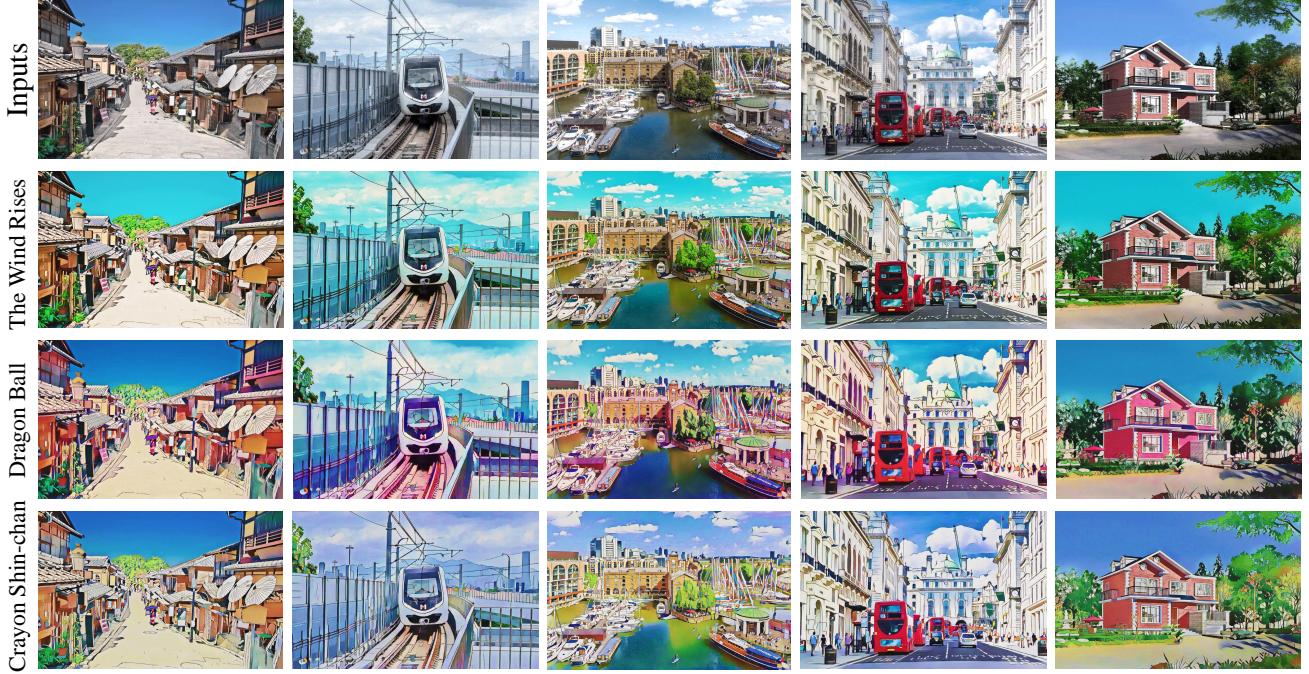


Figure 5. Example image cartoonization results tested over high-resolution real-world-scene input images. Results are evaluated on our model trained over different cartoon datasets, including “The Wind Rises” (the second row), “Dragon Ball” (the third row), and “Crayon Shin-chan” (the bottom row). **Please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

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.

4.3. 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 2×10^{-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 2×10^{-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.

4.4. 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) (Gatys et al., 2016), flow-based image abstraction (FBIA) (Kang et al., 2008), and CycleGAN (Zhu et al., 2017), as well as advanced image cartoonization methods including CartoonGAN (Chen et al., 2018), AnimeGAN (Chen et al., 2019), and White-Box (Wang & Yu, 2020). 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 WhiteBox 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,



Figure 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.



Figure 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.

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. 1. Our method achieves much lower FID than related GAN-based methods, indicating that the style distribution of our results



Figure 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.

Table 1. 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	115.25	112.97	124.66

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

Table 2. Influence of K to FID ↓.

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

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

Table 3. 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)	115.25	112.97	124.66

Table 4. 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)	115.25	112.97	124.66

$L_{ep_adv_global}$ denotes the edge-promoting adversarial loss proposed in CartoonGAN (Chen et al., 2018).

L_{gram} and L_{mean_std} are Gram loss (Gatys et al., 2016) and mean-variance loss (Huang & Belongie, 2017) respectively.

4.5. Ablation study

Ablation studies of the components of our model are qualitatively and quantitatively shown in Fig. 8 and Tab. 3 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 to transfer salient cartoon texture features. When removing the image-level learning branch, we observe that the model with only patch-level branch suffers from training instability. The adversarial learning fails to converge and

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. 2 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.

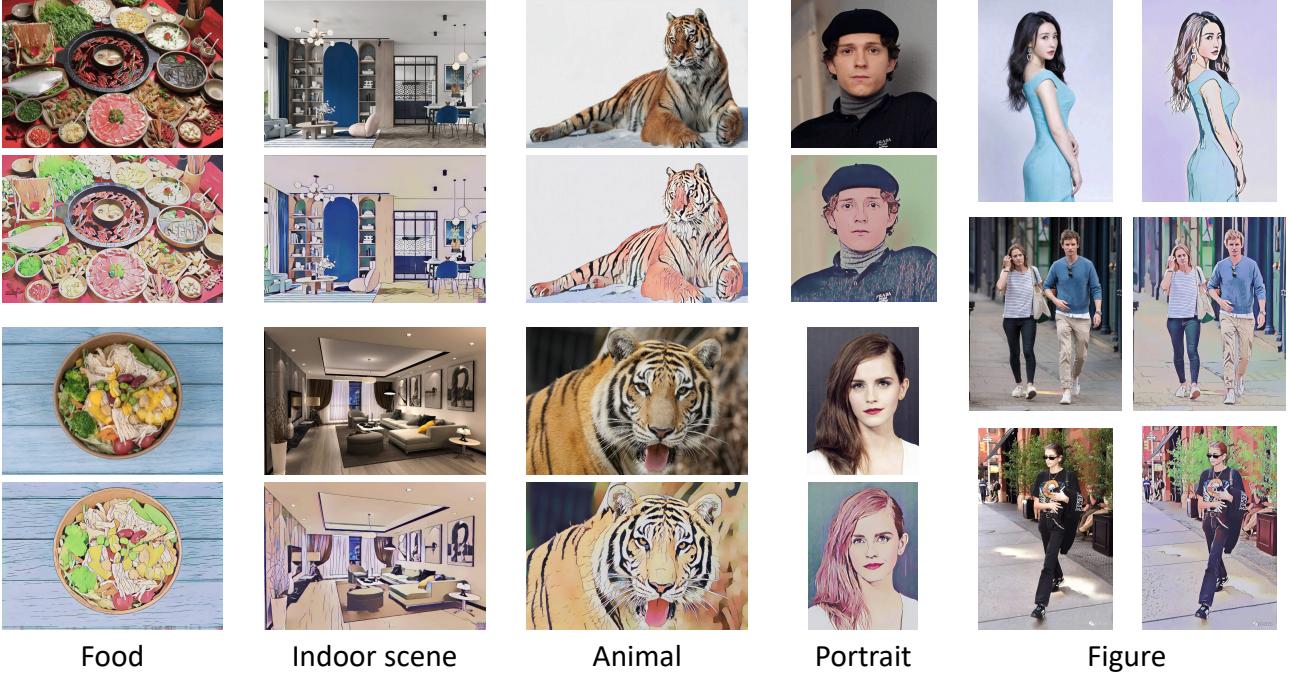


Figure 9. Qualitative cartoonization results evaluated on “Crayon Shin-chan” dataset in more scenarios including foods, indoor scenes, animals, portraits, and figures. **Better to zoom in for higher resolution.**

the corresponding results do not exhibit any cartoon styles. This indicates the fundamental role of image-level branch in maintaining the 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 loss functions in image cartoonization, results are reported in Tab. 4. 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 (Chen et al., 2018) brings very limited performance gains at the cost of a separate edge-smoothing data preparation stage. Following the idea of AnimeGAN (Chen et al., 2019) 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} (Gatys et al., 2016)

or the first-order one L_{mean_std} (Huang & Belongie, 2017). 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 (Li et al., 2017a), 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, and thus bypass the need of prior data pre-processing stage and additional style transfer losses.

5. Conclusion

This paper proposes a simple method for image cartoonization. We supplement the normal image-level adversarial learning with a patch-level one which is adaptively constrained on cartoon-texture-salient local regions via our proposed CTSS module. By incorporating such texture saliency adaptive attention to adversarial learning, our method can transfer noticeably more abstract and vivid cartoon style, with end-to-end architecture and only basic adversarial loss.

6. Acknowledgments

This work has been partially supported by grants from: National Natural Science Foundation of China (No.12071458, 71731009)

References

- Alsaifi, A., Rustagi, R., Thomas, M. M., Forbes, A. G., et al. Image denoising using a generative adversarial network. In *2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT)*, pp. 126–132. IEEE, 2019.
- Chang, H., Lu, J., Yu, F., and Finkelstein, A. Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 40–48, 2018.
- Chen, D., Yuan, L., Liao, J., Yu, N., and Hua, G. Stylebank: An explicit representation for neural image style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1897–1906, 2017.
- Chen, J., Liu, G., and Chen, X. Animegan: A novel lightweight gan for photo animation. In *International Symposium on Intelligence Computation and Applications*, pp. 242–256. Springer, 2019.
- Chen, Y., Lai, Y.-K., and Liu, Y.-J. Cartoongan: Generative adversarial networks for photo cartoonization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9465–9474, 2018.
- Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8789–8797, 2018.
- DeCarlo, D. and Santella, A. Stylization and abstraction of photographs. *ACM transactions on graphics (TOG)*, 21(3):769–776, 2002.
- Dumoulin, V., Shlens, J., and Kudlur, M. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016.
- Gao, X., Tian, Y., and Qi, Z. Rpd-gan: Learning to draw realistic paintings with generative adversarial network. *IEEE Transactions on Image Processing*, 29:8706–8720, 2020.
- Gatys, L. A., Ecker, A. S., and Bethge, M. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2414–2423, 2016.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- He, B., Gao, F., Ma, D., Shi, B., and Duan, L.-Y. Chipgan: A generative adversarial network for chinese ink wash painting style transfer. In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 1172–1180, 2018.
- He, K., Sun, J., and Tang, X. Guided image filtering. In *European conference on computer vision*, pp. 1–14. Springer, 2010.
- He, Z., Zuo, W., Kan, M., Shan, S., and Chen, X. Attgan: Facial attribute editing by only changing what you want. *IEEE transactions on image processing*, 28(11):5464–5478, 2019.
- Huang, X. and Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510, 2017.
- Huang, X., Liu, M.-Y., Belongie, S., and Kautz, J. Multi-modal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 172–189, 2018.
- Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.
- Jiang, Y., Lian, Z., Tang, Y., and Xiao, J. Dcfont: an end-to-end deep chinese font generation system. In *SIGGRAPH Asia 2017 Technical Briefs*, pp. 1–4. 2017.
- Johnson, J., Alahi, A., and Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.
- Kang, H., Lee, S., and Chui, C. K. Flow-based image abstraction. *IEEE transactions on visualization and computer graphics*, 15(1):62–76, 2008.
- Kolkin, N., Salavon, J., and Shakhnarovich, G. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10051–10060, 2019.
- Kyprianidis, J. E. and Döllner, J. Image abstraction by structure adaptive filtering. In *TPCG*, pp. 51–58, 2008.
- Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., 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*, pp. 4681–4690, 2017.

- Lee, H.-Y., Tseng, H.-Y., Huang, J.-B., Singh, M., and Yang, M.-H. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 35–51, 2018.
- Li, C. and Wand, M. Combining markov random fields and convolutional neural networks for image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2479–2486, 2016.
- Li, S., Xu, X., Nie, L., and Chua, T.-S. Laplacian-steered neural style transfer. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1716–1724, 2017a.
- Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., and Yang, M.-H. Universal style transfer via feature transforms. In *NIPS*, 2017b.
- Li, Y., Wang, N., Liu, J., and Hou, X. Demystifying neural style transfer. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 2230–2236, 2017c.
- Liu, M.-Y., Breuel, T., and Kautz, J. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pp. 700–708, 2017.
- Liu, M.-Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., and Kautz, J. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10551–10560, 2019.
- Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802, 2017.
- Mechrez, R., Talmi, I., and Zelnik-Manor, L. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 768–783, 2018.
- Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- Tian, Y. zi2zi: Master chinese calligraphy with conditional adversarial networks, 2017. *Retrieved Jun, 3:2017*, 2017.
- Ulyanov, D., Lebedev, V., Vedaldi, A., and Lempitsky, V. S. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, volume 1, pp. 4, 2016.
- Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8798–8807, 2018.
- Wang, X. and Yu, J. Learning to cartoonize using white-box cartoon representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8090–8099, 2020.
- Winnemöller, H., Olsen, S. C., and Gooch, B. Real-time video abstraction. *ACM Transactions On Graphics (TOG)*, 25(3):1221–1226, 2006.
- Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

A. User Study

Considering the subjective nature of style transfer problem, we conduct user study to evaluate our model against related methods. Based on our collected 300 real-world-scene high-resolution photos, we generated cartoonized results with different methods. First, we showed participants a content image, i.e., a real photo. Second, we showed them two cartoonization results generated by our method and a random contrast method. Finally, we asked the participants which result has better cartoon effect. We separately repeated the above process for “The Wind Rises”, “Dragon Ball”, and “Crayon Shin-chan” datasets, respectively. For results of each dataset, we collected 2400 votes from 40 participants and present the voting results in Fig. 11, which shows percentage of preference of our method against related methods on different datasets. Overall, our method gained the most user preference votes, indicating the superiority of our model from subjective perspective.

Besides, we sample 40 high-resolution photos and evaluate corresponding cartoonized results trained over “The Wind Rises” dataset using CartoonGAN, AnimeGAN, WhiteBox, and our method. Then, we allow 20 participants to score them with 1-10 ratings from three dimensions: (1) style saliency, (2) abstraction degree, (3) content integrity. Finally, we calculate the average score across all samples over each dimension. Results reported in Fig. 10 indicate that our method has obvious advantages in presenting cartoon abstraction and vividness than related advanced methods. On the other hand, it reflects that despite significant superiority in style saliency, our method compromises some content integrity, due to the inherent content-style trade-off in style transfer related problem. We will focus on improving this in future work.

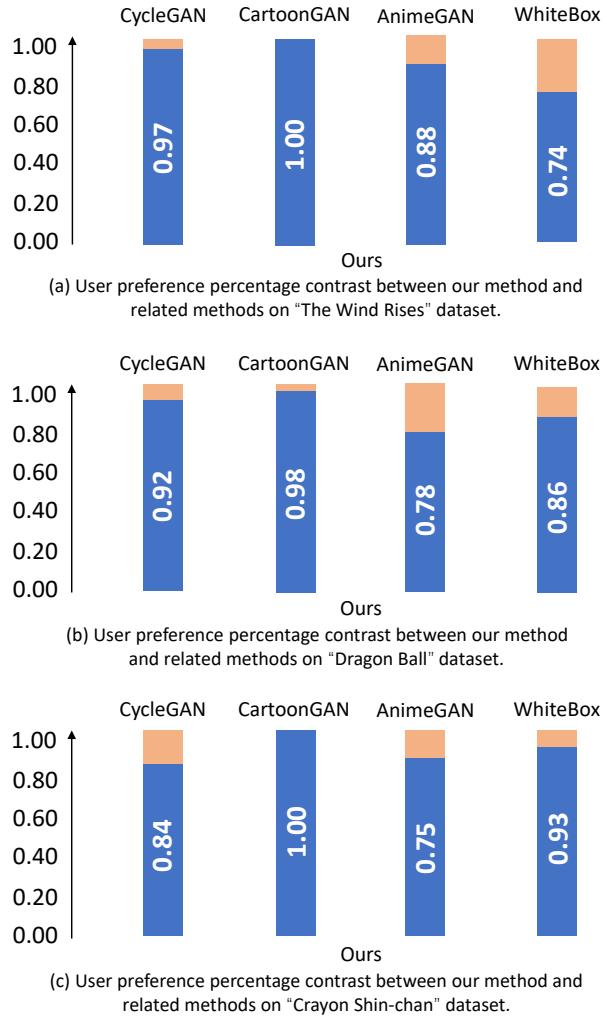


Figure 11. Percentage of user preference voting of our method against related methods on different cartoon datasets.

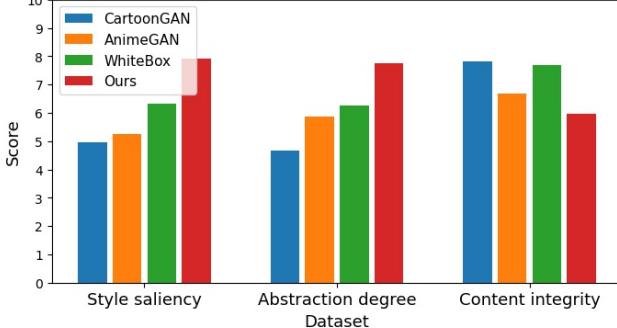


Figure 10. User preference scores of cartoonization results of different methods with respect to “style saliency”, “abstraction degree”, and “content integrity”.

B. Network Details

The generator G follows an autoencoder structure where we use four residual blocks to bridge a downsampling path and an upsampling path. Both the image-level discriminator D_{img} and the patch-level discriminator D_{patch} adopt the PatchGAN (Isola et al., 2017) structure. Notations and network details are listed in Fig. 12.

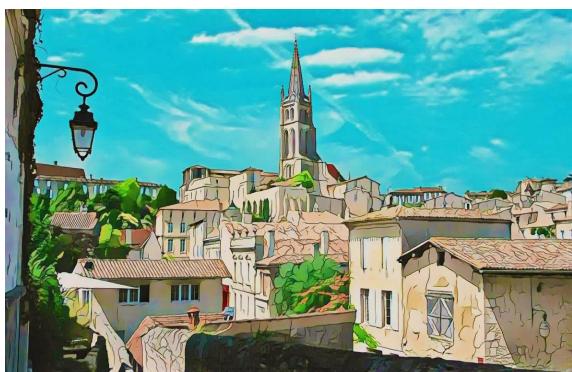
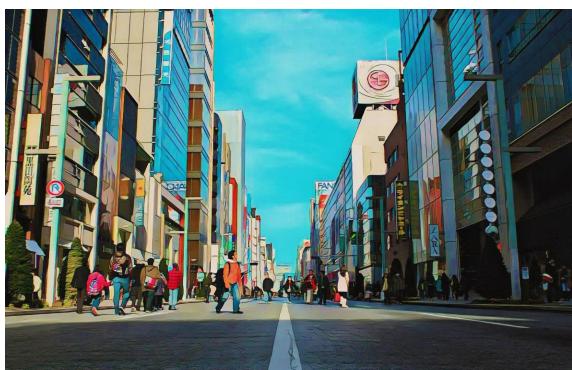
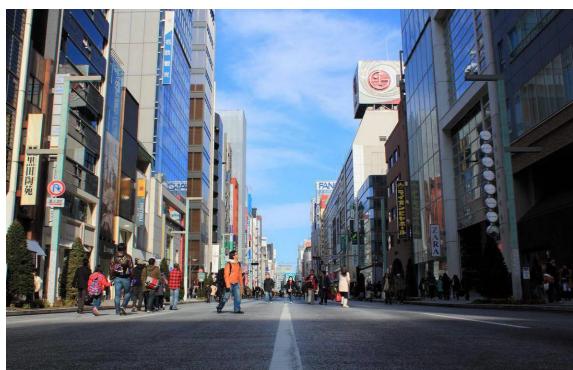
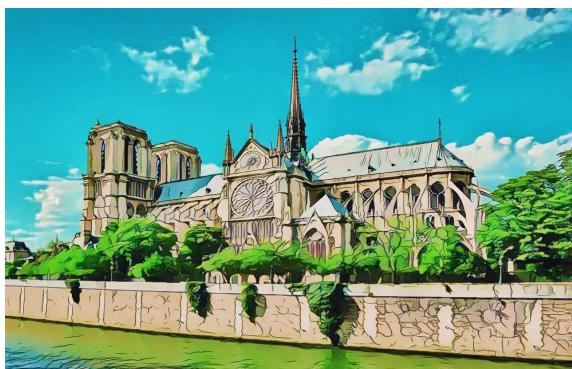
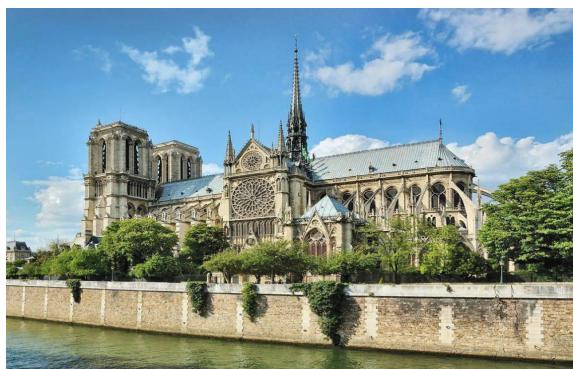
C. More Qualitative Results

We append more high-resolution cartoonization results of our model evaluated on different datasets in this section. For all the produced results, the typical cartoon styles are sufficiently transferred.

Notations			
LN: layer normalization		lRelu: 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→lRelu			
DConv_n(N)k(K)s(S): Depthwise convolutional layer with N filters, K×K kernel size, and stride S			
Generator (G)		Image-level discriminator (D_{img})	
Layers	Shape	Layers	
Input	256x256x3	Input	256x256x3
Conv_n(32)k(7)s(1), LN, lRelu	256x256x32	Conv_n(32)k(3)s(1), LN, lRelu	256x256x32
Conv_n(64)k(3)s(2), LN, lRelu	128x128x64	Conv_n(64)k(3)s(2), LN, lRelu	128x128x64
Conv_n(64)k(3)s(1), LN, lRelu	128x128x64	Conv_n(128)k(3)s(2), LN, lRelu	64x64x128
Conv_n(128)k(3)s(2), LN, lRelu	64x64x128	Conv_n(128)k(3)s(2), LN, lRelu	32x32x256
Conv_n(256)k(3)s(1), LN, lRelu	64x64x256	Conv_n(1)k(3)s(1)	32x32x1
ResBlock	64x64x256	Patch-level discriminator (D_{patch})	
ResBlock	64x64x256	Layers	Shape
ResBlock	64x64x256	Input	96x96x1
ResBlock	64x64x256	Conv_n(16)k(3)s(1), LN, lRelu	96x96x16
Conv_n(128)k(3)s(1), LN, lRelu	64x64x128	Conv_n(32)k(3)s(2), LN, lRelu	48x48x32
Upsample_n(128)	128x128x128	Conv_n(64)k(3)s(2), LN, lRelu	24x24x64
Conv_n(128)k(3)s(1), LN, lRelu	128x128x128	Conv_n(128)k(3)s(2), LN, lRelu	12x12x128
Upsample_n(64)	256x256x64	Conv_n(1)k(3)s(1)	12x12x1
Conv_n(64)k(3)s(1), LN, lRelu	256x256x64	ResBlock	
Conv_n(64)k(3)s(1), LN, lRelu	256x256x64	Layers	Shape
Conv_n(32)k(7)s(1), LN, lRelu	256x256x32	Input	64x64x256
Conv_n(3)k(1)s(1), Tanh	256x256x3	Conv_n(128)k(1)s(1), LN, lRelu	64x64x128
		DConv_n(128)k(3)s(1), LN, lRelu	64x64x128
		Conv_n(256)k(1)s(1), LN	64x64x256
		Add input	64x64x256

Figure 12. Network details of the generator G , the image-level discriminator D_{img} , and the patch-level discriminator D_{patch} .

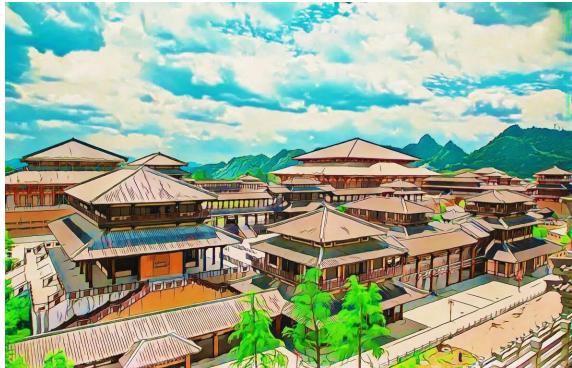
More cartoonization results on “The Wind Rises” dataset



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

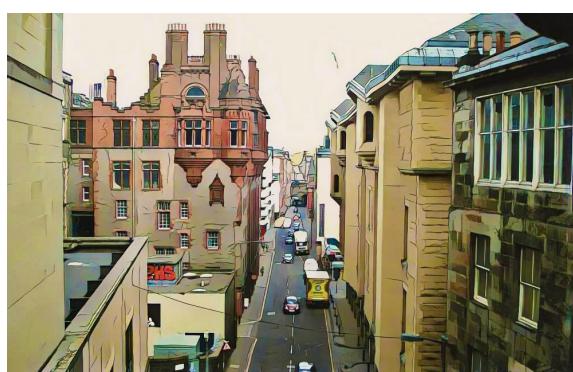
More cartoonization results on “The Wind Rises” dataset



Input Photo

Cartoonized Results

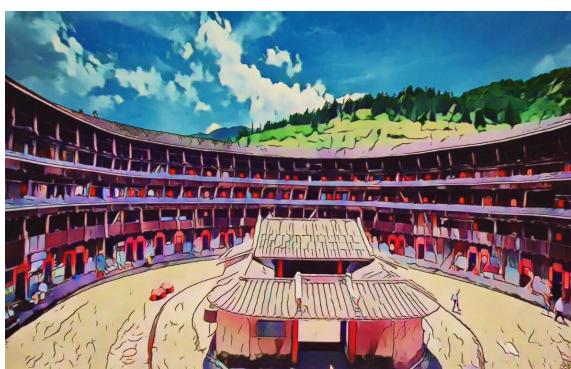
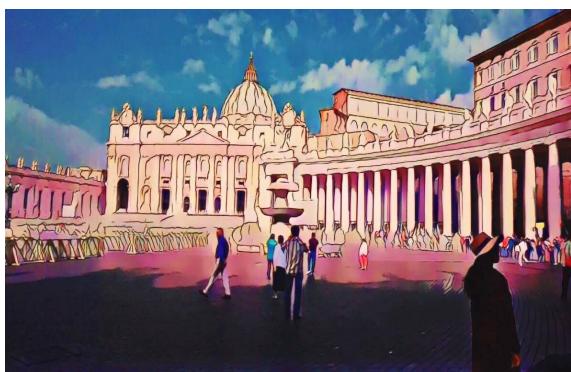
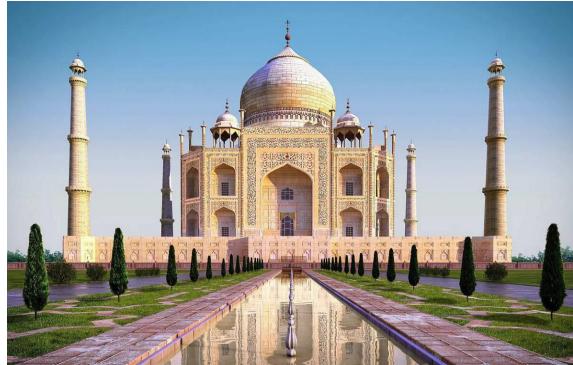
More cartoonization results on “The Wind Rises” dataset



Input Photo

Cartoonized Results

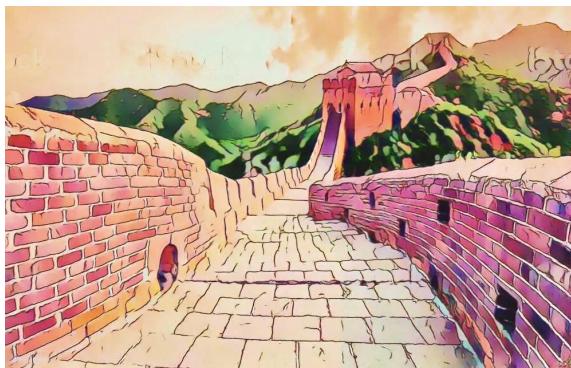
More cartoonization results on “Dragon Ball” dataset



Input Photo

Cartoonized Results

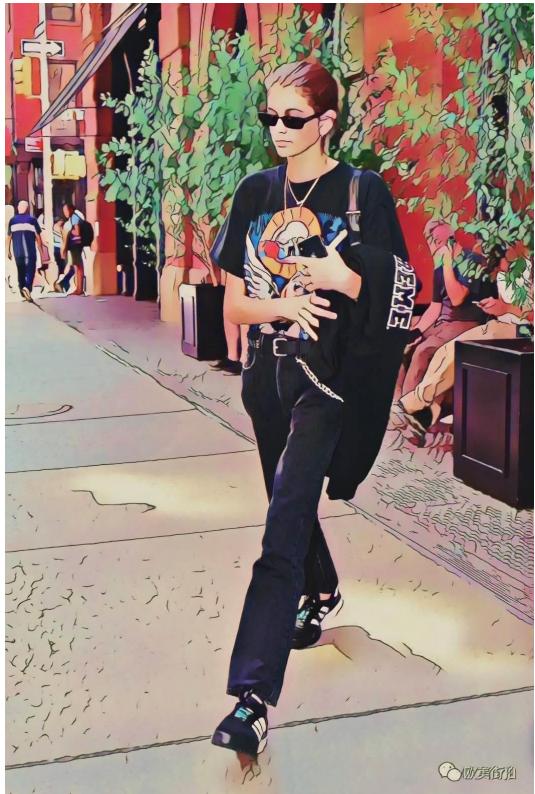
More cartoonization results on “Dragon Ball” dataset



Input Photo

Cartoonized Results

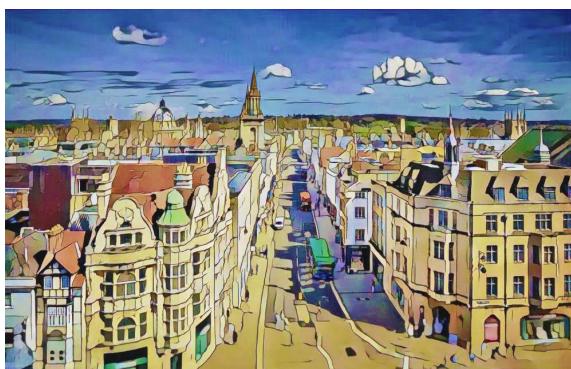
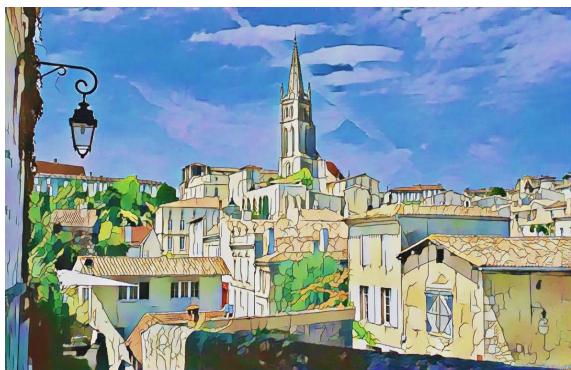
More cartoonization results on “Dragon Ball” dataset



Input Photo

Cartoonized Results

More cartoonization results on “Crayon Shin-chan” dataset



Input Photo

Cartoonized Results

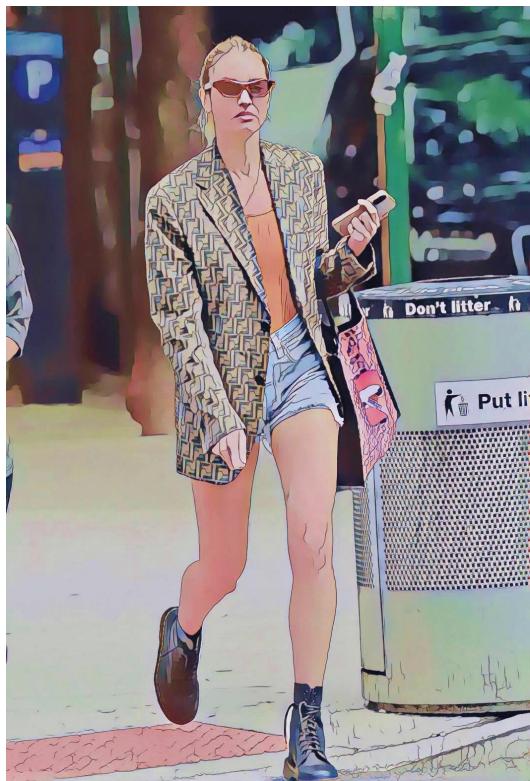
More cartoonization results on “Crayon Shin-chan” dataset



Input Photo

Cartoonized Results

More cartoonization results on “Crayon Shin-chan” dataset



Input Photo

Cartoonized Results