Information Fusion 112 (2024) 102572

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

FISTNet: FusIon of STyle-path generative Networks for facial style transfer

Sunder Ali Khowaja [a,b,*,1], Lewis Nkenyereye [c,1], Ghulam Mujtaba [d], Ik Hyun Lee [e,f,*], Giancarlo Fortino [g], Kapal Dev [h,i,j]

[a] School of Computing, Faculty of Computing, Digital and Data, and ADAPT Centre Technological University Dublin, D07ADY7, Ireland
[b] Faculty of Engineering and Technology, University of Sindh, 76080, Pakistan
[c] Department of Computer Science, and Information Security, Seoul, 05006, South Korea
[d] Anderson College of Business and Computing, Regis University, Denver, 80221, United States of America
[e] IKLab Inc., Seoul, 08513, Gyeonggi-do, Republic of Korea
[f] Department of Mechatronics Engineering, Tech University of Korea, Siheung-Si, 15073, Gyeonggi-do, Republic of Korea
[g] Department of Informatics, Modeling, Electronics and Systems, University of Calabria, Calabria, Italy
[h] Department of Computer Science and ADAPT Centre, Munster Technological University, Bishoptown, Cork, T12 P928, Ireland
[i] Department of Electrical and Computer Engineering, Lebanese American University, Byblos, Lebanon
[j] Department of institute of intelligent systems, University of Johannesburg, Auckland Park, 2006, South Africa

ARTICLE INFO

Keywords:
StyleGAN
Style of fusions
GANs
Face style transfer
Style transfer networks

ABSTRACT

With the surge in emerging technologies such as Metaverse, spatial computing, and generative AI, the application of facial style transfer has gained much interest from researchers and startups enthusiasts alike. StyleGAN methods have paved the way for transfer-learning strategies that could reduce the dependency on the vast data available for the training process. However, StyleGAN methods tend to need to be more balanced, resulting in the introduction of artifacts in the facial images. Studies such as DualStyleGAN proposed multipath networks but required the networks to be trained for a specific style rather than simultaneously generating a fusion of facial styles. In this paper, we propose a Fusion of STyles (FIST) network for facial images that leverages pretrained multipath style transfer networks to eliminate the problem associated with the lack of enormous data volume in the training phase and the fusion of multiple styles at the output. We leverage pretrained styleGAN networks with an external style pass that uses a residual modulation block instead of a transform coding block. The method also preserves facial structure, identity, and details via the gated mapping unit introduced in this study. The aforementioned components enable us to train the network with minimal data while generating high-quality stylized images, opening up new possibilities for facial style transfer in emerging technologies. Our training process adapts curriculum learning strategy to perform efficient, flexible style, and model fusion in the generative space. We perform extensive experiments to show the superiority of the proposed FISTNet compared to existing state-of-the-art methods.

1. Introduction

Facial style transfer is an art form that is extensively used in a wide variety of applications, including social media filters, virtual character creation, animation production, advertising, non-fungible tokens (NFTs), and the Metaverse [1]. Many artists in the past have endeavored to create exaggerated or simplified replicas of real-world figures, either in cartoon, anime, or arcane styles. Usually, it takes intensive professional skills and laborious efforts to recreate real-world people in the aforementioned styles for artists. The advent of generative adversarial networks (GANs) has caused a paradigm shift in skill acquisition and artistic media creation through facial style transmission.

From a macro perspective, the field of facial style transfer belongs to the category of image-to-image translation. GANs have been extensively used for such translation tasks, especially for image-to-anime or other style conversion [2]. Generally, facial style transfer techniques can be categorized into portrait [3–5] and scene [2,6,7]-based style transfer methods, which have been applied and realized

* Corresponding authors.
E-mail addresses: sunderali.khowaja@tudublin.ie (S.A. Khowaja), nkenyele@sejong.ac.kr (L. Nkenyereye), gmujtaba@ieee.org (G. Mujtaba), ihlee@tukorea.ac.kr (I.H. Lee), giancarlo.fortino@unical.it (G. Fortino), kapal.dev@ieee.org (K. Dev).
[1] Sunder Ali Khowaja and Lewis Nkenyereye contributed equally to this work.

https://doi.org/10.1016/j.inffus.2024.102572
Received 3 April 2024; Received in revised form 11 June 2024; Accepted 3 July 2024
Available online 9 July 2024
1566-2535/

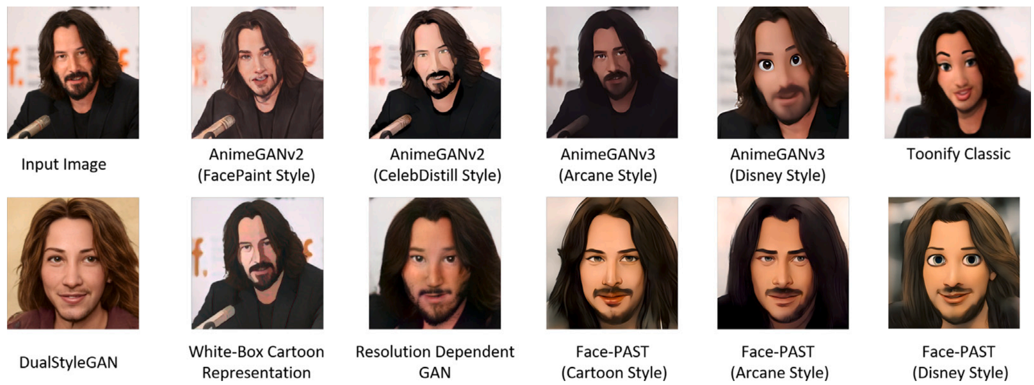

Fig. 1. The work proposes FISTNet that performs high-resolution facial style transfer, i.e., 1024 × 1024. The existing works either overfit styles onto the faces that do not preserve facial structure and characteristics, such as Resolution Dependent GAN, DualStyleGAN, and Toonify [15], or do not transfer diverse styles such as AnimeGAN and white-box cartoon representations [7]. In addition, studies such as AnimeGAN introduce artifacts into facial images. The aforementioned works can also generate different styles of images along with the proposed work, respectively.

in various applications. The former is best suited for producing caricature [3,8] and manga [4,5] inspired facial style images. The latter learns a transformation from photo to cartoon or any other style by leveraging pre-extracted representations and specialized losses. Both techniques rely heavily on induced facial landmarks and decomposed facial components, making the aforementioned methods unsuitable for applications centered around common scenes.

Another category for facial style transfer is the StyleGAN [9,10] or unsupervised image-to-image translation [11], which addresses the challenging task of selfie2anime. These methods are required to train for each specific style in order to generate a satisfactory output. Many existing studies have only considered using single-style facial transfer to generate a high-quality image. However, when switching to multi-style transfer or fusion of styles, the methods mostly generate unwanted artifacts with acceptable-quality images. A similar problem is also faced by studies that aim to apply customized styles, such as combining cartoon, anime, and arcane styles [12], Dreambooth [13], and Textual Inversion [14]. Another problem while performing multiple styles is that they cannot preserve facial poses, characteristics, and features accordingly. Misalignment between the facial features of target and source domains causes unwanted artifacts in the facial style transferred output images.

Fig. 1 shows a qualitative comparison between some existing works that completely transform the facial characteristics while compromising the original facial structure. Meanwhile, others retain the facial structure to a certain extent but fail to propagate high-quality and diversified styles. Furthermore, existing works train the network for each style separately to avoid incorporating multiple styles into the image. One of the seminal works is the multi-style cartoonization [12] that used multiple encoders for different style networks and multiple discriminators for identifying the cartoon style and selecting the appropriate style loss. Another work is the DualStyleGAN [16], which proposed the use of an external style pass and a progressive fine-tuning strategy to generate high-quality stylized images.

Recently, the study in [10] demonstrated that the hierarchical style training not only helps in stabilized training but also generates better visual results. The hierarchical style training process cascades the training process of two style paths with different parameters. However, training multiple style paths would neither preserve facial features when performing the style transfer nor handle diverse style transfers such as combining arcane, cartoon, and anime. Inspired by [16] that used multiple encoders for diversified styles, the fusion of pretrained style transfer models is hypothesized to provide better visual results and stabilized training while requiring less amount of training data.

To overcome the aforementioned issues, we propose a fusion of style-path generative networks (FISTNet) for facial style transfer, a multimodal generative network to introduce an effective multi-style transfer while realizing the control of facial characteristics in images. Inspired by the study DualStyleGAN and multi-style cartoonization, we leverage the hierarchical architecture from StyleGAN to embed multiple styles in fine and coarse-resolution layers via an extrinsic style path. This is done by leveraging the fusion of pretrained styleGAN networks to embed diverse styles in the form of intrinsic and extrinsic style paths. The intrinsic style path uses a base style (cartoonization) and residual blocks that retains the facial characteristics. In contrast to the DualStyleGAN, we use the pretrained encoders from existing state-of-the-art stylization networks to produce high-quality images along with gated mapping units to extract the domain-specific features. At times the extrinsic style path alters the behavior of pretrained networks, which might affect the style transfer as well as the quality of the output image. We show that the fusion of pretrained models helps in the non-alteration behavior of the extrinsic style path and positively affects the fine-tuning of convolutional layers. Furthermore, we maintain the facial details by considering identity, segmentation, and structural losses in the intrinsic style path. The main contributions of this work are as follows:

- We propose FISTNet for facial style transfer while preserving facial structure details.
- We adapt hierarchical style training such as extrinsic and intrinsic style transfer with pretrained encoders to generate better results.
- We perform the fusion of style networks to generate diverse yet facial preservation style transfer.
- Extensive quantitative and qualitative comparison has been carried out with current state-of-the-art methods.

The rest of the paper is structured as follows: Section 2 provides a consolidated review of image stylization in the context of the proposed study. Section 3 provides the working methodology of the proposed work. Section 4 presents experimental results to prove the effectiveness of the proposed approach. Section 5 concludes the work while highlighting future directions.

2. Related works

This section consolidates a brief review of existing works concerning style transfer with GANs, StyleGAN, and image-to-image translation techniques.

2.1. Non-photorealistic rendering

The NPR algorithms mainly render 3D shapes to create a cartoon-like effect through cel-shading techniques; however, adding cartoon style to existing photos is a much more challenging task [17]. The rendering of 3D shapes to impersonate cartoon-like effects leverages optimization and filtering methods. However, a high-level abstraction that exhibits an artistic effect is not achieved by doing so. Some studies used supplementary segmentation or added user interaction to generate cartoon portraits, respectively [18].

2.2. Style transfer networks

The use of style transfer was introduced to overcome the issues associated with non-photorealistic rendering. The traditional style transfer was proposed to generate stylized images using image training pairs. The use of such methods was able to create good results but was limited to a specific style. Researchers then proposed the use of the VGG network [19] for style transfer, which was considered to be good for extracting semantic features. The use of the VGG network also eliminated the issue of using training pairs. The use of neural style transfer was a good idea to start [20], but when using multiple faces, the style is transferred in a homogeneous manner that fails to produce smooth shading or clear edges. The study [21] used convolutional neural network (CNN) based feature maps and Markov Random Field for local matching to transfer the style. However, the transference yields semantically incorrect output. A deep analogy-based method was proposed in [22] to produce semantically correct output while extracting meaningful correspondences. But it was limited to a single style only and at times displayed undesirable results.

A dedicated CNN was recently used in the study [23] to classify between the comic and non-comic images to cope with the problem of undesirable results, but it was still limited to a single style. An alternate and popular approach for style transfer has been generative adversarial networks (GANs) [24]. In recent times, several studies have proposed the use of GANs to solve problems concerning pixel-to-pixel image synthesis. At the start, the GANs were considered impractical due to the need for a large set of paired images [25]. This fundamental issue was addressed by the study CycleGAN [26], which proposed a way for image translation while using unpaired images. The works such as GDWCT [27] and UNIT [28] perform similarly to CycleGAN while considering special characteristics like clear edges and high-level abstraction.

2.3. Image-to-image translation

Image-to-image translation is another technique to perform facial style transfer [29,30]. Such a method relies on the bi-directional mapping between the style domain and the facial image [26]. One of the seminal works on facial style transfer was the cartoonGAN [6], which proposed using content loss to preserve facial characteristics while performing the style transfer but was limited to a single style.

Some studies use attention mechanisms to find appearance discrepancies between key regions [11], while others use shared discriminator layers to extract features common in both domains [31]. The study [32] learns style and content image features to generate caricature-style images through facial deformation and image warping, respectively. The problem with bi-directional mapping is that it requires a long training time and generates low-resolution images [16]. Olivier et al. [33] focused on implementing an image-to-image translation method that can adopt 3D facial geometry to neutralize expressive faces. Subsequently, the method uses SpiralNet++ and FUNIT to deform and blendshapes on facial images for style transfer. They referred to their method as FaceTuneGAN. Chen et al. [23] extended the work of AnimeGAN [2] and proposed AnimationGAN thet uses bottlenecks in the residual network and hybrid attention mechanism to improve upon the toonification of the facial images. However, it is still limited to the style transfer of a single style in an implicit manner. Liu et al. [34], and Melnik et al. [35] conducted an extensive survey that highlights the studies performed for editing and generating faces using StyleGAN and manipulation of facial attributes. Both studies extensively layout the tasks existing works perform to manipulate facial attributes using image-to-image translation methods.

2.4. StyleGAN

Recently, StyleGAN was proposed to generate high-resolution images while performing facial style transfer [9]. Since the inception of StyleGAN, several studies have considered fine-tuning it to generate plausible results with limited data. The study [36] used fine-tuned StyleGAN to generate cartoon faces by extracting latent embedding and fine-tuned the model to perform semantic alignment for toonifying the facial image. Some works extended the toonifying approach by training the model on extremely limited data [37], efficient use of latent code [38], and embedding acceleration [39]. The study [40] introduced the use of exemplar style images instead of features extracted from fine-resolution-layer to generate better facial style transfer images. However, without valid supervision, the model alignment gets weakened, resulting in an efficient color transfer but does not perform well while preserving structural information. Some studies have proposed the use of cascaded StyleGANs, i.e. using the first StyleGAN to extract style codes and the second to generate the stylized image. These studies produce fixed-size images that hinder the generation of dynamic faces [15]. The StyleGAN3 [41] was proposed to solve the problem of unaligned faces, but it is still limited to the fixed-size image. Recently, a study [16] proposed DualStyleGAN that uses an extrinsic style path to overcome the model alignment problem and prior facial destylization. Our work is different from the aforementioned ones, as first, it uses pretrained encoders to generate base and secondary style images, and second, aims to preserve the facial structures and details. A comparison of the results generated using the proposed work with the existing ones is shown in Fig. 1.

2.5. GAN inversion framework

Recently, with the rise of generative AI techniques, many works have proposed the manipulation of facial attributes in order to embedded style transfer through GAN inversion frameworks. Liu et al. [42] proposed the use of face swapping and regional GAN inversion to perform style fusion. The method leverages the use of styleGAN but does not intrinsically applied style transfer to the base facial image, rather focuses on the swapping of any given image to the reference image. Lan et al. [43] proposed a self-supervised learning-based strategy to reconstruct 3D shapes and textures from a single 2D image. After a faithful 3D reconstruction of the 2D shape, facial style transfer is proposed through global latent codes. The style transfer is not performed to add effects in this study but to manipulate the facial attributes. Zheng et al. [44] proposed a new joint loss function for a generative adversarial based network that employs cross-fusion attention for style transfer. Their proposed network employed frequency domain loss to incorporate the contextual information between high-level and low-level features for reconstructing the image. Subsequently, the method will use the features to manipulate the image characteristics for the style transfer. Their proposed method works well but does not preserve the facial features as proposed in this method. Ren et al. [45] also proposed a cross-modal decoding framework by leveraging styleGAN to reconstruct faces from fMRI data. The method uses multi-level visual information from brain singles and extracts high-level features to be processed by transformation blocks and multi-stage refinement method in order to construct faces. Although the styleGAN have been employed in this study but more focus is redirected towards the use of fMRI data for reconstructing faces rather than performing facial style transfer.

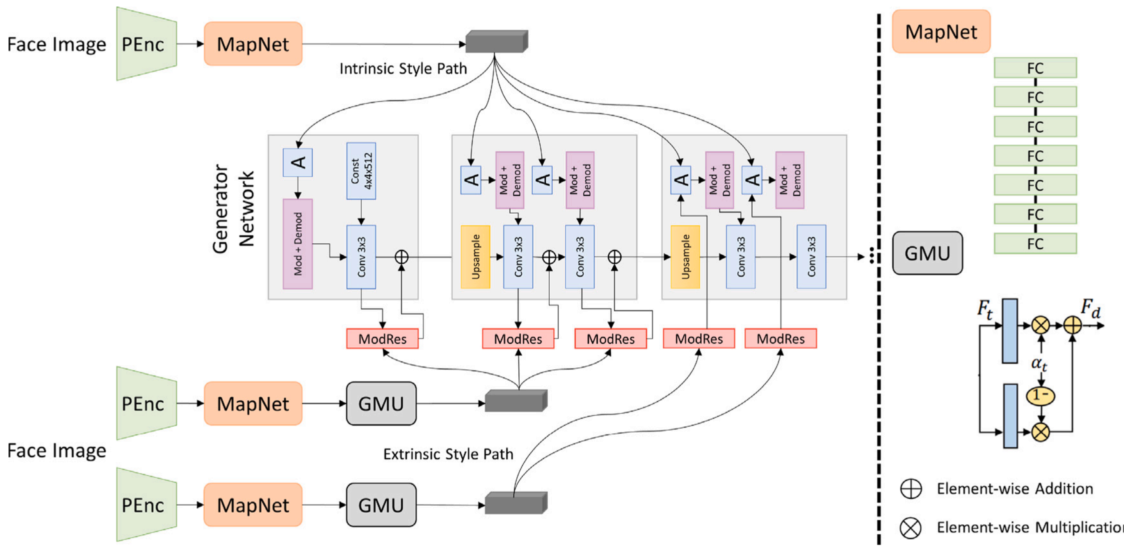

Fig. 2. The proposed FISTNet network architecture.

Peng et al. [46] proposed a dual branch style encoder, interpretable semantic generative adversarial networks (ISFB-GAN), that use inversion techniques to reconstruct facial geometry for manipulating appearance styles. This work focuses on preserving the background information and beautifying the facial characteristics like skin and face geometry rather than adding new styles to the faces.

2.6. Style hybrid methods

With the emergence of vision transformers and language models that can be integrated, researchers have tried to generate prompt-based stylized images. The study in [47] introduced the use of Contrastive Language-Image Pretraining (CLIP) models to stylize the image based on text-based prompts. A CLIP-based loss was introduced to optimize the input latent vector using the prompt provided by the user. The text prompts are then converted to input-agnostic directions in order to update the style space using StyleGAN. Some studies have followed upon the StyleCLIP work to integrate language models for generating stylized images. For instance, the study in [48] proposed a language-guided face animation that leverages motion and semantic information using the prompts and use StyleGAN to generate recurrent motion frames. Similarly, the study in [49] proposed DreamAnime that leverages the pretrained language models and fine-tuning on specific keywords to generate stylized anime images. The problem with such methods is that they rely on specific keywords rather than intuitive prompt languages to stylize the image. Furthermore, such methods rely highly on the manipulation strength and disentanglement threshold parameters, which must be varied every time an image is processed for stylization [47].

3. FISTNet

The goal of this work is to propose a facial style transfer that adopts pretrained networks, i.e., StyleGAN [9] and AnimeGAN [2], which would allow the users to characterize multiple styles using a single image while considering a few samples of data. The facial alignment problem is dealt with in the gated mapping unit module, while the problem of modeling multiple styles has been explicitly addressed using individual style paths. The proposed network is trained using a curriculum learning strategy to perform stable conditional fine-tuning. The details for each block are presented in the subsequent subsections.

3.1. Intrinsic style path

We use the transfer learning approach to train the Transformer-based StyleGAN model for our intrinsic style transfer path as shown in Fig. 2. The reasons for opting transfer learning approach are twofold: the first is to retain the facial structures and details, and the second is to work with limited data for training. Therefore, performing the fusion of pretrained StyleGAN and AnimeGAN for fine-tuning is an efficient approach to generate high-quality images. We adopt StyleGAN trained on the FFHQ dataset [9] as the pretrained encoder ($PEnc_{SG}$). Fine-tuning is performed using structural loss between the pretrained and transferred models. The formulated loss function is shown in Eq. (1).

$$loss_{suc} = \frac{1}{K}\sum_{k=1}^{K} \|\mathbb{G}^{k}_{base}(I) - \mathbb{G}^{k}_{tl}(I)\|^2 \tag{1}$$

where k is the index of the StyleGAN block and $\mathbb{G}$ represent the base and transfer generator model, respectively. We also use adversarial loss to fine-tune the StyleGAN formulated in Eq. (2).

$$loss_{adv} = \mathbb{E}_{I\sim p_I}[log(1-\mathbb{D}(\mathbb{G}_{tl}(I)))] + \mathbb{E}_{i\sim p_{data}}[log(\mathbb{D}(i))] \tag{2}$$

The fine-tuning of the StyleGAN module is performed by combining both of the aforementioned losses, i.e. $loss_{isp} = loss_{suc} + loss_{adv}$. However, even with fine-tuning, some artifacts are introduced that can affect the facial structure and expressions. In this regard, it is necessary to extract semantic features from the latent space of StyleGAN. The study [50] found the semantic features using a closed-form factorization, such as $\mathbb{G}(I) \triangleq \mathfrak{b} + \mathfrak{w}I = \mathfrak{z}$, where $\mathfrak{b}$ and $\mathfrak{w}$ are biases and weights, respectively. The same study suggested that the latent codes can be manipulated by using the formulation $\mathbb{G}(I + \sigma y) = \mathfrak{z} + \sigma\mathfrak{w}y$, where manipulation intensity is represented by the notation σ and semantic attribute representation in latent space is represented by y. It was discussed in the same study that the weights in mapping network contains information related to image variations, therefore, the discovery of semantic attributes can be performed by decomposing the weights, accordingly. The decomposition can be performed using the optimization function shown in Eq. (3).

$$y^* = \underset{y\in\mathbb{R}^d:\, y^T y=1}{\text{argmax}} \|\mathfrak{w}y\|_2^2 \tag{3}$$

The function $\|.\|_2$ represents the L_2 norm. Another problem that occurs during the manipulation of latent code through the mapping network is the change in facial identity, which might hinder the motivation of this work. Therefore, we use an identity loss that optimizes

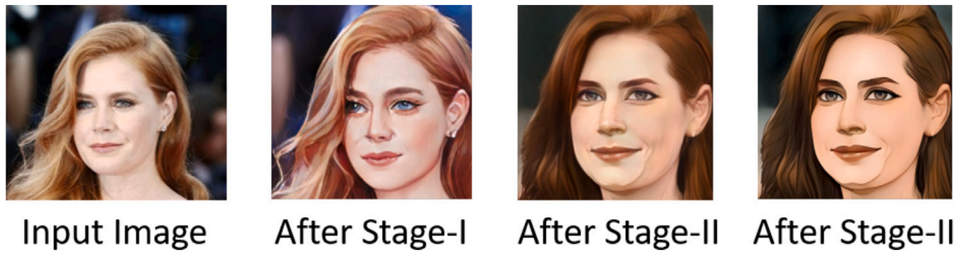

Fig. 3. Results of FISTNet after each fine-tuning stage.

the offset, i.e., σy^*, along with a pretrained recognition network [51] that is responsible for facial identity regularization. The formulation for the aforementioned optimization problem is shown in Eq. (4).

$$loss_{id} = \|f(\mathbb{G}_{base}(\sigma y)) - f(\mathbb{G}_{base}(\sigma y + \sigma y^*))\|^2 \tag{4}$$

There might be a condition that yields $\sigma y^* \rightarrow 0$. Such a condition may result in insignificant optimization. To overcome the said issue, we also add a constraint that aims to restore the low-level features between the manipulated and input modality. In order to do so, we use segmentation methods [52] to extract the facial area, represented by f_{seg}. The constraint is defined as a regularization function in Eq. (5).

$$loss_{seg} = \|f_{seg}(\mathbb{G}_{base}(\sigma y)) - f_{seg}(\mathbb{G}_{base}(\sigma y + \sigma y^*))\|^2 \tag{5}$$

The combined optimization function can then be given as $[loss_{mapnet} = loss_{id} - \alpha_{seg} loss_{seg}]$, where α_{seg} represents the hyperparameter to control the trade-off, accordingly.

3.2. Extrinsic style path

Our external style path simply performs the fusion of two pretrained networks, i.e. AnimeGANv2 trained on portrait and celeb distill styles to add semantic cues such as facial color, eye styles, hair color, and shapes, accordingly. The celeb distill style refers to the knowledge distillation based style transfer using celebA dataset. Different styles including arcane, Disney, and others can be leveraged by performing the fusion of pretrained networks to produce diverse results as shown in Fig. 1. We use the same optimization function for MapNet as discussed in the prior section to retain facial details and structure, respectively. The features extracted using MapNet will then undergo the Gated Mapping Unit (GMU) adopted from the study [53]. However, unlike the study, we only use domain-specific features instead of combining them with group-specific features. The domain-specific features are the ones extracted from the encoder dedicated to the reference image, while the group-specific features are the ones related to multiple styles. Let us denote the feature extracted from MapNet as F_t and domain-specific features as F_d. The domain-specific features can be formulated as shown in Eq. (6).

$$F_d = \gamma \cdot \varsigma_{d0}(F_t) + (1-\gamma) \cdot \varsigma_{d1}(F_t) \tag{6}$$

where ς_d refers to the branches in the GMU unit and γ represents the control factor and comprises of dichotomous values $0, 1$. The GMU outputs the style code from each of the pretrained AnimeGANv2 network. The last component of our architecture is the generator network that follows StyleGAN. Formally, the style transfer is achieved by $\mathbb{G}(PEnc_1(I), PEnc_2(I), \mathbb{W})$, where $PEnc_1$ and $PEnc_2$ represent pretrained encoders from AnimeGANv2 and $\mathbb{W}$ refers to the weight vector having size $\mathbb{R}^{18}$.

The coarse-resolution layers learn to map the high-level shape style transfer while the fine-resolution layers learn the low-level color style, respectively. We follow the training strategy from DualStyleGAN except for a few changes. The first change is that instead of using a single network in an extrinsic style path we use two pretrained encoders from AnimeGANv2 along with the GMU to extract style codes. The second difference is the use of modulative residual blocks (ModRes) throughout the network instead of using color transform block. The ModRes comprises Residual blocks (ResBlock) [54] and Adaptive Instance Normalization (AdaIn) [55] blocks for simulating changes and style conditioning, respectively.

3.3. Fine-tuning process

The fine-tuning scheme for style fusion in the proposed work follows the curriculum learning approach [56]. The fine-tuning process consists of three stages i.e., color, structure, and style transfer. The stage-I strives for color conditioning from one of the pretrained style networks by model initialization technique, such that the convolutional filters in ModRes from the style code generated by $PEnc_1$ are initialized with zeros, and the filters in ModRes from $PEnc_2$ are initialized with identity matrices. This stage generates a mixed style by transferring colors from the encoded styles while preserving the facial structures from the intrinsic style path. It was observed that the first stage kind of tones up the skin and hair color. It generates more of a bright-toned cartoonish portrait. We believe that the main reason for the said results is the use of AnimeGANv2 with a celeb-distill style, respectively.

The fine-tuning in stage-II fuses mid-level styles that manipulate facial structures in terms of make-up, eyes, and hair color. Random latent codes $\mathfrak{c}_1, \mathfrak{c}_2$ are drawn to approximate the style fusion through $\mathbb{G}(\mathfrak{c}_1, \mathfrak{c}_2, 1)$ with perceptual loss. The fine-tuning process is performed on $g(\mathfrak{c}_\mathfrak{l})$ and random latent codes, while gradually decreasing the layers $\mathfrak{l}$, accordingly. The concatenation of the latent codes with respect to layers is denoted by $\mathfrak{c}_\mathfrak{l}^+$ and g represents the output from $\mathbb{G}$ when $\mathbb{W} = 0$. The objective function for fine-tuning is shown in Eq. (7).

$$\min_{\mathbb{G}} \max_{\mathbb{D}} loss_{pl}(\mathbb{G}(\mathfrak{c}_1, \mathfrak{c}_2, 1), g(\mathfrak{c}_\mathfrak{l}^+)) \cdot \alpha_{pl} + loss_{adv} \cdot \alpha_{adv} \tag{7}$$

The objective function will be able to learn more structural styles along with colors by decreasing the layers, accordingly.

The third stage employs the identity loss and style loss (comprising of feature matching loss [55] and contextual loss [57]) which adds the abstractive styles from $PEnc_2$ and L_2 regularization on the ModRes blocks, accordingly. The regularizations help to preserve the facial structure while making the weight values for the residual features close to 0. The full objective function is formulated in Eq. (8).

$$\min_{\mathbb{G}} \max_{\mathbb{D}} loss_{content} + loss_{style} + loss_{pl} \cdot \alpha_{pl} + loss_{adv} \cdot \alpha_{adv} \tag{8}$$

Fig. 3 shows an example of stage-wise generation.

4. Experiments and results

4.1. Datasets

One of the motivations for conducting this study is to generate facial style transfer with a limited number of training data. We perform the fusion of the pretrained AnimeGANv2 to generate styles in face portraits and celeb distill styles. However, we also conducted experiments

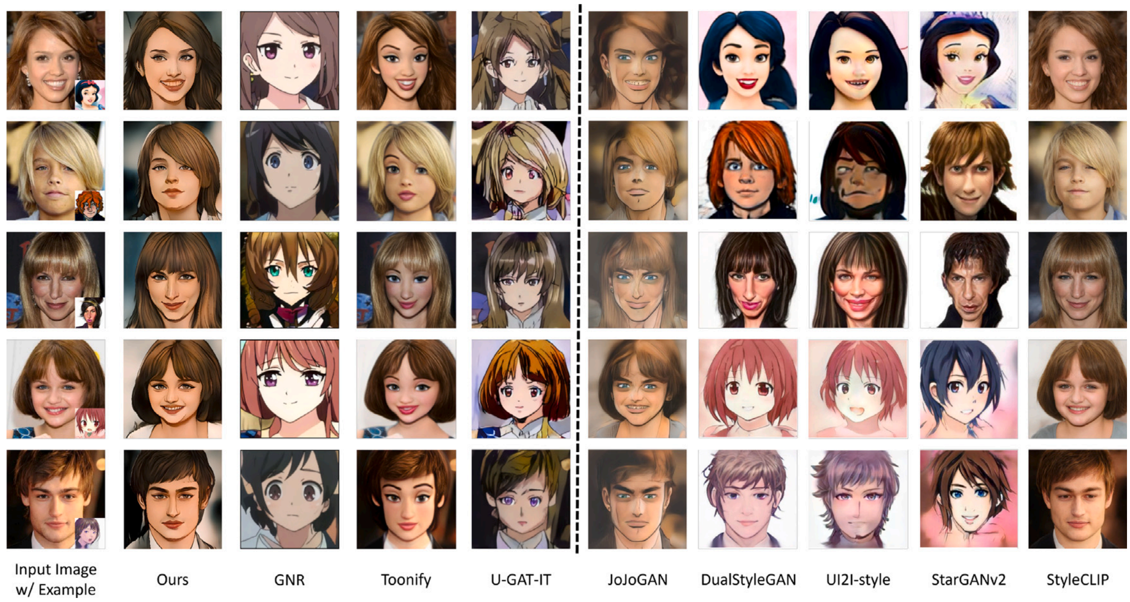

Fig. 4. Qualitative Comparison for style transfer with state-of-the-art works. The right side of the line shows the results from existing methods that require an input style while the left side of the line compares with the works that require exemplar images.

while considering the combinations of pretrained Disney and arcane styles to see the effect available with AnimeGANv3. Meanwhile, for the intrinsic style path, we use 317 images from the cartoon dataset [36], accordingly. The experiments are conducted on CelebA-HQ [58] dataset to conduct a fair comparison with state-of-the-art works.

4.2. Implementation details for intrinsic style path

As mentioned, we use the StyleGAN model to fine-tune the intrinsic style path. Our implementation is based on the PyTorch framework. All the images were resized to 256 × 256 resolution. The structural loss was applied on the first 2 blocks of the StyleGAN model. We used the same training strategy as [16], with an initial learning rate of 0.05. To obtain σy^{*}, we use 10 iterations for the optimization of each sample. The pretrained face embedding model [51], which uses ResNet18 [54] has been leveraged for $loss_{id}$ and $loss_{seg}$. The value of the regularization parameter α_{seg}, is set to 0.2, accordingly. All the aforementioned parameters are selected based on the experiential study while model training process.

4.3. Implementation details for extrinsic style path

The extrinsic style path uses pretrained encoders from AnimeGANv2 followed by the GMU and generator network. The GMU comprises fully connected domain-specific layers. We use two domain-specific layers in GMU, accordingly. The generator network comprises 18 modulative residuals (ModRes) blocks. There is one 4×4, two $8 \times 8, 16 \times 16, 32 \times 32, 64 \times 64, 128 \times 128, 256 \times 256, 512 \times 512$, and 1024×1024 convolution layers and one 1024×1024 to RGB layer in the generator network, respectively. The hyperparameters are the same as DualStyle-GAN study. We train the 5th layer for 2 k iterations while the 6th and 7th layers for 200 iterations each.

4.4. Comparison with state-of-the-art methods with user study

A qualitative comparison is presented with eight state-of-the-art methods performed in Fig. 4. Some of them require an example image for style transfer such as DualStyleGAN, UI2I style [40], JoJoGAN [59] and StarGAN v2 [60], while other methods either rely on style category or generate a specific style including GNR [32], Toonify [15], and U-GAT-IT [11], respectively. We have also included a method that leverages pretrained language models to stylize the image, i.e. Style-CLIP [47]. The results for the aforementioned are generated using their provided codes or APIs (Huggingface or grade io), accordingly. Methods such as U-GAT-IT, Toonify, and GNR do not consider multiple image styles, rather they only rely on domain level to learn the style transfer. JoJoGAN and StarGANv2 ignore the facial structures and details, which results in a kind of overfitting anime or cartoon style. DualStyleGAN and UI2I capture some structural details and color information, but they can need an example image which sometimes generates bad results. When using StyleCLIP, we added a prompt “A cartoon anime character” to the target string. We also set the manipulation strength and disentanglement threshold to 7.41 and 0.12, respectively. As it can be observed that the image does not stylize with respect to the cartoon or anime characters, rather it improves the quality of the image. Furthermore, none of the results generate a blend of artistic transfer while keeping the facial structure intact. Our proposed work FISTNet generates the best results without providing any example image. However, other styles such as arcane and Disney can be generated by replacing the pretrained encoders in extrinsic style paths, respectively. It should also be noted the proposed work considers the least amount of data for training.

We also provide a quantitative evaluation by conducting a user study of 67 subjects. Subjects were invited to rate the results generated by the aforementioned methods on three characteristics, which are the preservation of facial details, quality of generated image, and style transfer results. Ten examples were provided for each method during the evaluation. Furthermore, the subjects were allowed to vary three different examples for methods, including DualStyleGAN, UI2I-style, and StarGANv2. Average preference scores from the subjects

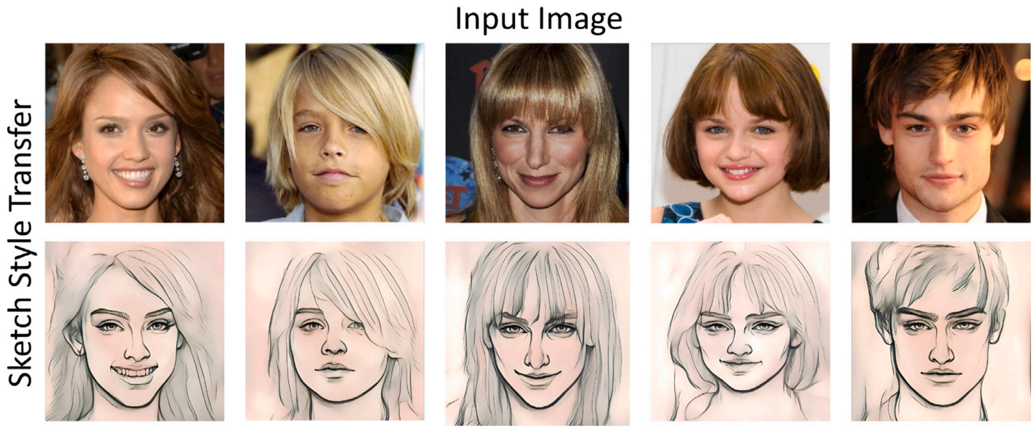

Fig. 5. Sketch style transfer using FISTNet.

are reported in Table 1. The proposed method scores best among all the existing methods for facial preservation, image quality, and style quality, respectively. It was also observed that the Toonify approach yields the second-best results.

Table 1
User scores for varying characteristics. The best scores are represented in bold. *FP → Facial Preservation, IQ → Image Quality, SQ → Style Quality, Avg → Average.

Method	FP	IQ	SQ	Avg
UI2I-style	0.07	0.04	0.03	0.047
StarGANv2	0.09	0.07	0.03	0.063
GNR	0.06	0.07	0.04	0.057
U-GAT-IT	0.06	0.08	0.03	0.057
DualStyleGAN	0.13	0.15	0.18	0.153
JoJoGAN	0.17	0.05	0.04	0.087
FISTNet	**0.24**	**0.33**	**0.38**	**0.316**

4.5. Comparison with state-of-the-art methods with FID

The Frechet Inception Distance (FID) [61] is extensively used in facial style transfer studies to evaluate the diversity and quality of images generated through population statistics. Studies such as [59,62] have also used FID for evaluating style mappers. The process for evaluating the facial style transfer quality through FID is as follows:

- Select a reference image and perform stylization.
- For generalized standard, its better to perform one-shot stylization.
- Compute FID between style and the generated result after applying style transfer.
- Compute FID using the testset.

The process is also compliant with the existing works [59]. For fair comparison, we replaced the pretrained encoder for the celeb-distill style with the sketch one and generated the images, accordingly. The results with sketch generative style using FISTNet are shown in Fig. 5, while the FID scores are reported in Table 2, respectively. The lower the FID score, the better facial structures are preserved. The proposed FISTNet achieves the lowest FID score after StyleCLIP [47] and Ojha et al. [62] in comparison to all the state-of-the-art works. The StyleCLIP was given the prompt of "generate a sketch of this face" to stylize the image, however, the method returned a colored image with similar face but with translated facial features as shown in Fig. 4, which illustrates and justifies the lowest FID score. The BlendGAN [63] achieves a lower FID than many state-of-the-art works, however, it was noticed that BlendGAN does not perform stylization well on the images. Meanwhile, the lowest FID was achieved by Ojha et al. [62], but the results showed that strong distortions were imposed on the faces. The observation concerning facial distortions for the method proposed in Ojha et al. [62] is compliant with the study [59].

Table 2
Comparative analysis with state-of-the-art works using FID score.

Method	FID
U-GAT-IT	183.2
GNR	167.4
BlendGAN	94.7
JoJoGAN	107.6
Ojha et al.	74.5
DualStyleGAN	155.2
Toonify	79.7
StyleCLIP	12.2
Intrinsic Style Path	95.9
Extrinsic Style Path	76.8
FISTNet	78.9

4.6. Ablation study

In this subsection, we provide the details regarding the component-wise performance of the proposed FISTNet. As mentioned, we employ an intrinsic and an extrinsic style path for the facial stylization. We provide a qualitative illustration of the ablation study in Fig. 3. It can be visualized that the stage-I (intrinsic style path) stylizes the image based on the pretrained encoder, thus transforms the facial structures and hair color to some extent. While the stage-II (extrinsic style path) preserves the facial structure and adds diverse style, it reduces the quality of generated images. Finally, the Stage-III (FISTNet) preserves the facial details from Stage-II while improving upon the quality of the image. The same is also reflected in the quantitative results shown in Table 2. The FID score is lesser in comparison to FISTNet, but is significantly lesser than the one generated using intrinsic style path only.

4.7. Comparison with state-of-the-art on multiple styles

We further compare our approach with state-of-the-art works such as Toonify, AnimeGANv3,[2] and DualStyleGAN for Cartoon, Arcane, and Disney styles. The Toonify and the proposed approach only need to select the style type, while the DualStyleGAN also needs an exemplar image. In this regard, we selected the image index of 142, 40, and

[2] https://github.com/TachibanaYoshino/AnimeGANv3

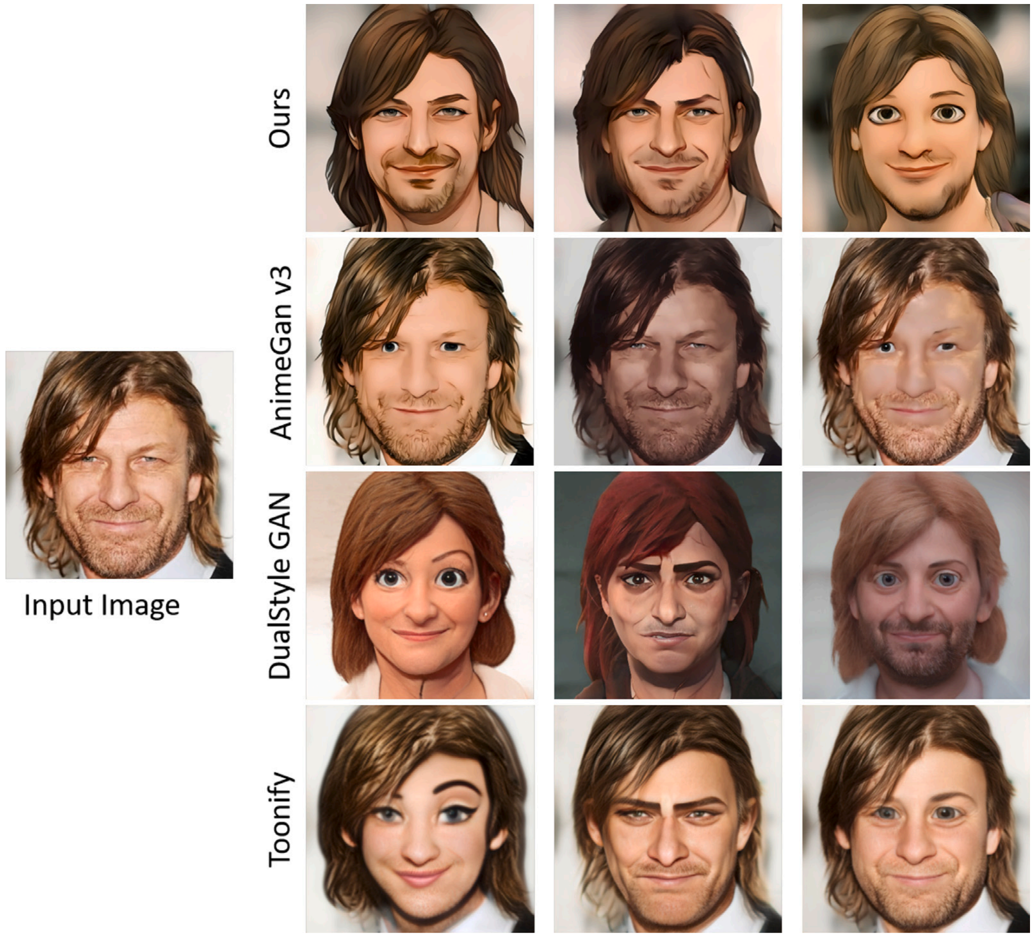

Fig. 6. Qualitative comparison on celebrity images with state-of-the-art works using multiple styles.

67, for cartoon, arcane, and Pixar (Disney) styles, respectively. The qualitative comparison is shown in Fig. 6. The results illustrate the weakness of AnimeGANv3 when tried with Disney style as it adds random artifacts. The Toonify approach retains the facial features better when applying the Disney style in comparison to the proposed work. However, the yielded image quality is on the downside as it is a little bit blurry. In summary, the proposed results yield high-quality images with better visual quality in comparison to the aforementioned existing works across multiple styles. It should also be noted that our work utilizes less data for training in comparison to the existing works.

We repeated the same experiment to test and compare the style transfer results on a randomly generated face.[3] Similar to the previous experiment, we selected exemplar samples for generating images using DualStyleGAN. The image index for Cartoon, Arcane, and Disney stiles are 183, 10, and 118, respectively. The results are shown in Fig. 7. The results highlight the limitation of AnimeGANv3 with randomly generated and sideways faces. Toonify generates good results, but the facial style transfer is limited in comparison to the results on celebrity images. The proposed work not only overcomes the problems associated with AnimeGANv3 but also successfully transfers diverse styles to the randomly generated image.

[3] https://this-person-does-not-exist.com/en

4.8. Limitations

A few limitations of the work are quite similar to AnimeGANv3 and DualStyleGAN, i.e., the facial features and structure are retained quite well. However, artifacts are introduced due to the props used with images such as caps, hats, glasses, and so forth. We believe the problem is due to the limited data used during the training phase, a potential reason that warrants further investigation. One such example is shown in Fig. 8. Another problem in Arcane style is shown in Fig. 7, where some artifacts are generated under the eyes when the facial image is sideways. The problem is quite consistent with AnimeGANv3, which we believe is due to the adoption of the pretrained network. Another limitation observed in Fig. 6 while performing sketch style transfer is that it does not well model the smile, especially the toothy smile.

5. Conclusion and future work

This paper proposes FISTNet, which performs style conditioning on facial images while preserving facial characteristics and structure. Inspired by DualStyleGAN, we use intrinsic and extrinsic style paths with StyleGAN to generate diverse stylized facial images. However, in contrast with DualStyleGAN, we use existing pretrained networks that not only help add unique stylization but also eliminate the problem of large-scale training data. We also replaced the transform coding blocks responsible for encoding and decoding the image with modulative residual blocks, which help generate high-image and style-quality

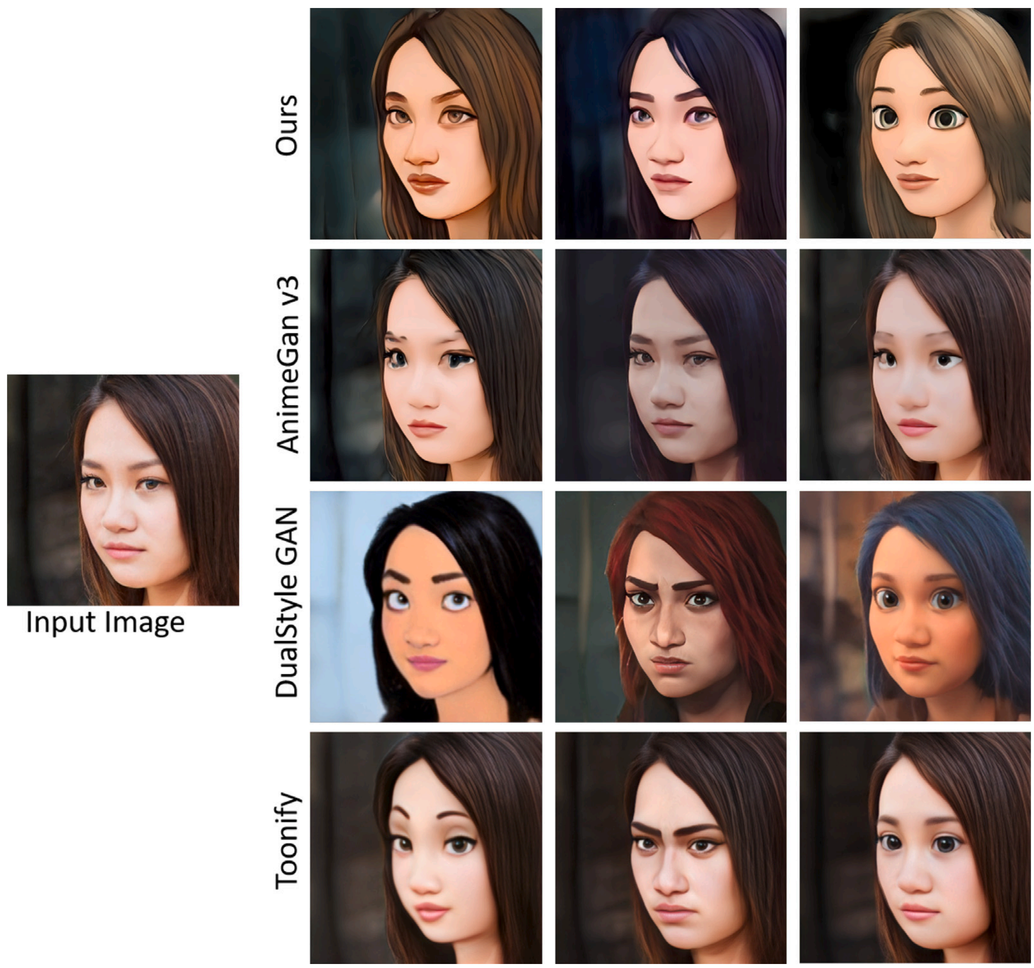

Fig. 7. Qualitative comparison on randomly generated images (style-in-the-wild experiment) with state-of-the-art works using multiple styles.

Fig. 8. Limitations of FISTNet. The artifacts are introduced when facial props are used in the input images.

images, respectively. We also show that the FISTNet networks can generate multiple styles, such as Arcane and Disney, by using existing pretrained networks. The proposed work can help emerge applications such as eXtended Reality (XR), Metaverse, avatar generation, and art generation for non-fungible tokens (NFTs). In the future, we would like to explore more diverse styles and add an extra path for the style transfer to observe the results. We also intend to add a localized stylization approach, allowing users to fuse multiple styles in a single image. Furthermore, AI tools are abundant for image generation and stylization. Therefore, it is necessary to build a tool that can quantitatively assess the quality of stylization and similarity. We aim to develop an evaluation metric that would help quantitatively assess the image generation method's effectiveness.

CRediT authorship contribution statement

Sunder Ali Khowaja: Writing – original draft, Validation, Methodology, Formal analysis, Conceptualization. **Lewis Nkenyereye:** Writing – review & editing, Visualization, Resources, Investigation, Data curation. **Ghulam Mujtaba:** Writing – original draft, Software, Investigation, Conceptualization. **Ik Hyun Lee:** Writing – review & editing, Supervision, Project administration, Data curation. **Giancarlo Fortino:** Writing – review & editing, Validation, Resources, Investigation, Conceptualization. **Kapal Dev:** Writing – review & editing, Supervision, Project administration, Formal analysis, Data curation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Sunder Ali Khowaja reports was provided by University of Sindh. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We have used publicly available datasets in this study.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1F1A1050022), and the Priority Research Centers Program (2017R1A6A1A03015562) funded by the Ministry of Education.

References

[1] M. Zawish, F.A. Dharejo, S.A. Khowaja, S. Raza, S. Davy, K. Dev, P. Bellavista, AI and 6G into the metaverse: Fundamentals, challenges and future research trends, IEEE Open J. Commun. Soc. 5 (2024) 730–778, http://dx.doi.org/10.1109/OJCOMS.2024.3349465.

[2] J. Chen, G. Liu, X. Chen, AnimeGAN: A novel lightweight GAN for photo animation, in: Artificial Intelligence Algorithms and Applications, Springer, Singapore, 2020, pp. 242–256.

[3] Y. Shi, D. Deb, A.K. Jain, Warpgan: Automatic caricature generation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2019, pp. 10762–10771.

[4] H. Su, J. Niu, X. Liu, Q. Li, J. Cui, J. Wan, Mangagan: Unpaired photo-to-manga translation based on the methodology of manga drawing, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35 (3) (2021) 2611–2619, http://dx.doi.org/10.1609/aaai.v35i3.16364.

[5] R. Yi, Y.J. Liu, Y.K. Lai, P.L. Rosin, Unpaired portrait drawing generation via asymmetric cycle mapping, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2020, pp. 8217–8225.

[6] Y. Chen, Y.K. Lai, Y.J. Liu, Cartoongan: Generative adversarial networks for photo cartoonization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2018, pp. 9465–9474.

[7] X. Wang, J. Yu, Learning to cartoonize using white-box cartoon representations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2020, pp. 8090–8099.

[8] W. Jang, G. Ju, Y. Jung, J. Yang, X. Tong, S. Lee, StyleCariGAN: caricature generation via stylegan feature map modulation, ACM Trans. Graph. 40 (4) (2021) 1–16, http://dx.doi.org/10.1145/3450626.3459860.

[9] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2019, pp. 4401–4410.

[10] S.A. Khowaja, S. Almakdi, M.A. Memon, P. Khuwaja, A. Sulaiman, A. Alqahtani, A. Shaikh, A. Alghamdi, Extending user control for image stylization using hierarchical style transfer networks, Heliyon 10 (5) (2024) e27012.

[11] J. Kim, M. Kim, H. Kang, K.H. Lee, U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation, in: International Conference on Learning Representations, 2020, pp. 1–19.

[12] Y. Shu, R. Yi, M. Xia, Z. Ye, W. Zhao, Y. Chen, Y.K. Lai, Y.J. Liu, GAN-based multi-style photo cartoonization, IEEE Trans. Vis. Comput. Graphics 28 (10) (2022) http://dx.doi.org/10.1109/TVCG.2021.3067201.

[13] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, K. Aberman, DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 22500–22510, http://dx.doi.org/10.1109/CVPR52729.2023.02155.

[14] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A.H. Bermani, G. Chechik, D. Cohen-or, An image is worth one word: Personalizing text-to-image generation using textual inversion, in: International Conference on Learning Representations, ICLR, 2023, pp. 1–14.

[15] S. Yang, L. Jiang, Z. Liu, C.C. Loy, VToonify: Controllable high-resolution portrait video style transfer, ACM Trans. Graph. 41 (6) (2022) 1–15, http://dx.doi.org/10.1145/3450626.3459860.

[16] S. Yang, L. Jiang, Z. Liu, C.C. Loy, Pastiche master: Exemplar-based high-resolution portrait style transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 7693–7702.

[17] D. Liu, M. Fisher, A. Hertzmann, E. Kalogerakis, Neural strokes: Stylized line drawing of 3D shapes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 14204–14213.

[18] Z. Wang, S. Qiu, N. Feng, H. Rushmeier, L. McMillan, J. Dorsey, Tracing versus freehand for evaluating computer-generated drawings, ACM Trans. Graph. 40 (4) (2021) 1–12.

[19] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015, pp. 1–14.

[20] L.A. Gatys, A.S. Ecker, M. Bethge, Image style transfer using convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 2414–2423.

[21] C. Li, M. Wand, Combining Markov random fields and convolutional neural networks for image synthesis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 2479–2486.

[22] J. Liao, Y. Yao, L. Yuan, G. Hua, S.B. Kang, Visual attribute transfer through deep image analogy, ACM Trans. Graph. 36 (4) (2017) 1–15.

[23] M. Chen, H. Dai, S. Wei, Z. Hu, Linear-ResNet GAN-based anime style transfer of face images, Signal, Image Video Process. 17 (2023) 3237–3245, http://dx.doi.org/10.1007/s11760-023-02553-8.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.

[25] P. Isola, J.Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017, pp. 1125–1134.

[26] J.Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision, IEEE, 2017, pp. 2223–2232.

[27] W. Cho, S. Choi, D.K. Park, I. Shin, J. Choo, Image-to-image translation via group-wise deep whitening-and-coloring transformations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10639–10647.

[28] M.Y. Liu, T. Breuel, J. Kautz, Unsupervised image-to-image translation networks, in: Advances in Neural Information Processing Systems, 2017, pp. 1–9.

[29] X. Shao, W. Zhang, SPatchGAN: A statistical feature based discriminator for unsupervised image-to-image translation, in: Proceedings of the IEEE Conference on Computer Vision, IEEE, 2021, pp. 6546–6555.

[30] Y. Zhao, R. Wu, H. Dong, Unpaired imageto-image translation using adversarial consistency loss, in: Proceedings of the European Conference on Computer Vision, Springer, 2020, pp. 800–815.

[31] B. Li, Y. Zhu, Y. Wang, C.W. Lin, B. Ghanem, L. Shen, AniGAN: Style-guided generative adversarial networks for unsupervised anime face generation, IEEE Trans. Multimed. 24 (2021) 4077–4091.

[32] M.J. Chong, D. Forsyth, GANs N'Roses: Stable, controllable, diverse image to image translation, 2021, pp. 1–9, arXiv preprint arXiv:2106.06561.

[33] N. Olivier, K. Baert, F. Danieau, F. Multon, Q. Avril, FaceTuneGAN: Face autoencoder for convolutional expression transfer using generative adversarial networks, Comput. Graph. 110 (2023) 69–85, http://dx.doi.org/10.1016/j.cag.2022.12.004.

[34] Y. Liu, Q. Li, Q. Deng, Z. Sun, M.H. Yang, GAN-based facial attribute manipulation, IEEE Trans. Pattern Anal. Mach. Intell. 45 (12) (2023) 14590–14610, http://dx.doi.org/10.1109/TPAMI.2023.3298868.

[35] A. Melnik, M. Miasayedzenkau, D. Makaravets, D. Pirshtuk, E. Akbulut, D. Holzmann, T. Renusch, G. Reichert, H. Ritter, Face generation and editing with stylegan: A survey, IEEE Trans. Pattern Anal. Mach. Intell. (2024) 1–21, http://dx.doi.org/10.1109/TPAMI.2024.3350004, Early Access.

[36] J.M. Pinkey, D. Adler, Resolution dependent GAN interpolation for controllable image synthesis between domains, 2020, pp. 1–7, arXiv preprint arXiv:2010.05334.

[37] U. Ojha, Y. Li, J. Lu, A.A. Efros, Fewshot image generation via cross-domain correspondence, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2021, pp. 10743–10752.

[38] G. Song, L. Luo, J. Liu, W.C. Ma, C. Lai, C. Zheng, T.J. Cham, Agilegan: stylizing portraits by inversion-consistent transfer learning, ACM Trans. Graph. 40 (4) (2021) 1–13.

[39] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzam, Y. Azar, S. Shapiro, D. Cohen-Or, Encoding in style: a stylegan encoder for image-to-image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2021, pp. 2287–2296.

[40] S. Kwong, J. Huang, J. Liao, Unsupervised image-to-image translation via pre-trained stylegan2 network, IEEE Trans. Multimed. 24 (2021) http://dx.doi.org/10.1109/TMM.2021.3065230.

[41] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, T. Aila, Alias-free generative adversarial networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2021, pp. 852–863.

[42] Z. Liu, M. Li, Y. Zhang, C. Wang, Q. Zhang, J. Wang, Y. Nie, Fine-grained face swaping via regional GAN inversion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 8578–8587.

[43] Y. Lan, X. Meng, S. Yang, C.C. Loy, B. Dai, Self-supervised geometry-aware encoder for style-based 3D GAN inversion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 20940–20949.

[44] X. Zheng, X. Yang, Q. Zhao, H. Zhang, X. He, J. Zhang, X. Zhang, CFA-gan: Cross fusion attention and frequency loss for image style transfer, Displays 81 (2024) 102588, http://dx.doi.org/10.1016/j.displa.2023.102588.

[45] Z. Ren, J. Li, L. Wu, X. Xue, X. Li, F. Yang, Z. Jiao, X. Gao, Brain-driven facial image reconstruction via stylegan inversion with improved identity consistency, Pattern Recognit. 150 (2024) 110331, http://dx.doi.org/10.1016/j.patcog.2024.110331.

[46] T. Peng, M. Li, F. Chen, Y. Xu, Y. Xie, Y. Sun, D. Zhang, ISFB-GAN: Interpretable semantic face beautification with generative adversarial networks, Expert Syst. Appl. 236 (2024) 121131, http://dx.doi.org/10.1016/j.eswa.2023.121131.

[47] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, D. Lischinski, StyleCLIP: Text-driven manipulation of StyleGAN imagery, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 2065–2074, http://dx.doi.org/10.1109/ICCV48922.2021.00209.

[48] T. Hang, H. Yang, B. Liu, J. Fu, X. Geng, B. Guo, Language-guided face animation by recurrent StyleGAN-based generator, IEEE Trans. Multimed. 25 (2023) 9216–9227, http://dx.doi.org/10.1109/TMM.2023.3248143.

[49] C. Xu, Y. Xu, H. Zhang, X. Xu, S. He, DreamAnime: Learning style-identity textual disentanglement for anime and beyond, IEEE Trans. Vis. Comput. Graphics (2024) 1–12, http://dx.doi.org/10.1109/TVCG.2024.3397712.

[50] Y. Shen, B. Zhou, Closed-form factorization of latent semantics in gans, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 1532–1540.

[51] Y. Shi, D. Agarwal, A.K. Jain, Lifting 2D StyleGAN for 3D aware face generation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 6258–6266.

[52] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, N. Sang, BiseNet: Bilateral segmentation network for real-time semantic segmentation, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 325–341.

[53] Y. Men, Y. Yao, M. Cui, Z. Lian, X. Xie, X.S. Hua, Unpaired cartoon image synthesis via gated cycle mapping, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 3501–3510.

[54] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[55] X. Huang, S. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization, in: Proceedings of the IEEE Conference on Computer Vision, 2017, pp. 1501–1510.

[56] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriuclum learning, in: International Conference on Machine Learning, 2009, pp. 41–48.

[57] R. Mechrez, I. Talmi, L. Zelnik-Manor, The contextual loss for image transformation with non-aligned data, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 768–783.

[58] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of the International Conference on Computer Vision, 2015, pp. 3730–3738.

[59] M.J. Chong, D. Forsyth, Jojogan: One shot face stylization, in: Proceedings of the European Conference on Computer Vision, 2022, pp. 128–152.

[60] Y. Choi, Y. Uh, J. Yoo, J.W. Ha, Stargan v2: Diverse image synthesis for multiple domains, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 8188–8197.

[61] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs trained by a two time-scale update rule converge to a local nash equilibrium, in: Advances in Neural Information Processing Systems, 2017, pp. 1–12.

[62] U. Ojha, Y. Li, J. Lu, A. A.Efros, Y.J. Lee, E. Shechtman, R. Zhang, Few-shot image generation via cross domain correspondence, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 10743–10752.

[63] M. Liu, Q. Li, Z. Qin, G. Zhang, P. Wan, W. Zheng, BlendGAN: Implicitly GAN blending for arbitrary stylized face generation, in: Advances in Neural Information Processing Systems, 2021, pp. 29710–29722.