Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.039182
Article

ECGAN: Translate Real World to Cartoon Style Using Enhanced Cartoon Generative Adversarial Network

Yixin Tang*

Department of Cooperative Course of Performance, Film & Animation, Sejong University, Seoul, 05006, Korea

*Corresponding Author: Yixin Tang. Email: 21170957@sju.ac.kr

Received: 13 January 2023; Accepted: 10 April 2023; Published: 09 June 2023

Abstract: Visual illustration transformation from real-world to cartoon images is one of the famous and challenging tasks in computer vision. Image-to-image translation from real-world to cartoon domains poses issues such as a lack of paired training samples, lack of good image translation, low feature extraction from the previous domain images, and lack of high-quality image translation from the traditional generator algorithms. To solve the above-mentioned issues, paired independent model, high-quality dataset, Bayesian-based feature extractor, and an improved generator must be proposed. In this study, we propose a high-quality dataset to reduce the effect of paired training samples on the model's performance. We use a Bayesian Very Deep Convolutional Network (VGG)-based feature extractor to improve the performance of the standard feature extractor because Bayesian inference regularizes weights well. The generator from the Cartoon Generative Adversarial Network (GAN) is modified by introducing a depthwise convolution layer and channel attention mechanism to improve the performance of the original generator. We have used the Fréchet inception distance (FID) score and user preference score to evaluate the performance of the model. The FID scores obtained for the generated cartoon and real-world images are 107 and 76 for the TCC style, and 137 and 57 for the Hayao style, respectively. User preference score is also calculated to evaluate the quality of generated images and our proposed model acquired a high preference score compared to other models. We achieved stunning results in producing high-quality cartoon images, demonstrating the proposed model's effectiveness in transferring style between authentic images and cartoon images.

Keywords: GAN; cartoon; style transfer; deep learning; Bayesian neural network

1 Introduction

Cartoons are widely used in education and entertainment to transform real-world activities into artistic domains. Likewise, cartoons are often made of real-world scenes through handcrafted visual transformation techniques. Cartoonization of real-world scenes involves recreating every line and shade of each color in artistic form, which may take a substantial amount of time. Therefore, several companies introduced cartoon editing software to generate cartoon images from real-world images. However, existing techniques and software do not provide high-quality image transformation. Due to the recent advances in artificial intelligence technology, specifically in deep learning techniques, researchers are utilizing deep learning techniques to extract feature information and recreate real-world scenes in the cartoon domain.

Cartoonization of real-world images can be described as image-to-image translation problems, which are currently addressed by deep learning techniques. Deep learning is a subfield of machine learning which uses layers of the artificial neural network to extract high-quality features from a given data, in our case, an image. Recently, Deep learning models such as Convolutional Neural Networks (CNNs) have been used to perform the image-to-image translation and acquire good results than the traditional methods. CNN-based methods use the correlation of the features between different style images to encode visual style. However, traditional CNN-based approaches did not get satisfactory results; therefore, researchers are utilizing Generative adversarial networks (GANs) based on CNN to improve the image-to-image translation performance.

The image-to-image translation techniques based on GANs are examples of offline learning [1]. To trick the discriminator, the generator automatically learns the target style. While using a two-player minimax game, the discriminator tries to discern between actual and fraudulent images. The local style patterns may be learned more effectively using the adversarial loss instead of the Gram-matrix-based style loss [2]. However, the existing GANs-based solution still poses some issues, such as (1) lack of paired training samples, (2) low feature extraction from the previous domain images, (3) lack of high-quality images translation from the traditional generator algorithms, which need to be solved to improve the performance of the image-to-image translations.

To solve the above-mentioned issues, this study proposes an enhanced cartoon generative adversarial network (ECGAN), which can produce high-quality cartoon images from real-world images. A novel generator is proposed to improve the feature translation from the real-world input image and generate carton style image. The proposed method does not require paired samples to transform real-world images into cartoon images. In addition, Bayes by back propagation method is introduced in the feature extraction part to improve the model's regularization statistically. The main contribution of this paper can be seen as follows,

1. A new dataset is proposed, composed of two different cartoon styles, including the traditional Chinese cartoon style (TCC) and the modern Chinese cartoon style (MCC). Each cartoon style has 1800 images and 1800 smooth-edge images.
2. An enhanced depthwise residual-based generator is proposed to generate better fake images for the discriminator to learn efficiently.
3. Bayes by backpropagation-based Very Deep Convolutional Network (VGG) feature extractor is introduced to improve the feature extraction and regularize the proposed model effectively.

2 Related Work

This section discusses the recent development in style transfers using the neural network-based style transfer method and GAN-based style transfer methods.

2.1 Neural Network-Based Style Transfer

Deep learning-based methods are frequently being used in image-processing tasks with considerable success. Due to the success of the deep learning-based method in other fields, researchers proposed deep learning-based methods to transfer styles. For the first time, VGG based deep learning model was used to extract feature information from an image and transform it into an artistic style [3]. However, their proposed method requires a lot of time to transfer real-world images to artistic-type images. Kalischek et al. proposed a central moment discrepancy-based optimization method to solve the partially aligned feature distribution that occurred in the traditional methods [4]. Their proposed method also inherits the problem of high time consumption and cannot be used in real-time applications. Virtusio et al. proposed a neural style palette method to enable multimodal style imager transfer from one image and user interaction during the transformation [5]. Their method is the first in the interactive style transfer field using the neural network style method. Chandran et al. proposed adaptive convolution to represent the local geometric structure in style images, which were often ignored in traditional methods [6]. An et al. reduce the content leak during the style transfer by proposing the ArtFlow method, which uses forward and backward inferences to preserve content during transfer [7]. Xu et al. proposed self-supervised space-time convolutional neural network to transfer video style [8]. They also proposed style-coherence loss to improve the performance of the model. Luan et al. reduced distortions and content mismatches by modifying the objective function and introducing optional guidance inspired by the semantic segmentation method [9]. Liu et al. proposed VGG-19-based style transfer algorithms by merging content and style images to improve the model's performance [10]. However, their proposed method is time-consuming and cannot be used in real-time applications. Kumar et al. proposed a forgery detection technique using multiple light source directions to identify image manipulations based on elevation angle α obtained from a source of light and surface normal [11]. The method is effective in detecting forgeries in both outdoor and indoor images but relies on certain assumptions about surface properties and illumination parameters.

2.2 GAN-Based Style Transfer

A generative adversarial network is mainly used to produce realistic images using the adversarial training procedure of two networks, such as a generator and discriminator. Xu et al. proposed multiple dynamic ResBlock and style collection conditional discrimination-based GAN to provide arbitrary and collection style transfer [12]. Chen et al. proposed a Dual style-learning network for artistic style transfer (AST) framework to learn holistic and specific style images [13]. They also proposed a style control block to modify the image style to give the user more control in the transfer process. Chen et al. proposed GAN-based cartoon style transfer and also proposed semantic and adversarial loss specific to cartoon style transfer [14]. Their proposed method achieved good performance in the cartoonization of real-world images. Dong et al. proposed a novel loss function-based CartoonGAN to improve the cartoonization of real-world images by smoothing the surface of the generated cartoon images [15]. Chen et al. proposed gated transformer-based GAN and auto-encoder reconstruction loss to enable multiple style transformations from a single image [16]. Li et al. proposed AniGAN to generate anime faces by introducing a novel generator, discriminator, and normalization function [17]. Their method achieved state-of-the-art performance compared to

other existing methods. Shu et al. proposed multi-style GAN by introducing hierarchical semantic loss, edge-promoting adversarial loss, and style loss to achieve multiple styles from a single image [18]. Zhang et al. proposed Convolutional Block Attention Generative Adversarial Networks (CBA-GAN) model to adaptively optimize features, adjust proportions of edge, texture, and smoothness, and can handle shadows to produce high-quality cartoon images from real photos [19]. Experimental results demonstrate better performance than the three existing methods on various image types, and the method is easily transferable to video cartoonization using the provided style image dataset. However, the image quality of the proposed method is more similar to real-world images compared to cartoon images. Feng et al. proposed a multi-scale training structure with a progressive growth generator and novel Cross-CBAM mechanism that improves unsupervised image-to-image translation tasks, achieving superior results to state-of-the-art algorithms on multiple datasets [20]. However, the proposed model faces generalization and an incomplete parallel strategy problem, which degrades the image quality of the generated images. Preeti et al. utilized pre-trained GANs for implementing deepfake and discuss the main techniques for manipulation and detection of deepfake [21]. The study includes comparative analyses of proposed GAN with other existing models and future trends in the field.

3 ECGAN

3.1 Dataset Preparation

Four different styles of cartoon images, including traditional and modern, Shinkai, and Hayao were acquired for the proposed study. Traditional Chinese Cartoon Style (TCC) is a style of animation that is rooted in traditional Chinese art forms, such as Chinese painting, calligraphy, and folk art. This style emphasizes a strong sense of Chinese culture and often features motifs and themes from Chinese mythology and history. The animation in TCC is typically done using a 2D hand-drawn approach, and the color palette used is often muted and subdued. Modern Chinese Cartoon Style (MCC), on the other hand, is a more contemporary style of animation that has emerged in China in recent years. MCC often features vibrant and bold colors and is characterized by a more exaggerated and dynamic animation style. This style is heavily influenced by the anime and manga styles of Japan, as well as the Western cartoon styles. TCC and MCC are two distinct styles of animation that reflect different eras and cultural influences. While TCC seeks to preserve and promote traditional Chinese culture through animation, MCC seeks to create a more contemporary and globalized form of animation that can compete on a global scale. This two-cartoon style dataset is proposed first time in this study. On the other hand, Shinkai and Hayao are two different animation styles that have emerged in Japan, both of which are well-known and highly regarded in the world of anime.

Cartoon images were collected from the following cartoons: The legend of a sealed book, lotus lantern, white snake, Ne Zha, New Gods: Yang Jian, Green Snake, Spirited Away, and Paprika. A total number of 60000 images with 1920×1080 resolution was collected by extracting every frame of the selected cartoons. Then we selected 10000 images based on the quality of the images to train our proposed model, where 1800 were traditional Chinese cartoons, 1800 were modern Chinese cartoons, 3200 were Shinkai cartoons and 3200 were Hayao cartoon-style images. Moreover, we also acquired 7343 real-world images for training and testing the proposed model. All images were resized to 256×256 before training the proposed model.

3.2 Data Preprocessing

The acquired images were randomly cropped with a resolution of 512×512 to remove redundant features from the images. Then the images were resized to 256×256, and 10000 images were selected

to train the proposed model. Fig. 1 shows the image sample of our proposed traditional and modern style image dataset.

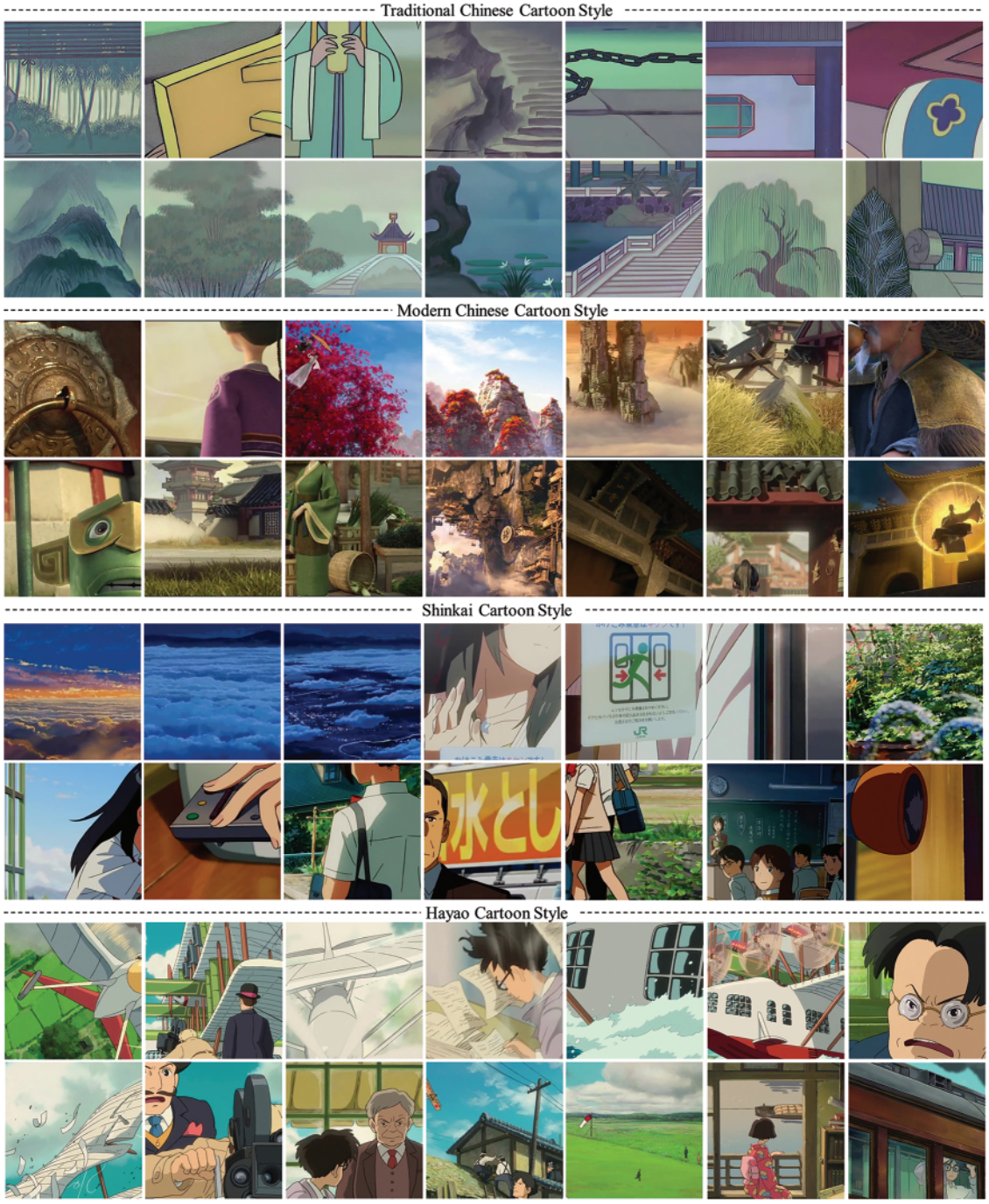

Figure 1: A bunch of samples of the implemented dataset which includes traditional Chinese cartoon and modern Chinese cartoon style, Shinkai cartoon style, and Hayao cartoon style

3.3 ECGAN Architecture

This paper proposed a robust and effective GAN to transform real-world images into cartoon images. The proposed model includes two parts: a generator network and a discriminator network. The generator is responsible for transforming real-world images into cartoon images, and the discriminator is responsible for deciding whether the generated images are real or fake. Fig. 2 shows the overall architecture of the proposed model.

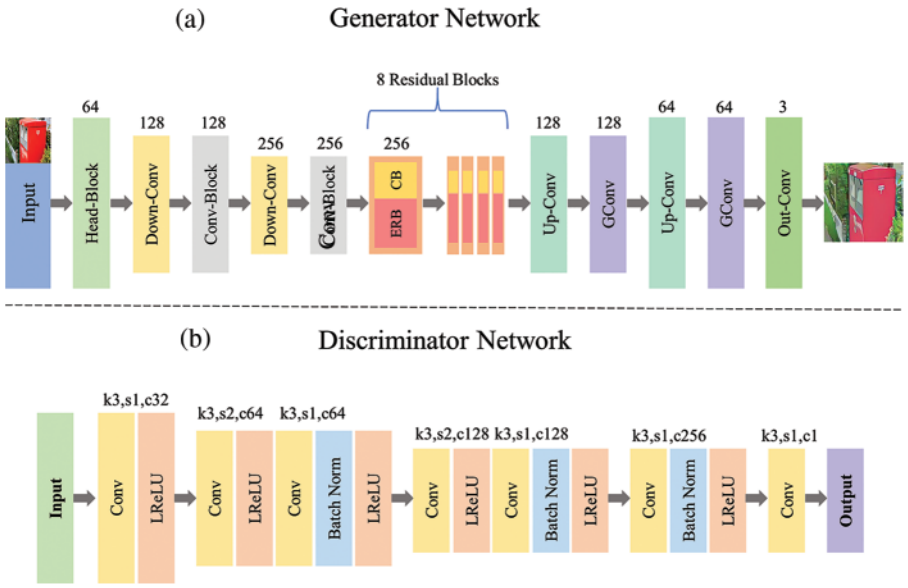

Figure 2: The overall structure of the proposed ECGAN. It has two parts: generator network and discriminator network

Fig. 2a is the architecture of the proposed generator, which is an encoder-decoder network. The generator network consists of a head block, downsampling followed by convolution (Down-Conv) block, Convolution Block (Conv-Block), Enhanced Residual Block (ERB), Convolution block with batch normalization (CB), upsampling followed by convolution (Up-Conv), and Group convolution (GConv) block. The last output layer consists of a 7×7 convolution layer with a tanh activation function to generate the cartoon image.

The discriminator network of the proposed model consists of a standard convolution layer, Leaky rectified linear unit (ReLU) activation function and Batch normalization. The detailed structure of the discriminator network can be seen in Fig. 2b.

The structure of each block used in the generator network can be seen in Fig. 3. Fig. 3a shows the head block's structure consisting of a 7×7 convolution layer followed by batch normalization and ReLU activation function. The structure of the Down-Conv can be seen in Fig. 3b. We used depthwise convolution with stride 2 followed by batch normalization and ReLU, then used standard 3×3 convolution layer followed by batch normalization and ReLU. Lastly, depthwise convolution with stride 1 followed by batch normalization and ReLU activation function is used to reduce feature information loss. The Conv-Block, which can be seen in Fig. 3c, is similar to the head block with the exception of kernel size of 3×3 is used in the normal Conv-Block.

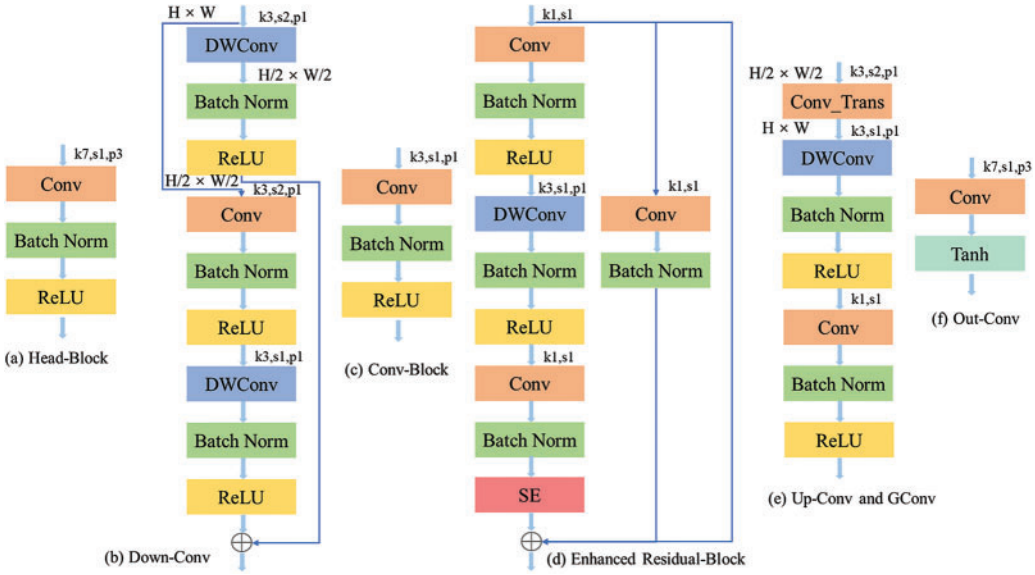

Figure 3: The detailed structure of the generator network including (a) head-block, (b) down-Conv block, (c) Conv-block, (d) enhanced residual-block, (e) up-Conv and GConv block, (f) out-Conv block

The structure of the ERB can be seen in Fig. 3d. We used standard convolution and depthwise convolution layer followed by batch normalization and ReLU activation function to extract features. Moreover, we also used the channel attention mechanism to extract more valuable feature information from the input. This study uses the Squeeze-and-Excitation (SE) attention mechanism, which is lightweight and effective [18]. We also used Conv-Block on the input along with the output of the SE attention mechanism to sum all the features information. This study uses 8 identical ERB blocks sequentially to increase the feature information of the generator. The proposed ERB block can acquire more feature information with a low computational burden compared to the standard residual block [22].

The Up-Conv and GConv consist of the Convolution transformation which is responsible for upsampling the feature maps, the 3×3 depthwise convolution layer to process the feature in groups, and the standard 1×1 convolution layer followed by batch normalization and ReLU activation function.

3.4 Loss Function

This study utilizes two loss functions to transfer style between the real world to the cartoon style, including adversarial loss and feature loss. The adversarial loss is responsible for the desired transformation of the style using the generator network. The feature loss is responsible for preserving features during the transformation of real-world images to cartoon images. The relevant equation of the combined loss of the proposed model can be seen as follows,

$$\mathcal{L}(G, D) = \mathcal{L}_{adv}(G, D) + \omega \mathcal{L}_{fea}(G, D) \tag{1}$$

where, ω represents the weight balance of the content loss. Larger weight results in preserving more features from the training, which may degrade the performance of the style transfer. Therefore, a weight balance value of 10 is used to preserve a balance between style and feature loss.

3.4.1 Adversarial Loss

The adversarial loss is used for the generator and discriminator network to indicate the style transformation quality between real-world images and cartoon images. This study utilizes adversarial loss from the CartoonGAN to improve the performance of the model [23]. The clear edge of the training data TCC and MCC is removed to generate the edge-smoothed data to train the model. The model first detects edge pixels using a standard Canny edge detector [24]; then, the images' edge regions are dilated; lastly, perform the Gaussian smoothing method on the dilated edge region to generate edge-smoothed images.

The transformation of edge-smoothed image from training can be seen in Fig. 4. The discriminator is responsible for labeling the generated image as real or fake. A well-trained discriminator can improve the performance of the model by guiding the generator to generate images effectively. The relevant adversarial loss function can be seen below,

Figure 4: Edge smoothed image is generated from the cartoon image by removing clear edges

$$\mathcal{L}_{adv}(G, D) = \mathbb{E}_{c_i} \sim S_{data}(c)[\log D(c_i)] + \mathbb{E}_{e_j} \sim S_{data}(e)[\log(1 - D(e_i))] + \mathbb{E}_{p_k} \sim S_{data}(p)[\log(1 - D(G(p_k)))] \quad (2)$$

where, $\mathbb{E}_{c_i} \sim S_{data}(c)$ is the real cartoon image, $\mathbb{E}_{e_j} \sim S_{data}(e)$ is the cartoon without clear edges, D is the discriminator, p_k is each image from image manifold P, $G(p_k)$ is the generated image.

3.4.2 Feature Loss

The Feature loss is used to indicate the preservation of the feature information of input images during style transformation between real-world images and cartoon images. This study introduces the Bayesian VGG network to attain high-level feature maps from the images. The relevant feature loss can be seen below:

$$\mathcal{L}_{fea}(G, D) = \mathbb{E}_{p_i} \sim S_{data}(p)[||BVGG_l(G(p_i)) - BVGG_l(p_i)||_1] \quad (3)$$

where, $BVGG_l$ refers to Bayesian VGG network, l refers to the feature maps of the model.

The ℓ_1 regularization method is used in the feature loss due to the huge characteristics difference between the two styles. Researchers show that ℓ_1 regularization adopts better than other regularization methods when there is a massive style difference, which can be observed especially in local regions. Layer 4 of the Bayesian VGG is used to compute the feature loss in this study.

3.5 Initialization Method

The traditional GAN-based model uses a random initialization method which may affect the convergence of the model. To overcome this issue, CartoonGAN proposed initiation phases to improve the GAN-based model's convergence. This study utilized the same initialization phase proposed by CartoonGAN, which pre-trains the generator with feature loss ($\mathcal{L}_{fea}(G, D)$) to preserve the feature during style transformation. This loss function is essentially a measure of the similarity between the feature representations of the real and generated images, as captured by the Bayesian VGG network's feature extractor at a specific layer. By minimizing this loss, the generator is encouraged to generate images that have similar high-level features as real images, which in turn can help improve the quality of generated images. Therefore, it can help reduce artifacts and improve the overall visual quality of the generated images. The learning rate was set to 0.0002 with the Adam optimizer. Moreover, the initialization epochs and feature information loss were set to 10 during training.

This study uses the Bayesian VGG-19 network in the initialization method to improve model performance and convergence. Traditional neural networks do not regard weights as randomized values. It presumes that the weights with such a real value are unknowable, and the data is interpreted as a stochastic process. Bayesian neural networks estimate model weights based on known or observed data. The Bayes Theorem enables us to define a probability across these weights in terms of observed likelihood, resulting in the posterior distribution of hyperparameters. The joint distribution of the prior beliefs of our neural networks can be defined as follows,

$$p(w|d) = \frac{p(d|w)\, p(w)}{p(d)}, \tag{4}$$

where $p(w|d)$ represents the likelihood of training data (d) for the parameter (w), $p(w)$ is the distribution of the weights, and $p(d)$ is the distribution of the data. The benefit of using Bayesian feature extractor in a GAN initialization method is that it can provide a principled way to model uncertainty in the weights and biases of the generator and discriminator networks. This allows for better regularization and can prevent overfitting of the GAN model to the training data. By incorporating Bayesian inference, the initialization method can provide more robust and accurate estimates of the parameters, which can result in improved convergence, stability, and performance of the GAN model.

The reconstruction of the image using the initialization part can be seen in Fig. 5. The images were acquired after training 10 epochs with the initialization method. Researchers showed that using the initialization method in the GAN-based model can also improve the performance of the style transfer [25].

4 Experiments

The proposed model was written in python with Pytorch deep learning library. All experiments were done on Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20 GHz processor and the NVIDIA(R) GV100GL (Tesla V100 DGXS 32GB) GPU.

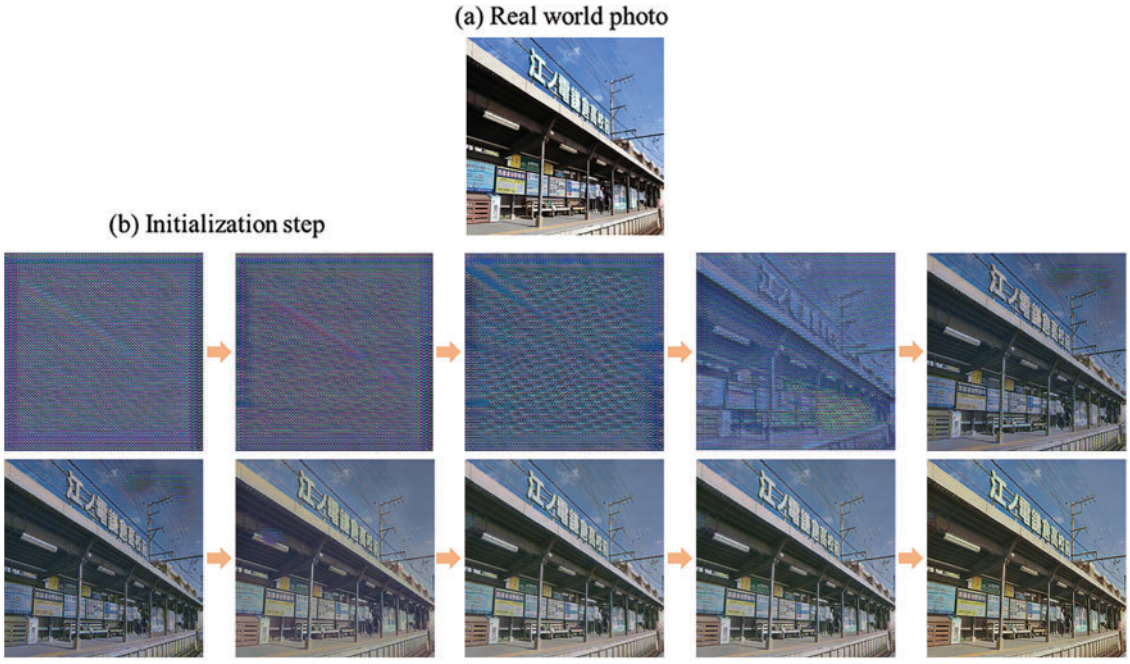

Figure 5: The step of the Initialization method over the 10 epochs for a given real-world image

Extensive experiments were done to evaluate the performance of the proposed model. A detailed description of the training data is presented in Section 4.1, qualitative experiments to compare with state-of-the-art models are presented in Section 4.2, and quantitative experiments are presented in Section 4.3.

4.1 Qualitative Experiments

4.1.1 Performance Evaluation

The proposed model was trained on the cartoon dataset to transform the real image into a cartoon to evaluate the performance of the proposed model. Fig. 6 shows the different cartoon-style images generated by the proposed model from the real-world images. It can be seen that the proposed model can generate high-quality cartoon images with smooth surfaces. The generated images' colors are similar to the cartoon images of actual movies. The proposed feature extraction block with the Bayes backpropagation method performs well in extracting features that can be seen from the generated cartoon images. It can also be seen that the initialization strategy used helps the model to train with more stability.

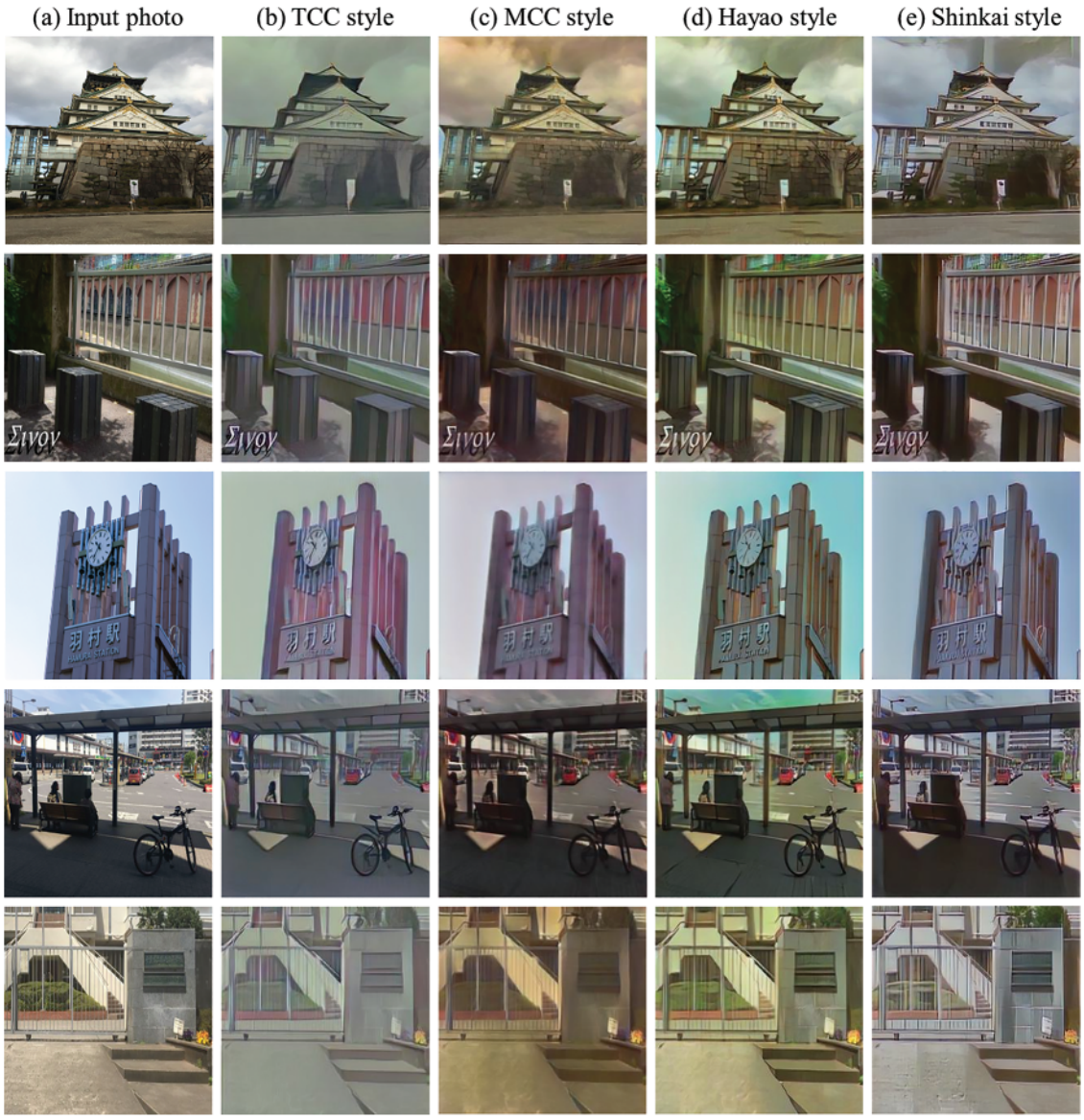

Figure 6: The results of the different style images generated by our proposed model. (a) Real-world image, (b) TCC style, (b) MCC style, (c) Hayao style, (d) Shinkai style

We further evaluate the performance of our proposed model by training the model with the initialization strategy and without the initialization strategy. Fig. 7 shows the initialization strategy experiments, where six images from six different epochs were collected to perform an ablation study by removing the initialization strategy. The upper part of Fig. 7 represents the results with the initialization method, and the lower part represents the results without the initialization method. It can be seen that the model generates good cartoonized images with the initialization method faster compared to without the initialization method. The model started to generate relatively good cartoon images at the 5th epoch, which can be seen in Fig. 7f, and started to generate good cartoon images after 30 epochs. This is due to the initialization method in our model, which helps the generator

generate content images before the training step, and the training step is only responsible for generating cartoonized images. On the other hand, the model performs worse without the initialization method. It can be seen that the model is not able to produce good cartoonized images even after 30 epochs. Therefore, this study uses the initialization method to train the dataset to improve the performance of the proposed model.

Figure 7: The result of the ablation experiment of our proposed initialization method. The upper part indicates that the initialization method is used, and the lower part indicates that the initialization method is not used. (a) Input real-world photos; (b)–(f) The experimental results of the first 5 epochs; (g) generated image after 30 epochs

4.1.2 Comparison with State-of-the-Art Models

The proposed model is compared with the state-of-the-art cartoonization method to evaluate the quality of the generated images. First, CycleGAN is selected to compare with the proposed model because CycleGAN is the standard artistic style transfer method. It utilizes the neural style transfer and GAN method to transfer styles for a given input. Two versions of CycleGAN were used to train the results with our proposed model, such as CycleGAN without identity loss and CycleGAN with identity loss. The identity loss is responsible for generating cartoon images with more content preservation from the input image. The CycleGAN was trained in the TensorFlow platform with the default configuration. Second, CartoonGAN is selected to compare the results with the proposed model as it is the state-of-the-art model for cartoonized style transfer. CartoonGAN was trained in the Pytorch platform with the default configuration provided by the author.

Fig. 8 shows the implemented model's experiment results and the proposed model. Two style images, including TCC and Hayao, were used to compare the trained model. CycleGAN without identity loss performs worse in generating cartoonized images for both styles, which can be seen in Fig. 8b. However, the CycleGAN with identity loss performs better than its predecessor version by preserving the more semantic content of the generated carton images, but the stylization is still far from satisfactory. In addition, CartoonGAN performs better than CycleGAN in terms of stylization and preserving the semantic content. However, the generated images still lack semantic content level preservation and color accuracy to reach a satisfactory level. On the other hand, it can be seen that our proposed model performs well compared to the other models in preserving semantic content, stylization, and color accuracy, which can be seen in Fig. 8e. Therefore, our proposed model can be used to produce high-quality cartoon images from real-world images.

4.2 Quantitative Experiments

This study uses Frechet Inception Distance (FID) to evaluate the performance of the proposed model [26]. FID is responsible for extracting features using a pre-trained inception model to calculate the distance between image distributions. The lower FID score indicates the closer to the reference images. In this study, we calculated the FID score for cartoon and content images to evaluate the performance of the proposed model. We also compared the FID score with the other two style transfer models, CycleGAN and CartoonGAN. This study used two cartoon styles, such as TCC and Hayao. The PyTorch platform was used to calculate the FID score per the author's instruction. First, the FID score for generated and cartoon images were calculated to compare the results in terms of cartoonization. Second, The FID score for generated and real-world images was calculated to compare the results in terms of semantic feature information preservation. Tables 1 and 2 show the results of the experiments for two different styles. It can be seen from the two tables that CycleGAN performs worse compared to CartoonGAN and our proposed model, which was obvious as per Fig. 8. CycleGAN generated cartoon images with distortions and less semantic feature information. The FID score of real-world images was very high due to the lack of accurate feature extraction from real-world images. On the other hand, our proposed model outperformed other models for TCC style by acquiring less FID score compared to CycleGAN and CartoonGAN. The difference between our proposed model and other models in terms of FID to cartoon is significantly better, demonstrating that our proposed can generate better-cartoonized images. However, the difference between our proposed model and CartoonGAN in terms of FID to the real world is better by a close margin. This demonstrates that CartoonGAN can also preserve better semantic information during style transfer. Although CartoonGAN and our proposed model use the VGG architecture, this study introduced Bayesian

CNN to the VGG network to acquire better performance. Nevertheless, it can be said that the proposed model can effectively transfer real-world images to cartoon images with high quality.

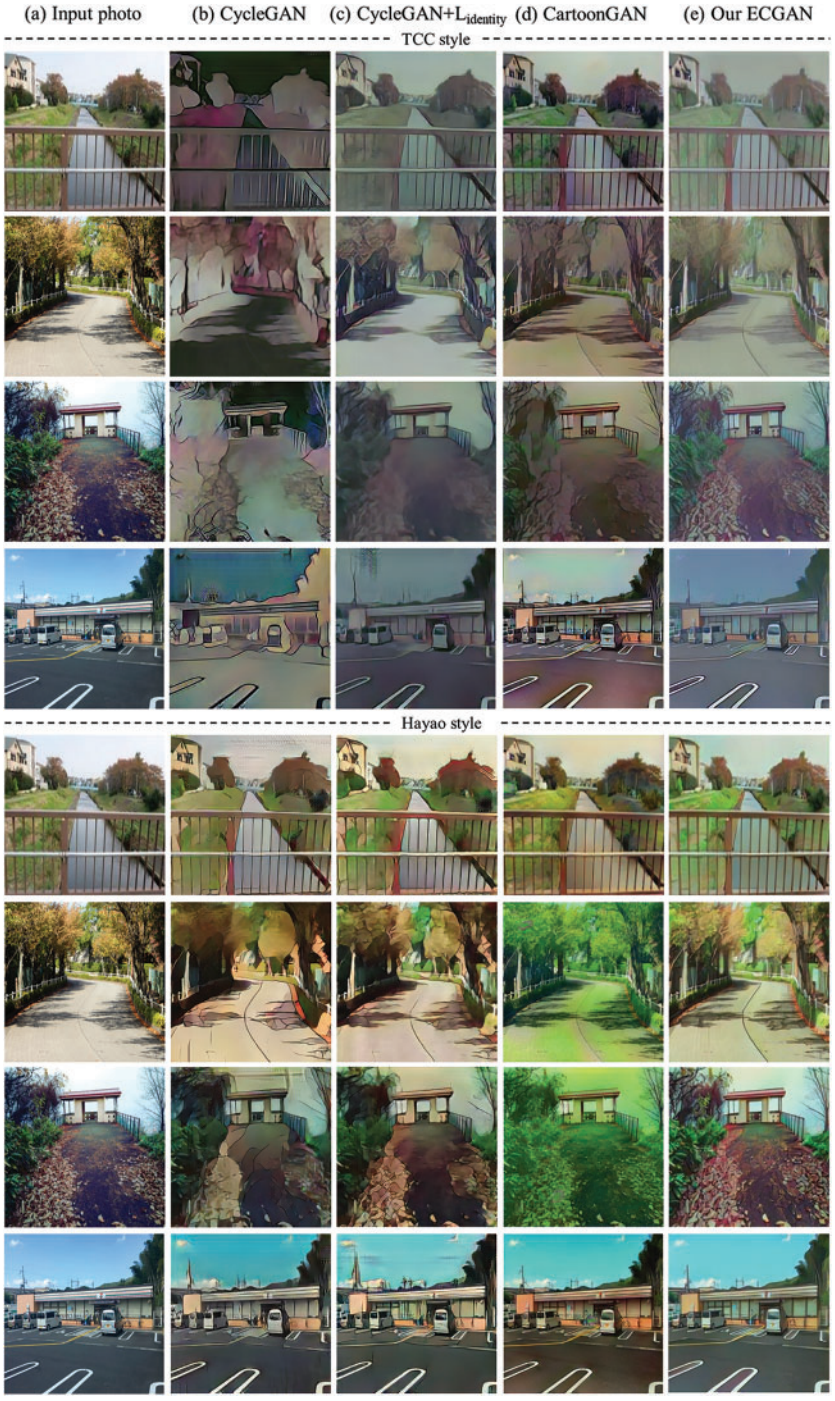

Figure 8: The comparison of different models to evaluate the quality of the generated images for the TCC style (top) and Hayao style (down)

Table 1: The results of the FID score of different models for the TCC style

Model	FID to cartoon images	FID to real-world images
Real-world images	213	N/A
CycleGAN [1]	142	212
CycleGAN + $\mathcal{L}_{\text{identity}}$ [1]	149	172
CartoonGAN [24]	175	81
ECGAN	**107**	**76**

Table 2: The results of the FID score of different models for the Hayao style

Model	FID to cartoon images	FID to real-world images
Real-world images	173	N/A
CycleGAN [1]	144	141
CycleGAN + $\mathcal{L}_{\text{identity}}$ [1]	148	158
CartoonGAN [24]	162	59
ECGAN	**137**	**57**

Table 3 shows the results of cartoon image generation time for four different models. The time is measured in seconds and represents the average time it takes for each model to generate a single cartoon image. It can be seen from the table that the average generation time for CycleGAN is 0.229 s and the generation time for CycleGAN + $\mathcal{L}_{\text{identity}}$ is 0.236 s. On the other hand, the average generation time for CartoonGAN and Proposed ECGAN is 0.034 and 0.025 s, respectively. Overall, these results suggest that the ECGAN and CartoonGAN models are faster at generating cartoon images compared to CycleGAN and CycleGAN + $\mathcal{L}_{\text{identity}}$. This could be since ECGAN and CartoonGAN have more efficient architectures and optimized training techniques.

Table 3: The results of cartoon image generation time of different models

Model	Time/s
CycleGAN [1]	0.229
CycleGAN +$\mathcal{L}_{\text{identity}}$ [1]	0.236
CartoonGAN [24]	0.034
ECGAN	**0.025**

We conduct user research trials to demonstrate the effectiveness of our proposed model in more detail. Similarly, to draw realistic paintings with a generative adversarial network (RPD-GAN) [27], this experiment takes place in a controlled lab environment. In this work, we use four models—CycleGAN, CycleGAN with Identity loss, CartoonGAN, and our proposed model—to transfer the real-world input images into cartoon-style images and assign them a number. We have considered four parameters used to define the user preference score for our proposed model which include: Style Strength, Content preservation, Quality, and Diversity. The style strength represents the degree to

which the target style of the reference image is applied to the output image. Content Preservation represents how well the output image preserves the content of the input image. Quality represents the measurement of the overall quality of the output image, based on factors such as sharpness, color balance, and visual artifacts. Diversity represents the diversity of the output images generated by the model. A higher diversity score indicates that the model is capable of producing a wider range of output styles. Then the mean of these parameters' scores was calculated to evaluate the performance of the model. To ensure that all participants in the user research have an equal experience, we randomly assign numbers to the cartoon-style images created by each model. We randomly choose a series of images for various cartoon styles and then ask users to score the batch of images. We take the user score and determine how often each model was chosen as a percentage of overall selections. Fig. 9 shows the user preference score for cartoon-style images produced by the four distinct algorithms. As shown in Fig. 9, our proposed model outperformed other models in both styles. Although CycleGAN and CartoonGAN both produce cartoon-style images, their generated cartoon-style images are distorted, which influences the user's decision during the experiment.

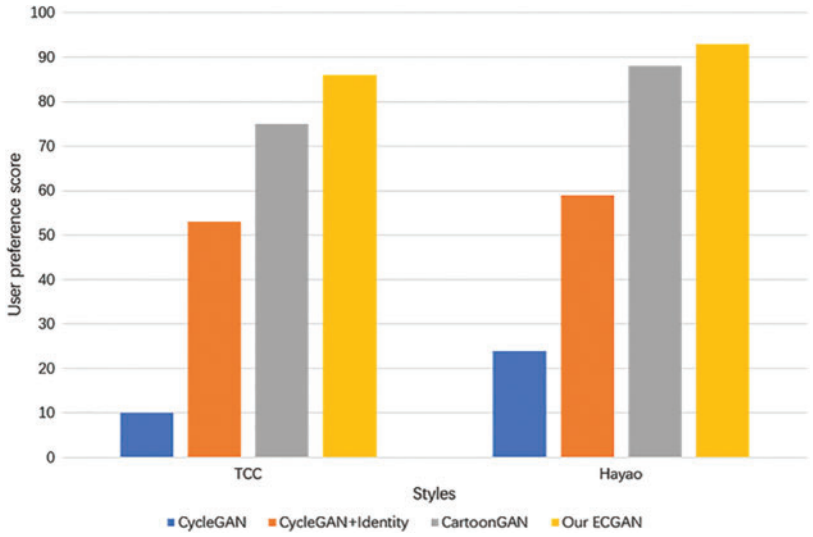

Figure 9: User preference score of different models for TCC and Hayao style

5 Conclusion and Future Work

In this study, we proposed an enhanced generative adversarial network to effectively transfer real-world images to cartoon images. To acquire more feature information from the real-world image, an enhanced generator is proposed, which can effectively acquire high-quality feature information from the input image. The Bayesian convolutional neural network is introduced to replace the standard convolutional neural network to improve the model's performance. A new dataset is also proposed with two different new styles, such as traditional and modern Chinese cartoons, to diversify the cartoonization of real-world images. Extensive experiments were done to evaluate the performance of the proposed model. The results indicate that the proposed model outperformed other models in producing high-quality cartoon images and a very good FID score.

In the future, we would like to work on extracting more styles to transfer real-world images into diversified cartoon styles. We also like to investigate the feature extraction process to acquire better

feature information from real-world images. We also like to work on implementing the proposed model in transferring human faces and videos to cartoon styles. In the future, we would like to explore the potential of incorporating attention mechanisms and reinforcement learning to improve the performance of the proposed model in generating diverse and high-quality cartoon images. Moreover, the integration of user feedback and customization options can be explored further for more interactive and personalized cartoonization translation, which could have significant applications in various fields such as entertainment and advertising.

Funding Statement: The author received no specific funding for this study.

Conflicts of Interest: The author declares he has no conflicts of interest to report regarding the present study.

References

[1] J. Zhu, T. Park, P. Isola and A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE Int. Conf. on Computer Vision (ICCV)*, Venice, Italy, pp. 2223–2232, 2017.

[2] Z. Zheng and J. Liu, "P^2-GAN: Efficient style transfer using single style image," arXiv preprint arXiv:2001.07466, 2020.

[3] L. A. Gatys, A. S. Ecker and M. Bethge, "Image style transfer using convolutional neural networks," in *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 2414–2423, 2016.

[4] N. Kalischek, J. D. Wegner and K. Schindler, "In the light of feature distributions: Moment matching for neural style transfer," in *2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 9382–9391, 2021.

[5] J. J. Virtusio, J. J. Ople, D. S. Tan, M. Tanveer, N. Kumar *et al.,* "Neural style palette: A multimodal and interactive style transfer from a single style image," *IEEE Transactions on Multimedia*, vol. 23, pp. 2245–2258, 2021.

[6] P. Chandran, G. Zoss, P. Gotardo, M. Gross and D. Bradley, "Adaptive convolutions for structure-aware style transfer," in *2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 7972–7981, 2021.

[7] J. An, S. Huang, Y. Song, D. Dou, W. Liu *et al.,* "ArtFlow: Unbiased image style transfer via reversible neural flows," in *2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 862–871, 2021.

[8] K. Xu, L. Wen, G. Li, H. Qi, L. Bo *et al.,* "Learning self-supervised space-time CNN for fast video style transfer," *IEEE Transactions on Image Processing*, vol. 30, pp. 2501–2512, 2021.

[9] F. Luan, S. Paris, E. Shechtman and K. Bala, "Deep photo style transfer," in *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 4990–4998, 2017.

[10] Y. Liu, F. E. Munsayac, N. T. Bugtai and R. G. Baldovino, "Image style transfer with feature extraction algorithm using deep learning," in *2021 IEEE 13th Int. Conf. on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, Manila, Philippines, pp. 1–5, 2021.

[11] M. Kumar, S. Srivastava and N. Uddin, "Forgery detection using multiple light sources for synthetic images," *Australian Journal of Forensic Sciences*, vol. 51, no. 3, pp. 243–250, 2017.

[12] W. Xu, C. Long, R. Wang and G. Wang, "DRB-Gan: A dynamic ResBlock generative adversarial network for artistic style transfer," in *2021 IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 6383–6392, 2021.

[13] H. Chen, L. Zhao, Z. Wang, H. Zhang, Z. Zuo *et al.,* "DualAST: Dual style-learning networks for artistic style transfer," in *2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 872–881, 2021.
[14] J. Chen, G. Liu and X. Chen, "AnimeGAN: A novel lightweight GAN for photo animation," in *Int. Symp. on Intelligence Computation and Applications*, Singapore, pp. 242–256, 2020.
[15] Y. Dong, W. Tan, D. Tao, L. Zheng and X. Li, "CartoonLossGAN: Learning surface and coloring of images for cartoonization," *IEEE Transactions on Image Processing*, vol. 31, pp. 485–498, 2022.
[16] X. Chen, C. Xu, X. Yang, L. Song and D. Tao, "Gated-GAN: Adversarial gated networks for multi-collection style transfer," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 546–560, 2019.
[17] B. Li, Y. Zhu, Y. Wang, C. W. Lin, B. Ghanem *et al.,* "AniGAN: Style-guided generative adversarial networks for unsupervised anime face generation," *IEEE Transactions on Multimedia*, vol. 24, pp. 4077–4091, 2022.
[18] Y. Shu, R. Yi, M. Xia, Z. Ye, W. Zhao *et al.,* "GAN-based multi-style photo cartoonization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 10, pp. 3376–3390, 2022.
[19] F. Zhang, H. Zhao, Y. Li, Y. Wu and X. Sun, "CBA-GAN: Cartoonization style transformation based on the convolutional attention module," *Computers and Electrical Engineering*, vol. 106, no. 1, pp. 108575, 2023.
[20] L. Feng, G. Geng, Q. Li, Y. Jiang, Z. Li *et al.,* "CRPGAN: Learning image-to-image translation of two unpaired images by cross-attention mechanism and parallelization strategy," *PLOS ONE*, vol. 18, no. 1, pp. e0280073, 2023.
[21] Preeti, M. Kumar and H. K. Sharma, "A GAN-based model of deepfake detection in social media," *Procedia Computer Science*, vol. 218, no. 1, pp. 2153–2162, 2023.
[22] J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 7132–7141, 2018.
[23] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770–778, 2016.
[24] Y. Chen, Y. K. Lai and Y. J. Liu, "CartoonGAN: Generative adversarial networks for photo cartoonization," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 9465–9474, 2018.
[25] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, 1986.
[26] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, Long Beach, CA, USA, pp. 6629–6640, 2017.
[27] 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.