Article in Press IJIMAI journal

A Diverse Domain Generative Adversarial Network for Style Transfer on Face Photographs

Rabia Tahir[1], Keyang Cheng[1]*, Bilal Ahmed Memon[2], Qing Liu[1]

[1] School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang (China)
[2] Ghulam Ishaq Khan Institute of Engineering Science and Technology Topi, Swabi (Pakistan)

Received 19 December 2021 | Accepted 21 June 2022 | Early Access 01 August 2022

unir LA UNIVERSIDAD EN INTERNET

Abstract

The applications of style transfer on real time photographs are very trending now. This is used in various applications especially in social networking sites such as SnapChat and beauty cameras. A number of style transfer algorithms have been proposed but they are computationally expensive and generate artifacts in output image. Besides, most of research work only focuses on some traditional painting style transfer on real photographs. However, our work is unique as it considers diverse style domains to be transferred on real photographs by using one model. In this paper, we propose a Diverse Domain Generative Adversarial Network (DD-GAN) which performs fast diverse domain style translation on human face images. Our work is highly efficient and focused on applying different attractive and unique painting styles to human photographs while keeping the content preserved after translation. Moreover, we adopt a new loss function in our model and use PReLU activation function which improves and fastens the training procedure and helps in achieving high accuracy rates. Our loss function helps the proposed model in achieving better reconstructed images. The proposed model also occupies less memory space during training. We use various evaluation parameters to inspect the accuracy of our model. The experimental results demonstrate the effectiveness of our method as compared to state-of-the-art results.

Keywords

CycleGAN, Gated-GAN, PReLU, Smooth L1 Loss, Style Transfer.

DOI: 10.9781/ijimai.2022.08.001

I. Introduction

Style transfer means to apply style of an image to another image by keeping the original content the same. Style transfer lies under the category of Image-to-Image translation. Some other examples of Image-to-image translation are transfer from satellite images to Google maps, winters to summers, night to day, etc. [1]. Our work is specifically about applying diverse style transfer from distinct style images to real photographs. These styles can be paintings of artist, animated images or cartoon, sketches etc. There is a limited research work which describes painting style transfer to human faces [2]. Painting style transfer also known as artistic style transfer that means transferring a real photograph into the painting style of some artist [3]. However, painting transfer techniques can be divided into three categories; texture transfer, stroke transfer and section transfer techniques. Texture transfer technique means to follow the texture pattern of painting and then transfer it to the content image. When we apply most of painting transfer techniques to human faces, it results in deformation. Artistic style translation or painting style translation field is facing many issues including appearance of artifacts on generated image. Moreover, if content image contains some organ of human body such as face or head portraits then it is more difficult to perform style translation, because it may destroy structure of the face [4].

Deep learning is getting popular day by day in many social media apps such as DeepArt.io and PRISMA which are most popular examples of deep learning involved in style transfer applications [5]. Therefore, in this work, we propose Diverse Domain Generative Adversarial Network (DD-GAN) for style transfer from famous paintings to human faces. Our model is based on [6], and we aim to perform fast training and better visual results without loss of semantic content. The loss functions used in this model helped to reduce artifacts on generated images. It is faster in training process and generates reconstructed images with preserved content i.e. face. We reduce complexity of the model by using smaller number of residual blocks without sacrificing the accuracy. Furthermore, we use various evaluation parameters to check the efficiency of the proposed model. Hence, our proposed model decreases training time, improves visual results after style translation, generates better reconstructed images is simpler to implement. The main contributions in this work are:

- A novel method DD-GAN is proposed which transfers diverse painting styles to human face photographs.
- A loss function based on SmoothL1 is used in the model that preserves the identity in reconstructed images.
- A simple training strategy and small number of residual blocks enable the model to reduce the training time.

II. Related Work

Style translation is a significant field of computer vision which is being studying from the last two decades [7]. There are many algorithms which were proposed for style translation with different

* Corresponding author.

E-mail address: kycheng@ujs.edu.cn

Please cite this article in press as:
R. Tahir, K. Cheng, B. A. Memon, Q. Liu. A Diverse Domain Generative Adversarial Network for Style Transfer on Face Photographs, International Journal of Interactive Multimedia and Artificial Intelligence, (2022), http://dx.doi.org/10.9781/ijimai.2022.08.001

types of deep neural network such as CNN or VGG network. However, a few works encapsulate the significance of Generative Adversarial Networks (GANs) in this field. As our work is related to style transfer with GAN therefore, this section explains some recent work of style transfer with various models of GAN.

A. Generative Adversarial Networks

Generative Adversarial Networks or GANs were proposed in 2014 [8] and brought a revolution in the field of machine learning and computer vision for fake image generation. They are generative models and consist of two parts; Generator and Discriminator. The generator is a generative model which generates fake data similar to the training data while discriminator network detects among real and fake images. The uniqueness of GAN is that both generator and discriminator train simultaneously and improve their performance with time.

GANs use different loss functions such as adversarial loss and cycle consistency loss in order to generate fake data. From 2014 till now, many variants of GAN have been proposed and they are used for various purposes such as image generation, style translation, super-resolution, and text-to-image generation [9]. Many GAN-based methods have been proposed for style translation specifically for multi domain tasks. This work also aims to produce a GAN variant which can perform fast style translation on human photographs. Although GAN and its variants have achieved high accuracy results in style translation tasks but still there are many challenges which are high computational time, need of rich resources, complexity of the model and unstable GAN training. There is need to sort these challenges in new variants of GAN for multi-domain style translation.

B. Image-to-Image Translation

Image-to-Image translation (I2I) methods are of two types; methods with paired data and methods without paired data. Firstly, Isola et al. [10] proposed pix2pix GAN with paired training samples. There are some other examples of I2I methods with paired training data in [11] and [12]. CycleGAN [1], ComboGAN [13], and StarGAN [14] are some examples of image-to-image translation with GAN in an unsupervised way. They performed style translation on multiple domains without paired examples and used as base work for other papers. Our method is also based on CycleGAN because it serves as a general purpose solution for various style translation tasks. In addition, we use transformer module of Gated-GAN to perform multi-domain style translation as CycleGAN requires multiple discriminators and generator module to perform multi-domain style translation. Hu et al. [15] proposed a style transfer model based on CycleGAN and VGG model. However, they used only one GAN structure instead of using two GAN. In order to preserve the semantic content of image, they used VGG network as a feature map. I2I methods have produced improved visualized results in style of various domains but resultant images still contain artifacts and blurriness. Moreover, these methods are unable to preserve the complete identity of the original images especially face. I2I methods need improved loss functions to produce high quality stylized images and reconstructed images with preserved identity.

C. Painting Style Transfer

Painting style transfer or artistic style transfer is another type of image-to-image translation. There are various works which perform artistic style transfer with different deep learning models such as CNN and GAN. Gay et al. [16] proposed a CNN-based style transfer technique which performs style translation on a content image by transferring style of an image. However, this work was computationally expensive which is replaced by recent works [17]-[20]. Zhang and Dana [21] proposed MSG-Net introducing a CoMatch layer in the model for style transfer. This model not only transfers the style to the target image, but also removes the artifacts. They produced better results regarding processing time and visual quality. Most of research is based on different types of style loss function used in style translation. For example, adversarial loss [8], perceptual loss [19], content loss [16], [22]. Huang et al. proposed a brush-based approach that inherits the spirit of the stroke rendering. They transform small patch of images into brushstroke of the target style. Only texture and color are changed while keeping the geometrical shape preserved [23]. The above mentioned methods mostly use common painting style images for style translation. However, a little research work is done which focused on diverse and unique style translation with the help of GAN. Therefore, this article is a contribution in this field. As we consider diverse multiple domains and perform style translation using one GAN model with fast training.

III. Overview of the Proposed Model

We propose Diverse Domain Generative Adversarial Network (DD-GAN) for multi-domain style translation on human photographs. This model specifically focuses on how we can apply different painting styles of artists on human faces and convert them to charming portraits. For this purpose, we adopt an architecture that consists of an encoder, decoder and number of transformers to perform style translation. However, we adopt an efficient training strategy with new loss functions in both generator and discriminator of our model. To stabilize training, we use PReLU in generator and LeakyReLU in discriminator of our model. Moreover, we use Smooth L1 function in reconstruction loss formula. Because Smooth L1 loss function has more benefits over L1 and L2 loss function as it combines advantages of both L1 and L2. Furthermore, it speeds up the training process. This loss function gives better results for reconstructed images as compared to other state-of-the-art methods. Further, we adopt 2D-Instance normalization to speed up the process of stylization. We modify ordinary weight initialization method in discriminator and generator model with Xavier weight initialization method. Our model is simple to adopt yet fasten for diverse style translation on human face photographs. The training time of the proposed model is decreased because small number of residual blocks. Moreover, we use various evaluation metrics to inspect the efficiency of model. For content images, we use Helen dataset as portrait of human. For style images, we use four diverse styles to train our model i.e. wall mural, iconography, painting and Albrecht Durer. All four style images categories are different from each other in terms of texture and pattern. Fig. 1 explains the working of the proposed model. The generator consists of encoder, list of transformers, decoders and instance normalization with PReLU activation function that is used in generator. The encoder contains three convolution layers and the decoder contains four residual blocks. The discriminator model we adopt is commonly used PatchGAN with two loss functions. Details of these components are explained in the next sections.

A. Auto-encoder With Transformers

In DD-GAN, we have a generator G which consists of an encoder E, decoder D and a number of transformers T, and a discriminator D. The number of T depends upon number of styles i.e. you can extend or decrease T based on choice of your style domain. The transformer is basically a number of the residual block and output of encoder (feature maps) is input of the residual block in Transformer T. The residual block consists of a stack of layers and it provides output of a specific layer to any other deep layer in the block. The basic purpose of using the gated transformer is to add multiple styles in a single generator. In our case, we have two domains of images namely; Human face photographs $x_i \in X$ and famous artistic painting images $y_i \in Y_k$ where k means number of style domains. In our case, we set k=4 because there are four diverse domains. There are no paired examples for these two

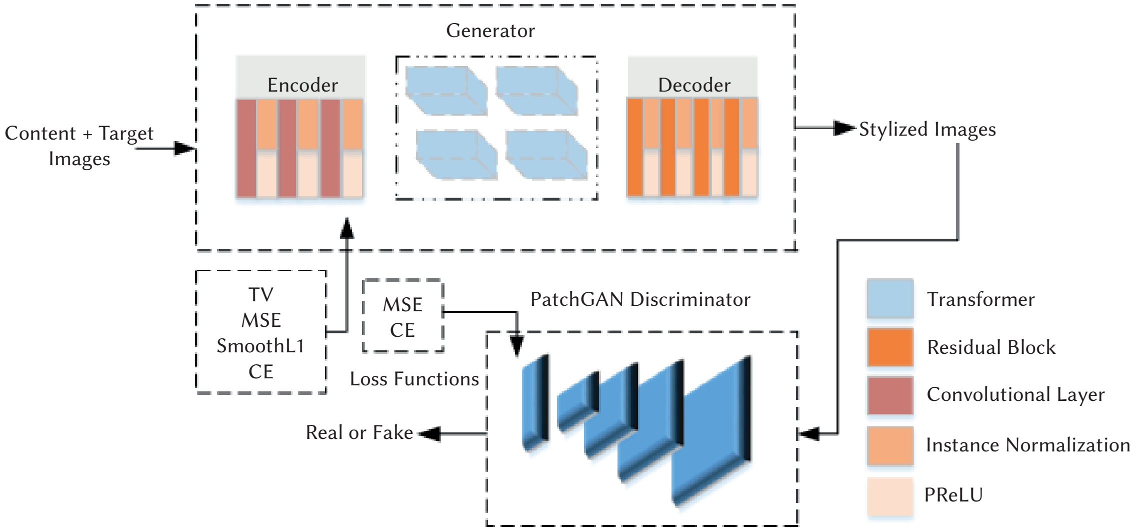

Fig. 1. The architecture of DD-GAN. It consists of an encoder, decoder and four transformers. The discriminator takes the real images and generated images and identifies whether they are real or not.

domains as it is a kind of unsupervised style translation like Cycle-GAN. There are two mapping functions in generator i.e. H and F. The aim of H mapping function is to generate a fake image y by translating painting style to human face photographs i.e. H : x⟶y. Then, we have an inverse mapping function F which converts the translated image back to its original state i.e. F : y⟶x. There is one encoder E in our generator G which encodes the important features of the input image into the feature space E(x) and gives it to Transformer T. This encoder comprises of several convolutional layers, while a convolutional layer is the main component of any deep neural network and comprised of various kernels or filters. The transformer T consists of 1 residual block and the decoder consists of four residual blocks. The transformer T takes the encoded input from encoder and assigns a specific style k to that input. It aims to give an output like G(x, k). The output of these 5 residual blocks is activation T(E(x)). Next, we have a decoder Dco which consists of fractionally strided convolution layers. The purpose of decoder D is to transform the T(E(x)) into output image G(x) i.e. Dco (T(E(x)) = G(x).

B. Discriminator of DD-GAN

The role of discriminator D is to distinguish among real y and fake samples G(x). Therefore, in DD-GAN, we have two separate discriminators D_y and D_x for both mapping functions H and F. The discriminator D_y learns to discriminate real paintings and fake generated painting, while D identifies among real photographs and reconstructed photographs. We train generator with PReLU and discriminator with LeakyReLU which helps to fasten and stabilize the training process. The loss functions in GAN play a very important role for stabilizing training procedure and better quality generation of images. In our model, we use four different loss functions: auto-encoder loss, total variation loss (TV), mean square error (MSE) and cross entropy (CE) loss.

C. Loss Functions

We adopt four loss functions in our model i.e. Mean Square Error (MSE), Smooth L1 reconstruction loss, total variation loss (TV), and cross entropy (CE) loss. First loss function is Least Square Generative Adversarial Network (LSGAN) loss which trains D and G simultaneously like a minimax game [24]. LSGAN can be implemented with the help of Mean Square Error (MSE). LSGAN helps to get non-saturating and smooth gradient in the discriminator D and it is defined as:

$$\mathrm{Loss}_{least}=\mathbb{E}_{x\epsilon X}\,[D(G(X))^2]+\mathbb{E}_{y\epsilon Y}[(D(y)-1)^2] \tag{1}$$

where D(G(x)) means that discriminator is provided a fake input to identify it. And D(y) means that we give the target label to the discriminator to identify among real and fake labels. The second loss function is auto-encoder reconstruction loss which is defined between real input x and reconstructed image $\bar{x}$. We use this loss function by combining both encoder and decoder module i.e. E and Dco. Auto-encoder reconstruction loss reduces the possible mapping function i.e. provides unique solutions and diverse outputs. We use Smooth L1 loss function between reconstructed image and original image and it is defined as:

$$\mathrm{Loss}_{rec}=\mathbb{E}_{x\epsilon X}\,[||\ \mathrm{Dco}\ \mathbb{E}\ (X)\text{-}X\ ||_{smoothl1}] \tag{2}$$

where Dco is decoder and $\mathbb{E}$ (x) is encoded feature space. And Dco(E(x)) means identical output like input x. Smooth L1 loss function is also known as Huber Loss and it is less prone to outliers as compared to MSE loss function. It is a combination of L1 and L2 loss functions. When training with L2 loss functions, there are chances of gradient exploding. Smooth L1 loss [25] eliminates this limitation and it is defined as:

$$\mathrm{L1}_{smooth}(x,y)=\frac{1}{n}\sum_i Z_i \tag{3}$$

where

$$Z_i=\begin{cases}0.5\times(x_i-y_i)^2\,, & if\ |x_i-y_i|<1\\ |x_i-y_i|-0.5, & if\ |x_i-y_i|\geq 1\end{cases} \tag{4}$$

where x and y contain n different elements and have random shapes. Smooth L1 loss acts like a combination of L1 and L2 losses. When the absolute value is near to zero it acts like L2 loss and when its value is high it acts like L1 loss function. It combines two major benefits of L1 and L2 loss functions which are steady gradients for high values of x and low oscillations for small values of x. With the use of Smooth L1 loss, our model generates reconstructed images with small amount of outliers and fastens the training process. As there are multiple styles in DD-GAN, so the model may get confused between multiple styles. Therefore, we use an auxiliary classifier to discriminate the style categories [6]. To calculate auxiliary classifier loss, we use Cross Entropy loss (CE) to measure the performance of discriminator model whose output is among 1 and 0. It compares the label with discriminator prediction and it is defined as:

$$\min_G \mathrm{Loss}_{clc}\,(G)=-\,\mathbb{E}_{x\epsilon X}\,[\log C\ (\ \mathrm{Style}=c|G(x,c))] \tag{5}$$

TABLE I. Generative and Discriminative Network of DD-GAN

Encoder	Transformer	Decoder	Discriminator
Conv2d(C=3,F=32, Instance2D,K=7,S=1) PReLU	RB(F=128, Instance2D,K=3,S=2) PReLU	4RB(F=128,Instance2D,K=3,S=2) PReLU	Conv2d(C=3,F=64, Instance2D,K=4,S=2)LeakyReLU
Conv2d(F=64, Instance2D,K=3,S=2) PReLU		ConvT2d(F=128,Instance2D,K=3,S=1/2) PReLU	Conv2d(F=128, Instance2D,K=4,S=2)LeakyReLU
Conv2d(F=128,Instance2D,K=3,S=2) PReLU		ConvT2d(F=64,Instance2D,K=3,S=1/2) PReLU	Conv2d(F=256,Instance2D,K=4,S=2) LeakyReLU
		ConvT2d(F=3.Instance2D,K=7,S=1) tanh	Conv2d(F=512,Instance2D,K=4,S=2) LeakyReLU
			Conv2d(F=1,K=4, S=1)
			Conv2d(nstyles, K=1, S=1)

where c means the index of style collections K i.e. C ϵ 1, 2, 3. . . K. And c is an auxiliary classifier. More details of this loss function can be seen in [6]. The last loss function is total variation regularization loss or TV loss which helps to get smoother generated images i.e. G(x, c). It is defined as [19], [26], [27]:

$$L_{TV}=\sum_{i,j}[\ (G(X)_{i+1,j}-G(X)_{i,j})^2+(G(X)_{i,j+1}-G(X)_{i,j})^2]^{1/2} \quad (6)$$

The overall loss function for generator G is described as:

$$Loss_G\ (total)= Loss_{least}+\alpha L_{TV}+\beta Loss_{clc}+\gamma Loss_{rec} \quad (7)$$

Where α, β, γ are hyper-parameters of weight consistency.

D. PReLU-based Generator

The activation function plays a significant role in neural networks especially in GAN. In most GAN models, we see that ReLU activation is used in the generator while LeakyReLU activation is used in the discriminator. It is popular to use them as activation functions in many neural networks. In any deep learning model, the activation function plays a vital role. Therefore, it is very important to choose a suitable activation function while designing your own model. In our model, we choose Parametric Rectified Linear Unit (PReLU) instead of ReLU activation function in our generator model and LeakyReLU in the discriminator. PReLU adds additional parameters as compared to ReLU. The convergence rate of PReLU is faster as compared to other activation functions such as ReLU and sigmoidal. Therefore, the ultimate purpose of using PReLU activation function in the generator model is to automatically tune the parameters which helps in improving the accuracy rate [28], [29].

E. Network Architecture

We use the network architecture proposed by Chen et al. [6] but with some modifications. In DD-GAN, we have three basic modules named encoder, decoder and transformer. The generator contains three modules named encoder, decoder and transformer. The discriminator is used to identify that the image is real or fake. Table I shows the layers specification of our network. The encoder consists of three layers of Conv2D with instance normalization and PReLU as activation function. Zuo et al. proposed DPGAN [28] to use PReLU in generator while LeakyReLU in discriminator. Therefore, by following it, we are using the same in our model. We use one residual block in Transformer with PReLU and instance normalization. While decoder consists of 4 residual blocks, 2 transpose Conv2D, and 1 Conv2D layer along with PReLU and instance normalization. We reduced the number of residual block from five to four in decoder to make the model less complex and to fasten the training process. We use one up sampling and three down sampling layers in our encoder. We use Markovian Patch GAN architecture for the discriminator because it has a small number of parameters which can applied to various sizes of input [1]. This type of discriminator is effective because it assumes independence among all pixels separately, while these pixels are separated by a patch diameter. The discriminator contains five Con2D layers with instance normalization and LeakyReLU activation function.

F. Instant Normalization and Xaviar Weight Initialization

We use instance normalization in all layers of the encoder, residual block and decoder. Also, we use instance normalization in all layers of discriminator [30]. Replacing batch normalization with instance normalization produces better results especially for style generation tasks. It is better than batch normalization because it independently normalizes all elements of the batch. While training any neural network, the weight initialization is an important step. Too much small weights can lead to vanishing of gradient while too large size weights can lead to explosion of gradient. Xaviar weight initialization [31] method solves this problem by keep the variance same in each layer of the network. Therefore, we use Xaviar weight initialization method because it also gives good performance for style translations tasks [32]. Our model consists of Conv2D layers, therefore we initialize weight with Xaviar normal technique in both generator and discriminator models. As compared to normal weight initialization method, it selects the weight from Gaussian distribution with values zero mean and 1/n variance, where n denotes the number of neurons in input [33].

IV. Datasets and Experiments

A. Datasets

1. Helen Dataset

This paper proposes a model for painting style transfer with diverse domains and five diverse datasets are used for experiments. We use Helen dataset as a content resource. Helen dataset is a famous dataset for facial recognition task. We use images of these datasets as content images in our model. All images in this dataset are portrait images of human faces. This dataset contains 2000 training images and 330 testing images [34]. We use 856 images for training purpose. For style exemplar, we use four different datasets from Kaggle. They are Wall Mural[1], Iconography[2], The Work of Painting[2] and Albrecht_Durer[3].

2. Wall Mural

Mural is a kind of art which is applied directly on some wall, surface or ceiling. Wall Mural is a collection of wall mural painting collected from Kaggle. In this dataset, there are 10,200 images of

[1] https://www.kaggle.com/vbookshelf/art-by-ai-neural-style-transfer

[2] https://www.kaggle.com/thedownhill/art-images-drawings-painting-sculpture-engraving.

[3] https://www.kaggle.com/supratimhaldar/deepartist-identify-artist-from-art/data

human portraits in mural style. The size of images is 400×300 pixels. We take 500 images from this collection as our first style domain.

3. Albrecht Durer

Albrecht_Durer is a collection of a German artist Albrecht Durer (1471-1528). His paintings mostly consist of portraits, water colors and altarpieces. We use 324 images from his collection of drawing and engravings. This collection includes black and white, gray color based engraving drawings of this artist. We use these images as our second domain.

4. Iconography

The Iconography is a collection of icons and works of old Russian applied art, ranging from the artists of 10th to the 18th centuries. We take 500 images from this collection as our third style domain.

5. The Work of Paintings

The Work of Paintings is a collection of Russian Museum's paintings by artists of 18th, 19th and 20th centuries. We take 500 random images from this collection. All images are mostly the self-portrait with dark brown, red and gray texture. These paintings are very colorful, bright and clear in content. Therefore, our total number of images in style training dataset is 1824 while 500 images of Helen as training content. We use 330 images of Helen dataset for testing. Fig. 2 shows all style images that are used in experiments. All categories of painting possess diverse characteristics which can be seen in Fig. 2.

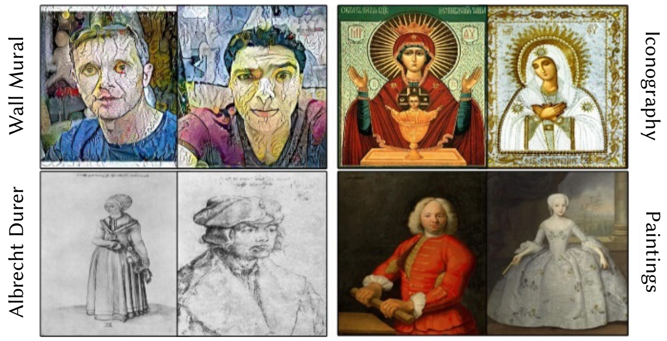

Fig. 2. All four diverse style painting images.

B. Training Strategy

We use batch size of 1 with 120 epochs for trainings. The load size of images is 128×128. The reason of small size of images is to reduce the computational cost. However, for testing phase, we use the original size of images i.e. 256×256 to evaluate the performance. Other details of parameters used in DD-GAN are given in Table II.

TABLE II. Experimental Settings for DD-GAN

Parameters	Value
Epoch	120
Batch size	1
Input size	128 ×128
λ_A	10
Learning rate	0.0002
TV weight	1e-6
No of styles	4
Decay epochs	80
Reconstruction weight	10
β_1 and β_2	0.5 and 0.999

V. Results and Analysis

A. Qualitative Results

Fig. 3 shows qualitative results of DD-GAN with Helen dataset and other four style datasets Wall Mural, Painting, Iconography, Albrecht Durer and reconstructed images. We take human face photographs from Helen dataset as a content image and apply style transfer process on these images after training. We can see the newly generated style transferred images along with reconstructed images. From Fig. 3, we can see that style transfer to iconography and painting images are visually less attractive as compared to Wall mural and Durer images. The reason of this difference is because of dynamic nature of both datasets iconography and painting as both of these datasets contain paintings from different artists. Therefore, it is difficult to train the model. Contrary, Wall Mural and Albrecht Durer are two datasets which contain paintings of one artist and are of same type. Therefore, the resultant images are more appealing and better as compared to the other two datasets. And, reconstructed images are very much similar to the original images because of Smooth L1 loss function.

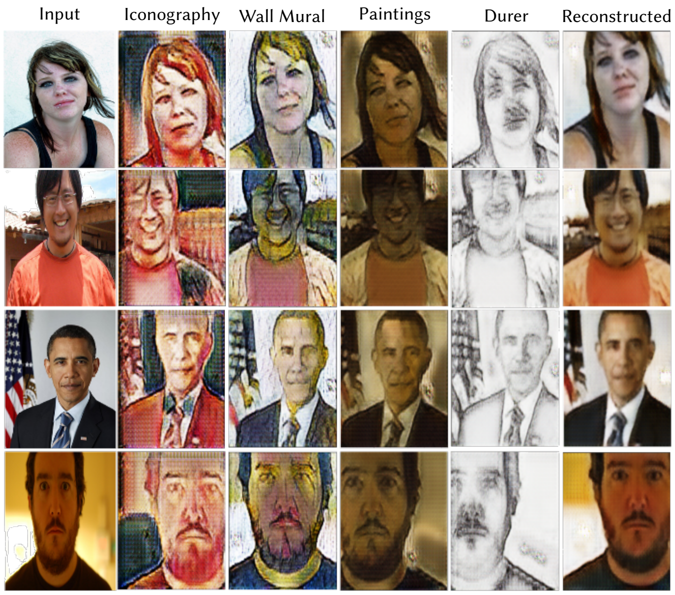

Fig. 3. Qualitative results of DD-GAN on Helen dataset.

Another important thing is the preservation of shapes and edges of human faces after style translation. We can observe that the important features of faces are preserved after style translation. The aim of DD-GAN is to make sure the preservation of face identity. As in style translation, we do not want to change the content and lose its original identity. In Fig. 4, we compared Gated-GAN with DD-GAN in reconstruction phase. We observed that black empty hole were present in most of reconstructed images of Gated-GAN. We zoomed and cropped these images and showed them in Fig. 4. The improvement in results is because of reconstruction loss formula which is comprised of Smooth L1. It helped in achieving better results by removal of distortion in generated images.

However, our results are quite better but little blurry as compared to Gated-GAN. In Fig. 5, we visually compare results of Gated-GAN with our model for style domain Iconography. Gated-GAN performed better as compared to our model in terms of texture appearance. It is obvious in this figure, that Gated-GAN learned texture and style of Iconography dataset in more efficient way and then implemented it on Helen dataset. However, it failed to preserve the content i.e. face of person. Contrary, our model preserved the identity but failed to transfer color and style texture of target dataset. Further, our resultant images contain less distortion as compared to the base model which shows superiority of our model up to some extent. In Fig. 6, we compared some style transferred images of DD-GAN and Gated-GAN. All four style domains i.e. wall mural, iconography, painting and Durer are compared in this figure. It is obvious from this comparison that our results are better as compared to the other model especially for mural, painting and Durer. However, Gated-GAN showed better results for iconography style domain as compared to DD-GAN. Our model generated style transferred images with less distortion with more clear representation of texture of target domain. However, Gated-GAN produced images with noise such as visible black dots.

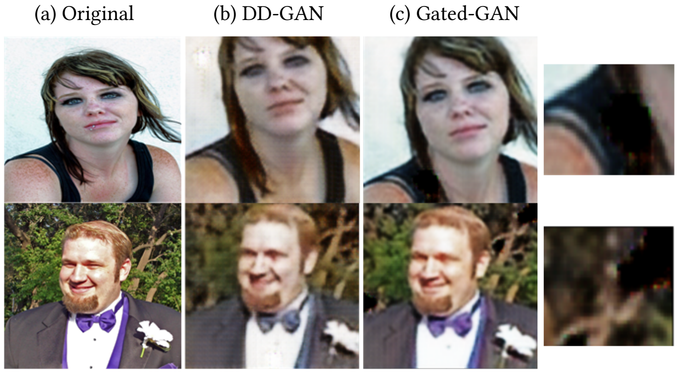

Fig. 4. Comparison between reconstructed images of Gated-GAN and DD-GAN.

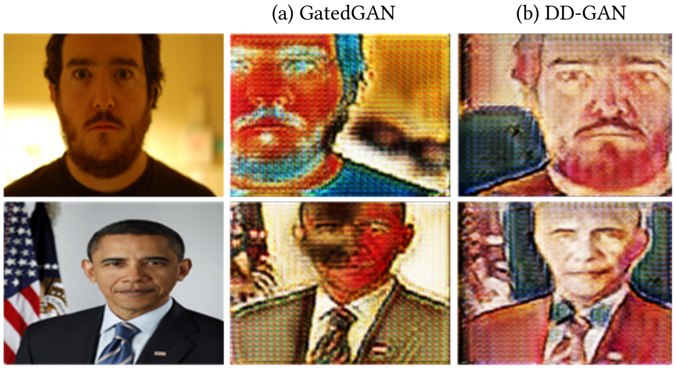

Fig. 5. Visual comparison of Iconography style transfer on Helen dataset.

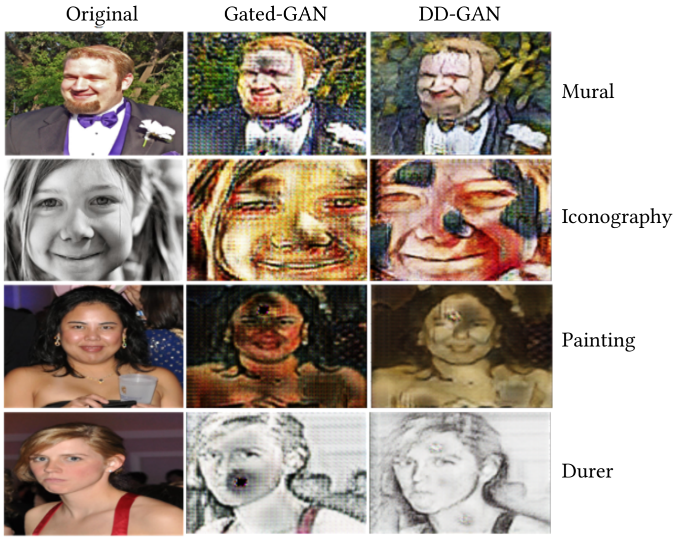

Fig. 6. Comparison of our model with the base model for all four style domains.

B. Discussion on Evaluation Metrics With Quantitative Results

In this section, we explain some popular evaluation metrics to inspect quality of generated images especially with GAN. There are some common evaluation metrics such as FID, MSE, PSNR, SSIM and MS-SSIM for the assessment of image quality [35]. Therefore, we use these five evaluation metrics to quantify our results. Mean Square Error (MSE) calculates the average of the square of difference among the target image and generated image. An MSE with small value shows higher similarity while MSE with high value shows less similarity. A smaller MSE value means that model is performing well. For example, zero MSE means that model is perfect that means the two images are identical. Peak Signal-to-Noise Ratio (PSNR) is an expression of ratio between signal and noise, where noise is the error produced by compressed image and signal depicts original image. The more the value of PSNR means better results. The more value of PSNR means that two images are more similar [35]. Structural Similarity Index (SSIM) [12] was proposed by Wang et al. to inspect the quality of an image. It is a perceptual evaluation metric and it calculates the image quality degradation. The values closer to 1 means high accuracy and values closer to zero mean less accuracy.

We use SSIM to check the similarity among original image and fake generated image after applying style transfer. Extensive version of SSIM is MS-SSIM [36] (Multi-Scale Structural Similarity Index) that calculates the similarity index among two images at different scales. It performs better than SSIM. All of above metrics are not enough to inspect efficiently the visual quality of images. Therefore, we use another state-of-the-art metric Frechet Inception Distance (FID) for our generated images. FID is a metric which is proposed to inspect quality of generated images especially by GAN. It is an improved version of Inception score. It takes a collection of original images and generated images by GAN. It basically calculates the distance among two different types of collection of images. FID for same collection of images becomes zero. The more the FID score, the more difference among two collections exits [37]. Fig. 7 shows FID score of all categories of our diverse style domains. We used two categories (original and style transferred) to calculate FID score. We can see the reconstructed images obtained the best FID score i.e. lowest score.

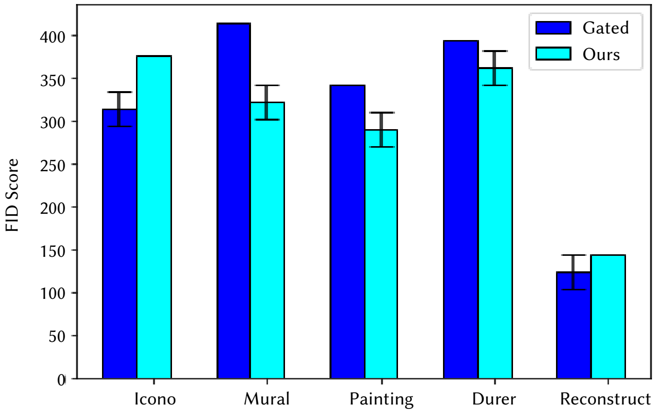

Fig. 7. Fid score for original images and style transferred images. Black lines on bars show the best fid sore.

However, Iconography obtained highest fid score which shows the average quality of style transferred images for this category. The reason of best fid score of reconstructed images is that the images are more similar to the original images. However, when we apply style to content images then texture and color of these images become change. Therefore, fid of style transferred images is slightly higher as compared to reconstructed images. If we compare all four style categories, we can observe that Painting images obtained best FID score as compared to rest of three categories. The reason of this lowest score is the less dynamic nature of this dataset. However, the Iconography paintings obtained highest fid score because the collection of these paintings possesses diverse styles. In Fig. 7, we compared FID scores of both models. Gated-GAN obtained less FID score for reconstructed images and Iconography as compared to our model. However, it produced high FID score for Durer, Wall Mural, and Painting datasets. Table III shows the comparative analysis among Gated-GAN and our model for reconstruction images. We take average of 15 images in testing phase for both Gated-GAN and DD-GAN.

TABLE III. Comparative Analysis Of Various Error Rates of Reconstructed Images

	MSE	PSNR	SSIM	MS-SSIM
Gated-GAN	699.3	19.97	0.32	0.10
Our Model	618.96	20.41	0.32	0.09

The values in Table III are comparison between original image and reconstructed image after applying style translation. Our model gives less MSE and higher PSNR value as compared to Gated-GAN which shows the superiority of our model. Because we used Smooth L1 loss function in our reconstruction phase, therefore the results are better as compared to Gated-GAN. However, our model obtained low accuracy values for MS-SSIM as compared to Gated-GAN, while SSIM values are the same for both models. In Table IV, we present MSE, PSNR, SSIM and MS-SSIM values of all four style categories. The Painting category achieved the best results as compared to remaining three style domains which shows its better visual quality. Among all datasets, Albrecht Durer obtained highest MSE and lowest PSNR which shows the complex nature of this dataset. There is always a trade-off between accuracy and computational time of any neural network. The best model focuses not only on achieving high accuracy but also on decreasing computational time. Therefore, we also compare different times for training and testing phases. Table V and VI compare time complexity of our model with the base model Gated-GAN.

TABLE IV. Comparative Analysis of Various Error Rate on All Four Style Transferred Images

	MSE	PSNR	SSIM	MS-SSIM
Wall Mural	960.62	18.43	0.15	0.04
Albrecht Durer	2139	15.29	0.21	0.02
Painting Images	854.59	19.47	0.24	0.05
Iconography Images	1383	16.55	0.10	0.003

TABLE V. Comparative Analysis of Training Time for 1 and 120 Epochs

	Gated-GAN	DD-GAN
Each Epoch	~4 to 5 minutes	~4 minutes
120 Epochs	~7 hours 54 minutes	~7 hours 30 minutes

TABLE VI. Elapsed Time for Reconstructon of Images During Testing Phase

	Gated-GAN	DD-GAN
MS-SSIM+ SSIM (ms)	20.86	19.46
MSE+PSNR(ms)	1.6	1.8

Table V shows time for first epoch and 120 epochs of Gated-GAN and DD-GAN during training phase. Our model completes training in less time as compared to the other. The reason is using a small number of residual blocks in generator and using of PReLU activation. This proves our model is a fast style transfer model for different types of images. Table VI compares time for reconstruction of images during testing period. During the testing phase, Gated-GAN achieved minimum time for the calculation of MSE and PNSR values and DD-GAN obtained minimum time for the calculation of SSIM and MS-SSIM. We also compared our model results with CycleGAN in terms of FID score. As CycleGAN is two domains generated network which can transfer to one style at a time. Therefore, we performed style transferred for two domains separately i.e. wall mural and Durer. Also, we checked the quality of reconstructed images after style translation. We noted that it took a lot of time to train CycleGAN with only one style domain. Results are comparative with our model and Gated-GAN but the drawback is domain limitation and computational time. Table VII compares FID score of our model with CycleGAN. Moreover, we conducted a short survey about quality of generated images of our model with base model. Results of this survey are given in Fig. 8. A total of 100 responses were received for this survey. We randomly choose images generated from Gated-GAN and DD-GAN from all style domains including reconstructed images. Then, we ask users to choose the best among two in each category. The purpose of this survey was to get feedback from people who do not belong to this field and only choose image according to its visual appearance. For our model, we received 64.52% positive responses while 42.2% positive response for base model. This shows better performance of our model. Category wise responses can be seen in Fig. 8.

TABLE VII. Comparison of FID Score for CycleGAN and DD-GAN

Style Domains	FID score
Wall Mural (CycleGAN)	255.60
Wall Mural (DD-GAN)	244.03
Durer (CycleGAN)	260.267
Durer (DD-GAN)	372.87
Reconstructed Images (CycleGAN)	89.86
Reconstructed Images (DD-GAN)	144.48

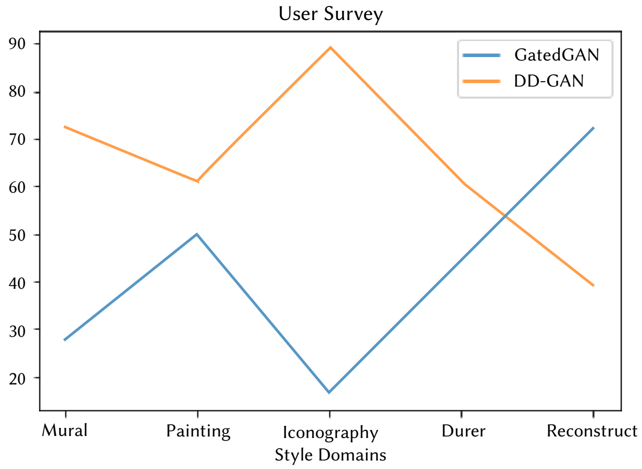

Fig. 8. User survey about style generated images between two models.

C. Ablation Study of Loss Function

In this section, we check the significance of all loss functions used in our model. The purpose is to ensure the usage of each loss function that either it is making some contribution in improvement of results or not. For this, we performed various experiments with removal of one loss function. Table VIII shows results of these experiments. Firstly, we check the significance of Total Variation (TV) loss in our model. We removed it and then accomplished our training. A clear fall in accuracy can be seen in Table VIII. The reason is that TV loss is used to remove noise by making sure the smoothness and spatial continuity in generated images. Therefore, when we removed it, increase in FID score and MSE, decrease in PSNR, SSIM and MS-SSIM can be seen. The second important loss function is reconstruction loss which plays an efficient role in performance of our model. As our model contains an encoder, decoder that makes it an auto-encoder. Basically, auto-encoder compressed an image to spatial features then reconstructs the image from these features. Hence, there is loss of some pixels and degradation of quality in generated image. For this purpose, reconstruction loss is proposed to measure the distance among original image and reconstructed image. There are many ways to implement this loss. In our model, we use Smooth L1 as a reconstruction loss function. When we remove this loss from our model, a sharp decrease of model's performance can be seen in Table VIII. This proves the importance and value of reconstruction loss in our model. We also compared training time with and without these loss functions. Table IX shows results of these experiments. The removal of TV and reconstruction losses leads to reduction the training time and accuracy. When we removed auxiliary classifier loss from our model, it resulted in no discrimination of style generated images. For example, we select Durer style to transfer on content image during testing and it gives us output image with painting style. Also, the quality of generated images is not good because loss of content structure i.e. face. Examples of some of these images are given in Fig. 9. In this figure, all three images were assigned Durer style domain at time of testing but the model failed to adopt this style and transferred mixture of other style domains. This proved the significance of auxiliary classifier loss in our model.

TABLE VIII. Ablation Study of Loss Functions in Our Model

	FID	MSE	PSNR	SSIM	MS-SSIM
DD-GAN	144.48	618.96	20.41	0.32	0.09
DD-GAN without TV loss	139.82	680.1	20.2	0.31	0.057
DD-GAN without reconstruction loss	468.30	1008	18.01	0.181	0.024

TABLE IX. Time Consumed for Training With and Without Loss Functions

	Time
Wall Mural (CycleGAN)	255.60
Wall Mural (DD-GAN)	244.03

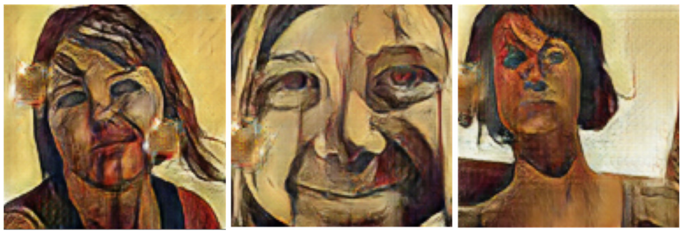

Fig. 9. Generated images with no AC loss in Durer style domain.

D. Discussion

The major issue in style transferred methods is their evaluation. There is no exact parameter of comparing style generated images. Some researchers use feedback from different people to compare images [38]. While some use evaluation parameters such as FID, MSE etc. But there is no guarantee that small MSE means good results. During our experiments, we observed that some poorly generated images produce small MSE and high PSNR. And some best style transferred images show high MSE and low PSNR. This is the reason we used multiple evaluation parameters to compare our results. We tried our best to present the results and comparison in an efficient way. Firstly, we compared our results with base work i.e. Gated-GAN at 120 epochs. Table V shows the comparison among DD-GAN and Gated-GAN for computational time. It is clear that our model performs faster on the same dataset. We used a different approach in our model and training strategy which results in fast computational time and results are almost the same like original Gated-GAN. Increasing training time may result in better visualization results. However, our method DD-GAN is fast as compared to original Gated-GAN. Our method can be applied to those problems where fast computation and results are required. The reason of fast time processing is our simple architecture and choice of loss function that results in rapid results. We faced two main limitations of our model DD-GAN. First is the production of artifacts in generated images especially for style domains Painting and Iconography. The reason is complex and dynamic nature of these two datasets. These two datasets contain paintings from multiple artists and possess different style and color texture. Hence, some generated images after style transfer contain artifacts on images which spoil the face of a person. The second limitation is production of blurry images in the reconstruction phase.

VI. Conclusion and Future Work

In this work, we proposed a novel and fast GAN variant named DD-GAN (Diverse Domain Generative Adversarial Network) for diverse painting style transfer on human face photographs. The DD-GAN applies different styles on human faces and converts them into realistic and beautiful art pieces using one GAN model. The purpose of this research is to add a contribution in the field of neural style transfer specifically for painting style transfer to human faces. We used a new loss function in our model in order to increase the accuracy and decrease the computational cost. Moreover, we used PReLU activation function in our model in order to improve the results. We have obtained a state-of-the-art qualitative and quantitative results which shows the efficiency of our model. In the future, we want to use more dynamic and complex datasets for training. Moreover, we want to improve the visual quality of newly generated images without increasing computational time. This work can be extended by performing training with more iteration which is possible by the availability of resources. As we lack rich resources, therefore we performed and compared results with small number of epochs. In the future, we aim to improve the visual results by using efficient resources with less complex model.

Acknowledgement

This research is supported by Natural Science Foundation of China (No. 61972183).

References

[1] J. Zhu., P.Taseng, I. Philip and E. Alexie, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision,* 2017, pp. 2223-2232.

[2] A. Selim, E. Mohamed and D. Linda, "Painting style transfer for head portraits using convolutional neural networks," *ACM Transactions on Graphics (ToG),* vol. 35, no. 129, pp. 1-18, 2016, doi: https://doi.org/10.1145/2897824.2925968.

[3] B. Rahul. and S. Jianlin, "Artist style transfer via quadratic potential," arXiv preprint arXiv, 2019, doi: https://doi.org/10.48550/arXiv.1902.11108.

[4] A. Khan et al., "Photographic painting style transfer using convolutional neural networks," *Multimedia Tools and Applications,* vol. 78, no. 14, pp. 19565-19586, 2019.

[5] L. Du, "How much deep learning does neural style transfer really need? an ablation study," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 3150-3159.

[6] C. Xinyuan, X. Chang, Y. Xiaokang, S. Li amd T. Dacheng, "Gated-gan: Adversarial gated networks for multi-collection style transfer," *IEEE Transactions on Image Processing,* vol. 28, no. 2, pp. 546-560, 2018, doi: https://doi.org/10.1109/TIP.2018.2869695.

[7] X. Li, S.Liu, K. Jan and M. Yang, "Learning linear transformations for fast image and video style transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 2019, pp. 3809-3817.

[8] I. Goodfellow et al., "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[9] K. Cheng, T. Rabia, L. Eric and M. Li, "An analysis of generative adversarial networks and variants for image synthesis on MNIST dataset," *Multimedia Tools and Applications*, vol. 79, no. 19, pp. 13725-13752, 2020, doi: https://doi.org/10.1007/s11042-019-08600-2.

[10] I. Phollip, J. Zhu, T. Zhou, and A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125-1134.

[11] S. Patsorn, J. Lu, C. Fang and F. Yu, "Scribbler: Controlling deep image synthesis with sketch and color," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2017, pp. 5400-5409.

[12] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P., "Simoncelli, Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600-612, 2004.

[13] A. Anoosheh, A. Eirikur, T. Radu and G. Luc, "Combogan: Unrestrained scalability for image domain translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 783-790.

[14] Y. Choi et al. "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789-8797.

[15] H. Chan, Y. Ding and Y. Li, "Image style transfer based on generative adversarial network," in *IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, vol. 1, pp. 2098-2102, 2020.

[16] L.A. Gatys, S. Alexander and B. Matthias, "A neural algorithm of artistic style," arXiv preprint arXiv:1508.06576, 2015.

[17] B. Mohammad and G. Golnaz, "Adjustable real-time style transfe, " arXiv preprint arXiv:1811.08560, 2018.

[18] G. Agrim, J. Justin, A. Alahi and F. Li, "Characterizing and improving stability in neural style transfer," in *Proceedings of the IEEE International Conference on Computer Vision.* 2017, pp. 4067-4076.

[19] J. Johnson, A. Alahi, and L. Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision.* 2016, pp. 694-711.

[20] K. Dmytro, S. Artsiom, P. Ma, S. Lang and B. Ommer, "A content transformation block for image style transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2019, pp. 10032-10041.

[21] Z. Hang and K. Dana, "Multi-style generative network for real-time transfer," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops.* 2018.

[22] H. Xun and B.Serge, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision.* 2017, pp. 1501-1510.

[23] H. Fay and C.L. Chein, "Patch-based Painting Style Transfer," in *IEEE International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan)*, pp. 1-2, 2020, doi: 10.1109/ICCE-Taiwan49838.2020.9258218.

[24] X. Mao et al, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision.* 2017, pp. 2794-2802.

[25] G. Sreekumar, "How to interpret smooth l1 loss? ," *StackExchange* Sep 2018. [Online]. Available: https://stats.stackexchange.com/questions/351874/how-to-interpret-smooth-l1-loss.

[26] L.I. Rudin, S. Osher and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: nonlinear phenomena*, vol. 60, no. 1-4, pp. 259-268, 1992.

[27] A. Hussein and E. Dubois, "Image up-sampling using total-variation regularization with a new observation model," *IEEE Transactions on Image Processing*, vol. 14, no. 10, pp. 1647-1659, 2005.

[28] Z. Fang and L. Xiaofang, "DPGAN: PReLU Used in Deep Convolutional Generative Adversarial Networks," in *Proceedings of the 2019 International Conference on Robotics Systems and Vehicle Technology.* pp. 56–61, 2019, doi: https://doi.org/10.1145/3366715.3366728.

[29] K. He, X. Zhang, S. Ren and J.Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision.* 2015, pp. 1026-1034.

[30] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," arXiv preprint arXiv:1607.08022, 2016.

[31] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249-256.

[32] T. Rabia, C. Keyang, E.K. Lubamba and M.S. Khokhar, "Multi-domain Cross-dataset and Camera Style Transfer for Person Re-Identification," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, pp. 2035-2041, 2019.

[33] N. Kumar, "Xavier Weight Initialization Technique in neural Networks," *The Professionals Point* 2019, [Online]. Available: http://theprofessionalspoint.blogspot.com/2019/06/xavier-weight-initialization-technique.html?

[34] V. Le, J. Brandt, Z. Lin, B. Lubomir and S.H. Thomas, "Interactive facial feature localization," in *European conference on computer vision*, 2012, pp. 679-692.

[35] U. Sara, M. Akter and M.S. Uddin, "Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study," *Journal of Computer and Communications,* vol. 7, no. 3, pp. 8-18, 2019.

[36] Z. Wang, E.P. Simoncelli and A.C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers,* vol. 2, pp. 1398-1402, 2003.

[37] H. Martin, R. Hubert, U. Thomas and B. Nessler, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in neural information processing systems.* vol. 30, 2017.

[38] N. Hulzebosch, S. Ibrahimi, and M. Worring. "Detecting cnn-generated facial images in real-world scenarios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops,* 2020, pp. 642-643.

Rabia Tahir

Rabia Tahir is currently a PhD student in Jiangsu University, Zhenjiang, China. Her interest areas are Image processing, Computer Vision, Deep Learning, Pattern Recognition and Computational Intelligence. She has published good impact factor papers in leading journals and conferences.

Keyang Cheng

Keyang Cheng is a member of CCF. He received the PHD Degree from School of Computer Science and Engineering, Nanjing University of Aeronautics & Astronautics in 2015. He has coauthored more than 30 journal and conference papers. He was at University of Warwick, UK, as a post-doctoral researcher in 2016. He is currently an associate professor and researcher in the School of Computer Science and Telecommunication Engineering of Jiangsu University. His current research interests lie in the areas of pattern recognition, computational intelligence and computer vision.

Bilal Ahmed Memon

Bilal Ahmed Memon is currently an Assistant Professor in GIKI University Swabi Pakistan. He has done his PhD in Management Sciences specialization in Finance from Jiangsu University, Zhenjiang China. He has published multiple impact factor papers in world's leading journal. His research area is complex networks in Finance. He is also reviewer of many journals.

Qing Liu

Liu Qing is currently pursuing the master's degree in computer application with Jiangsu University. His research interests include Computer Vision and Pattern Recognition.