ITC 3/53

Information Technology and Control
Vol. 53 / No. 3/ 2024
pp. 960-977
DOI 10.5755/j01.itc.53.3.35213

Face Attribute Transfer Fusing Feature Enhancement and Structural Diversity Loss Function

Received 2023/09/27 | Accepted after revision 2024/03/02

HOW TO CITE: Sun, Y., Zhang, C., Yu, F., Xu, H., Pan, Q. (2024). Face Attribute Transfer Fusing Feature Enhancement and Structural Diversity Loss Function. *Information Technology and Control*, 53(3), 960-977. https://doi.org/10.5755/j01.itc.53.3.35213

Face Attribute Transfer Fusing Feature Enhancement and Structural Diversity Loss Function

Yulin Sun
School of Computer Science and Technology, Changchun University of Science and Technology, Changchun, China

Chao Zhang
National & Local Joint Engineering Research Center for Special Film Technology and Equipment, Changchun, China; School of Computer Science and Technology, Changchun University of Science and Technology, Changchun, China

Fudong Yu
Jilin Sino Agriculture Sunshine Data Co., LTD, Changchun, China

Haonan Xu, Qunqin Pan
School of Computer Science and Technology, Changchun University of Science and Technology, Changchun, China

Corresponding authors: Chao Zhang zhangchao@cust.edu.cn

In the process of face attribute transfer, non-frontal and occluded face images often suffer from low generation quality, missing facial edges, and a lack of diversity. To address these challenges, we present the FES-StarGANv2, an unsupervised multi-domain face attribute transfer network. In the feature extraction phase, we incorporate an attention-guided feature fusion module aimed at enhancing image details while preserving the overall integrity of the transferred images. Moreover, a style code extraction module is presented, refining the style code of the target domain, enhancing the learning capabilities of the generator. To further augment image diversity and authenticity, a face image optimization module and a structural diversity loss function are integrated. Experimental results reveal that, in comparison with the baseline StarGANv2, our approach achieves substantial improvements of 23% and 3.9% in FID and LPIPS metrics, respectively, attaining optimal 13 and 0.453. Notably, in terms of visual quality, significant enhancements were observed, particularly in addressing issues of low image quality and edge deficiencies. The FES-StarGANv2 approach effectively addresses the challenges associated with non-frontal and occluded facial images.

KEYWORDS: Face Attribute Transfer, Unsupervised Learning, Feature Fusion, LSTM, SSIM.

1. Introduction

Face attribute transfer is the work of image-domain to image-domain work. Specifically, learn texture features from another face image while retaining the original main content of the image to change special facial properties, such as hairstyle, skin color, age, and so on. This technique is crucial for assisting facial recognition and is frequently utilized in digital entertainment and social media. In the past, only supervised image transfer was possible. Still, the development of Generative Adversarial Networks (GANs) [5] and their derivative models have enabled unsupervised image transfer without being constrained by one-to-one matching data.

Based on the mapping relationship between the source and target picture domains, unsupervised image transfer is divided into single and multiple image domain transfers. Single image domain transfer is the process of transferring only between two image domains. For example, Perarnau et al., in [29], proposed the IcGAN, which introduces two encoders to decompose the actual image into content and attribute feature information, and generate the specified image through the CGAN [25] framework. Zhu et al., in [39], proposed the CycleGAN, which reconstructs the generated transfer image using two pairs of generator and discriminator and constrains the image content by cycle consistency loss. An essential problem of single-domain transfer is that it can only learn the relationship between a pair of different image domains, and it is difficult to deal with the transformation between multiple image domains. When dealing with multi-domain image transfer, it needs to model and train each pair of image domains separately, which increases the difficulty and cost of training.

Multi-image domain transfer is an extension of single-image domain transfer, that is, only one training process can complete the transfer between multiple image domains. For example, Huang et al., in [7], proposed MUNIT, which considers latent vectors as content vectors and style vectors, and shares content vectors with different images to enhance image diversity. Lee et al., in [20], proposed DRIT to embed images into invariant domain content and domain-specific attribute space to capture cross-domain shared information to realize attribute transfer. Choi et al., in [1], proposed the StarGAN to transform multiple facial expressions and attributes only in a single GAN. He et al., in [8], proposed AttGAN, which takes the source image and the target attribute vector as input, and generates the image through the attribute classification constraint to ensure that the target attribute changes correctly. Liu et al., in [21], proposed STGAN, which uses Selective Transfer Units (STU) to transfer image content features from encoder to decoder so that the network can adaptively modify the features. Choi et al., in [2], proposed the StarGANv2 based on StarGAN, which provides multiple style information of the target domain without additional labels, reducing the dependence of attribute labels. Yang et al., in [36], proposed L2M-GAN, which introduced the orthogonality loss, separated the target attribute's associated style code from the unrelated style code, and carried out the attribute transfer. Mao et al., in [26], proposed SAVI2I using signed attribute vectors to achieve cross-domain image attribute transfer.

The effectiveness of facial image is highly subjective, as the concept of beauty varies across different cultures. These perceptions significantly influence the design and outcome of facial image processing. Diamant et al., in [28] proposed Bhoder-GAN, which generates facial images based on requested aesthetic scores. Wei et al., in [33], researched facial symmetry and attractiveness to quantify facial aesthetics. Donatas et al., in [22], using GAN to predict and enhance the aesthetic appeal of generated facial images, aiming to improve the prediction of facial attractiveness. Zheng et al., in [4], proposed FE-GAN, which incorporates emotions and expressions in facial generation, producing more distinctive images.

In the context of existing research, it becomes evident that the field of facial attribute transfer faces significant challenges. These challenges primarily revolve around non-frontal facial images and faces with occlusions, encompassing issues such as low image generation quality, attribute omissions, and a lack of diversity. These problems have hindered the effectiveness and scope of comprehensive facial attribute transfer.

This paper presents an innovative unsupervised multi-domain face attribute transfer network known as FES-StarGANv2, which is built upon the foundation of StarGANv2. The objective is to tackle these significant challenges. Our primary contributions are summarized as follows:

1 We propose an attention-guided feature fusion module that combines feature fusion and attention

mechanisms to eliminate redundancies in low-level features while enrich the local key features extracted from the face images.

2 To fully capture the style code of the target domain, we design a style code extraction module. It selectively allocates attention and extracts the style code that reflect key facial texture features.

3 We add a face image optimization module to simulate random changes in the real world like hair and wrinkles. To enhance the visual realism of the generated image.

4 To enhance diversity in the generated images, we employe a structural diversity loss function to maximize dissimilarity between generated images.

The structure of this paper is organized as follows: Section 2 gives a review of related works on face attribute transfer. Section 3 describes the proposed network model and a loss function in detail. Section 4 presents the experiments implements, such as datasets, evaluation metrics and setting. Section 5 presents the results, and analysis of the experiments. The conclusion is drawn in the final section.

2. Related Work

2.1. Face Attribute Transfer

Face attribute transfer is to change some attributes of a face image, such as age, gender, and facial expression, to maintain the same person while having new attribute characteristics. We can transform facial features from one person to another by training a network. This transformation can be modest, like changing hair color or lip shape, or more complex, like converting male to female facial features. The core concept of face attribute transfer lies in transferring the texture of the image.

Traditional image texture transfer is accomplished through non-parametric algorithms. Originally, Hertzmann et al., in [9] proposed an image texture synthesis and feature transfer algorithm based on an approximate nearest neighbor search nonparametric method, which involves resampling the source image to create a new texture. However, this non-parametric approach often results in the loss of high-level semantic information in the image.

With the widespread adoption of deep learning, transfer models based on Convolutional Neural Networks (CNNs), as proposed by Gatys et al., in [6], have emerged. These models leverage CNNs to simultaneously extract and model both the underlying texture information and high-level semantic content of the image. Many subsequent texture transfer algorithms are also built upon this foundation [10, 18, 32]. Although this method requires a lengthy training process and can produce relatively uniform results.

The advent of Generative Adversarial Networks (GANs) has provided a powerful tool for manipulating facial attributes. GANs generate realistic, high-definition face images through adversarial training, and algorithms relying on GANs for image texture transfer have emerged. Initially, Mirz et al., in [25], introduced conditional GAN (cGAN), which successfully use class labels to generate specific images randomly. After that, Pix2Pix [17], CycleGAN [39], DualGAN [37], and other networks are based on GAN to realize image texture transfer. These algorithms are supervised learning that requires paired data sets.

To overcome the limitations of datasets, numerous unsupervised algorithms relying on a single network to achieve texture transfer across multiple domains have emerged. Examples include MUNIT [7], DRIT [20], StarGAN [1], STGAN [8], AttGAN [21], L2M-GAN [36], and I2 I[26], which employ generated adversarial network for image texture migration. Choi et al., in [2], proposed StarGANv2 method represents an improved version of StarGAN, eliminating the need for labels while enabling transformations of multiple facial attributes through an MLP structure.

Both face attribute transfer and face recognition are important application areas in the field of face image processing. The former aims to modify face attributes, while the latter focuses on recognizing faces. Excellent face recognition technology can greatly assist in face attribute transfer. Saeed et al., in [31] proposed a framework for recognition of facial expression using HOG features, which significantly improved accuracy in recognizing expressions. Zeebareeet al., in [40], proposed a face mask detection using Haar Cascades classifiers, which can reduce the cost of real-time identity verification. These advancements provide a foundational contribution to the development of facial attribute transfer in later stages.

2.2. Feature Fusion

In the realm of convolutional neural networks utilized for feature extraction, feature fusion stands as a prevalent strategy to enhance model efficacy. There are primarily two techniques for fusing features: addition and concatenation. The method of feature concatenation involves amalgamating two or more feature maps along the channel dimension. Prominent models like DenseNet [11] and U-net [30] employ this approach, leveraging semantic data across feature maps of diverse scales to augment the number of channels, thereby elevating performance. On the other hand, feature addition entails a direct, pixel-wise summing of feature maps, as exemplified by models such as ResNet [12] and FPN [23]. This technique amplifies the descriptive power of the feature maps across each dimension without a change in dimensionality.

2.3. LSTM

The Long Short-Term Memory network (LSTM) [13] constitutes a specialized variant of Recurrent Neural Networks (RNNs) [14] tailored expressly for treating sequential data. In order to solve the problem of gradient disappearance and gradient explosion that affect the existence of traditional RNNs, the LSTM incorporates a nuanced gating mechanism. This mechanism employs a cell state alongside three specific gated units: the forget gate, input gate, and output gate, thereby aptly capturing long-term dependencies. Owing to these characteristics, LSTMs are exceptionally well-suited for a wide range of applications, including Natural Language Processing (NLP) [3], speech recognition [27], and image processing [24, 38].

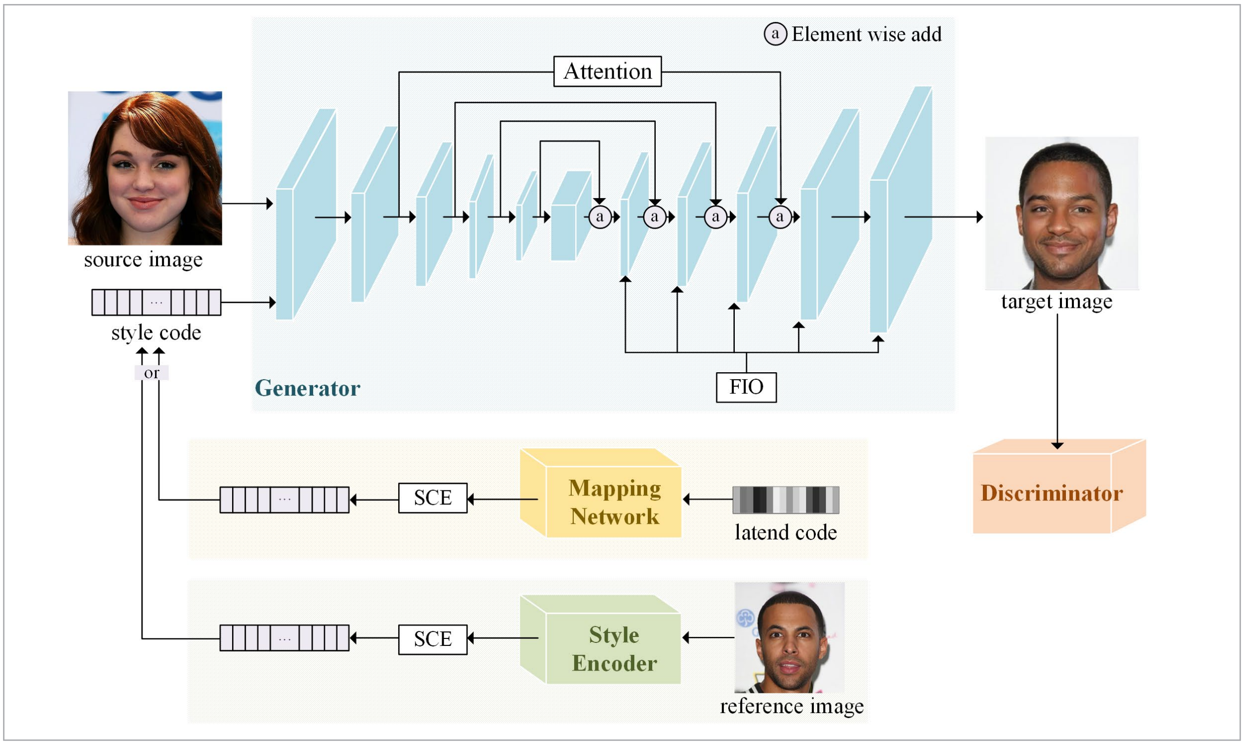
3. Approaches

3.1. Overall Architecture

The essence of face attribute transfer is to generate the desired image by changing some facial characteristics while keeping others the same. Based on the StarGANv2 method proposed in [2], we design an unsupervised multi-domain face attribute transfer network FES-StarGANv2 fusing feature enhancement and structural diversity loss function. FES-StarGANv2 has roughly four modules: Generator, Discriminator, Mapping Network, Style Encoder and specialized sub-modules, like Style Code Extraction module (SCE), Face Image Optimization module (FIO), etc. Figure 1 shows the overall structure of the network.

Figure 1

The overall structure of FES-StarGANv2



The Style Encoder take a reference image and the domain associated with that image, then it learns the style of the reference image and generates a corresponding style code. The Mapping Network takes random latent code and a target domain as inputs. It learns the style of the target domain and outputs style codes that vary within that domain. These style codes are then fed into the Generator.

The Generator employs an encoder-decoder architecture with six encoders and decoders. It receives the style codes generated by the Mapping Network and Style Encoder. The Generator then fabricates new images based on these style codes. The goal here is to alter specific facial attributes (like age, gender, hairstyle, etc.) while maintaining the integrity of other features.

The Discriminator comprises six pre-activated residual blocks. Its primary function is to distinguish between real and generated fake data.

The SCE module is designed to optimize the style codes, ensuring they accurately represent the desired attributes. The FIO module focuses on enhancing the diversity and fidelity of the generated images, making them more realistic and varied. The Attention Module: Improves the overall quality of the images by focusing on specific areas of the image that need refinement or greater detail.

By integrating these modules, FES-StarGANv2 efficiently accomplishes multi-domain face attribute transfers without needing paired data.

3.2. Attention-Guided Feature Fusion Module

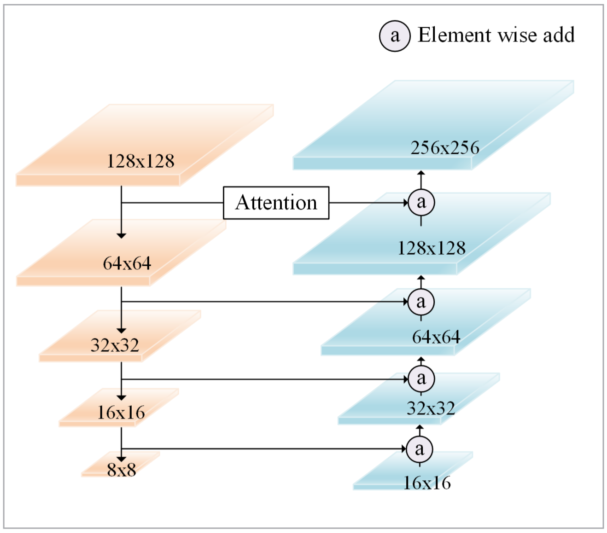
In image processing tasks, low-level features typically contain fine-grained details of the image, while high-level features contain more abstract semantic information. By combining low-level features with high-level features through skip connections, it ensures that the generated image retains fine details, thereby improving the quality of the generated image.

We add skip connections of element-wise addition between feature maps of different resolutions (128×128, 64×64, 32×32, and 16×16). This helps in preserving and transferring information from lower-resolution feature maps to higher-resolution ones, can enhance the overall understanding of the image, which can be crucial for generating high-quality images.

Low-level features often contain a significant amount of redundant information, which may not be particularly helpful for the task and may even disrupt the performance of the model. To address this, we introduce an attention module into the first layer of down-sampling (128×128 size, 128-channel feature map), which focus on task-related feature information. This fusion step enhances the precision and refinement of the features, making them more accurate. Figure 2 illustrates the improved encoder-decoder structure.

Figure 2

The encoder-decoder of Attention-guided feature fusion (AFF)



Guided by the attention module, the network not only gains access to more informative features but also effectively filters out the adverse effects of redundant features. This emphasis on relevant information and suppression of redundant details ensures that the subsequent feature fusion module benefits from a refined and focused set of features, leading to improved precision and feature integration in the final output.

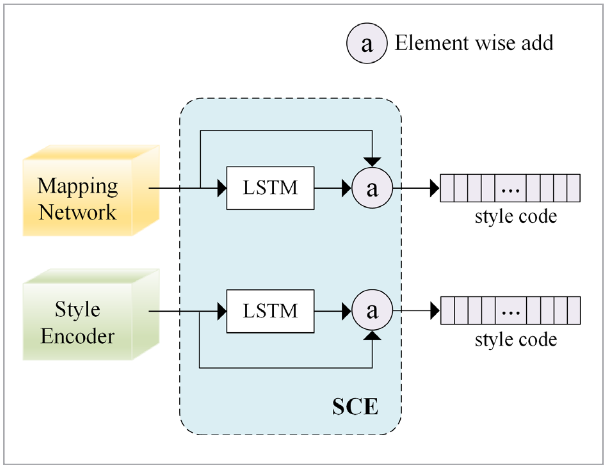
The attention module is implemented using the Convolutional Block Attention Module (CBAM) [34], which comprises both a Channel Attention Module (CAM) and a Spatial Attention Module (SAM). This combination effectively identifies spatially significant facial features. The introduction of the attention-guided feature fusion module allows the network to concentrate on critical facial characteristics while efficiently filtering out less relevant peripheral areas.

3.3. Style Code Extraction Module

After the mapping network and the style encoder, we introduce the Style Code Extraction Module (SCE) with the aim of enhancing the accuracy and precision of the style code. This SCE module processes the style data to generate a more precise style code, assisting the generator in gaining a nuanced understanding of the stylistic characteristics of the target domain. The SCE module is based on LSTM, as depicted in Figure 3.

Figure 3

The Style Code Extraction module (SCE)



The extracted style code is presented as a sequence. This sequence-based style code allows us to use LSTM to uncover the relationships among regional features. LSTM excels at capturing the temporal relationships within the sequence of target style codes, thus providing valuable contextual information about the target style image. By selectively extracting relevant information from the target image, without relying on additional supervision data, we enhance the accuracy of the style code representation.

In the SCE module, LSTM is computed as follows:

$$i_t = \sigma(W_{it} x_t + b_{ii} + W_{hi} h_{(t-1)} + b_{hi}), \quad (1)$$

$$f_t = \sigma(W_{if} x_t + b_{if} + W_{hf} h_{(t-1)} + b_{hf}), \quad (2)$$

$$g_t = \tanh(W_{ig} x_t + b_{ig} + W_{hg} h_{(t-1)} + b_{hg}), \quad (3)$$

$$o_t = \sigma(W_{io} x_t + b_{io} + W_{ho} h_{(t-1)} + b_{ho}), \quad (4)$$

$$c_t = f_t c_{(t-1)} + i_t g_t, \quad (5)$$

$$h_t = o_t \tanh(c_t), \quad (6)$$

where x_t is the input temporal sequence at the current time t, W and b are the weight matrix and bias vector; i_t is the input gate, f_t is the forget gate, o_t is the output gate, g_t is the output of the cell gate, c_t is the state of the cell at time t, h_t is the hidden layer state at time.

3.4. Face Image Optimization Module

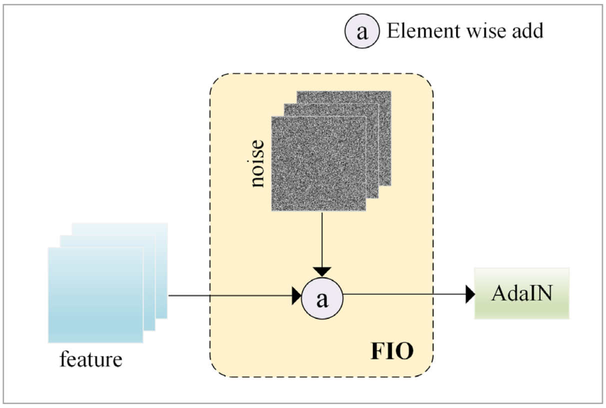
Conventional face attribute transfer models often produce similar images because they rely solely on latent codes, which might not capture the unique characteristics of each face image. As a result, the generated images can lack diversity and authenticity.

To overcome this limitation, the authors introduce the Face Image Optimization Module (FIO) into the decoder of the model[19]. The random variations introduced by the FIO module are meant to mimic the subtle changes in hair and skin texture that are commonly observed in real-world faces. These variations make the generated images more realistic and diverse. The specific implementation of the FIO module involves adding random noise to the feature map of the decoder. This noise is added with a certain weight, as depicted in Figure 4.

Random noise is incorporated ahead of AdaIN [15] to ensure that any stylistic alterations applied to the generated image remain subtle and do not adversely affect the overall composition of the image.

Figure 4

The Face Image Optimization module (FIO



3.5. Structural Diversity Loss Function

To enhance the expressiveness, diversity, and detail of the images generated by generator G, we add a structural diversity loss function as a supplement:

$$L_{sd}(x, \widetilde{y}) = 1 - SSIM(G(x, \widetilde{s}_1), G(x, \widetilde{s}_2)), \quad (7)$$

where x is the source image, $\widetilde{y}$ is the target domain, $\widetilde{s}_1$ and $\widetilde{s}_2$ are two different style codes belonging to the same target domain $\widetilde{y}$, $G(x,\widetilde{s}_1)$ and $G(x,\widetilde{s}_2)$ are two different transfer images generated by generator G under the guidance of $\widetilde{s}_1$ and $\widetilde{s}_2$, respectively.

Introducing this loss function L_{sd} guides generator G to take into account more structural changes when generating new images, resulting in transferred images with enhanced expressive richness. In our specific implementation, we employ the Structure Similarity Index Measure (SSIM)[35] loss function to constrain the maximum inconsistency in color and edge texture structure between two transfer images generated by two distinct style codes. During training, we aim to maximize the structure diversity loss L_{sd}. The SSIM metric quantifies the similarity of two images by considering factors like brightness, contrast, and structure. For two different images, denoted as x and y, the SSIM is defined as follows:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \quad (8)$$

where μ_x and μ_y are the means of x and y respectively, σ_x^2 and σ_y^2 are the variances of x and y respectively, σ_{xy} is the covariance of x and y, c_1 and c_2 are constants.

Finally, our comprehensive objective function can be summarized as follows:

$$L = L_{adv} + \lambda_{sty}L_{sty} - \lambda_{ds}L_{ds} - \lambda_{sd}L_{sd} + \lambda_{cyc}L_{cyc}, \quad (9)$$

where L_{adv} is the adversarial loss, which measures the difference between the fake data and the real. L_{sty} is the style reconstruction loss, constricting the generator G to better preserve the image style upon attribute transfer. L_{ds} and L_{sd} are the diversity sensitive loss and the structural diversity loss, respectively, which increase the diversity expression of the generated image. L_{cyc} is a cycle consistency loss, which guarantees the consistency of the generated image with the content of the source image. λ_{sty}, λ_{ds}, λ_{sd} and λ_{cyc} are hyperparameters for each term.

4. Implementation of Experiments

4.1. Datasets

We use the CelebA-HQ dataset for training and validation, which is an advanced version of the CelebA dataset that contains 30K high-quality face images with a maximum resolution of 1024x1024.

During training, the dataset is divided into male and female domains, consisting of 28,000 training images and 2,000 validation images. To maintain consistency in our experiments, we consistently utilized 256x256 resolution images.

4.2. Evaluation Metrics

The CelebA-HQ dataset uses FID [16] and LPIPS [41] as the evaluation metrics of model, which measure the accuracy and diversity of the generated images, respectively.

4.2.1. FID

FID (Fréchet Inception Distance) is a metric used to assess the disparity between two image distributions (typically between real and generated images). It is commonly employed to evaluate the quality of images produced by GANs.

Use the Inception-v3 model to extract features from an intermediate layer for both the real and generated images. This results in a feature vector for each image. Calculate the mean and covariance of the feature vectors for the real images and the generated images. We define the Gaussian distribution of the real images is defined as $X_1 \sim (m, C)$, and the Gaussian distribution of the generated images is $X_2 \sim (m_\omega, C_\omega)$, The FID calculation is formulated as:

$$\begin{aligned} &d^2((m,C),(m_\omega,C_\omega)) = \\ &\|m - m_0\|_2^2 + Tr(C + C_\omega - 2\sqrt{CC_\omega}), \end{aligned} \quad (10)$$

In the best case scenario, FID is 0, indicating that the two sets of images are identical. A lower FID is highly correlated with higher quality generated images.

4.2.1. LPIPS

LPIPS (Learned Perceptual Image Patch Similarity) is a metric used to evaluate the perceptual differences between two images from a human visual perspective. LPIPS calculates the discrepancies between outputs from various layers of a pretrained AlexNet network. By weighting these differences, LPIPS yields a scalar value that measures the perceptual disparity between the two images. The larger the LPIPS, the higher the diversity of the generated images. Define two generated images as x and x_0, and the LPIPS calculation is formulated as:

$$d(x, x_0) = \sum \frac{1}{H_l W_l} \sum \left\| w_l \otimes (\hat{y}^l_{hw} - \hat{y}^l_{0hw}) \right\|_2^2 , \tag{11}$$

4.3. Training Details

The experimental hardware environment is made of Intel i9-10980XE CPU, 64GB, and NVIDIA GeForceRTX3090, 24GB. The software contains Python3.6.5, CUDA11.3, cuDNN8.8.0, PyTorch11.3 framework, and Ubuntu22.04 operating system.

In terms of hyper-parameter, the batch size is set to 8, and the total number of iterations is set to 100K. The learning rate of the Generator, Discriminator, and Style Encoder is set to 0.0001, and the learning rate of the Mapping Network is set to 0.00001. In model optimization training, Adam is used as the gradient descent algorithm, with β_1 and β_2 set to 0 and 0.99. In terms of weight selection, set λ_{sty} , λ_{ds}, λ_{sd} and λ_{cyc} to 1, and for stable training, λ_{ds} and λ_{sd} linearly decays to zero in 100K iterations. Table 1 shows the influence of λ_{ds} on the experimental effect, so λ_{ds} is set to 1.

Table 1

The influence of λ_{ds} on LPIPS, LPIPS is related to image diversity

the value of λ_{sd}	LPIPS_latent↑	LPIPS_reference↑
λ_{sd}=0	0.436	0.383
λ_{sd} =0.5	0.445	0.379
λ_{sd} =1	**0.453**	**0.389**

In the testing and validation experiments of this study, target domain images are chosen to significantly differ from the source domain images. This includes selecting images with substantial skin color differences, vivid hair colors (like blonde or white), and faces with distinct attributes such as numerous wrinkles. This approach helps to better assess the capability of the network to handle and transform images with various attributes.

5. The Results

To validate the efficacy of the experiment, we conducted comparisons between the FES-StarGANv2 method and other face attribute transfer techniques, performing both qualitative and quantitative experiments. We analyze experimental data and image visualization results from two viewpoints: latent code-guided image generation and reference image-guided image generation.

5.1. Latent Code-Guided Generation

We present a quantitative comparison between our method and other approaches in Table 2. The FID value, inversely related to image quality, and the LPIPS value, directly linked to image diversity, our method achieves lower FID and higher LPIPS values compared to other methods, indicating its excellence in generating images of superior quality and diversity. Compared with the baseline StarGANv2, our method achieves an FID value of 13, a reduction of 11%, and an LPIPS value of 0.453, an improvement of 3.9%.

Table 2

Quantitative comparison of latent code-guided generation

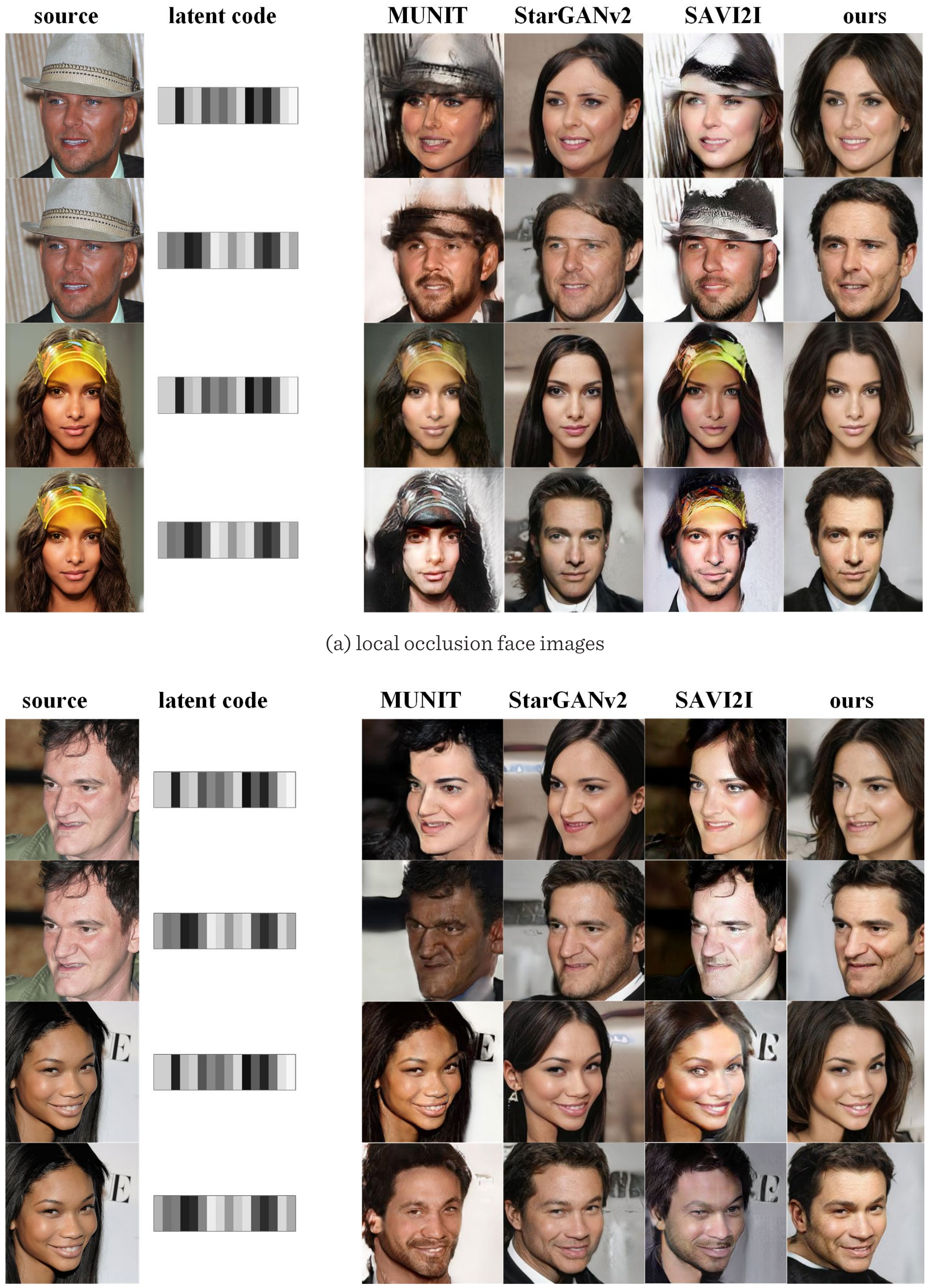
Datasets	Methods	FID↓	LPIPS↑
CelebA-HQ	MUNIT	31.4	0.363
	DRIT	52.1	0.178
	MSGAB	33.1	0.389
	StarGANv2	14.6	0.436
	SAVI2I	48.3	0.311
	ours	**13.0**	**0.453**

We also present the visualizations results of three methods applied to local occlusion and non-frontal face images in the Figure 5. Figure 5(a) displays the visualization results for local occlusion face images, while Figure 5(b) illustrates results for non-frontal face images.

Observing the visualizations in Figure 5, it becomes apparent that when the MUNIT and SAVI2I methods learn random image styles, the transferred images it generates fail to change attributes such as hairstyle and hat. Similarly, the images generated by the StarGANv2 method exhibiting noticeable deficiencies in facial edges, hairstyles, and overall realism. In contrast, our FES-StarGANv2 method excels, producing higher-quality images that effectively mitigate the issue of incomplete generation. These images successfully modify attributes like hairstyle and hat, with

Figure 5

Qualitative comparison of latent code-guided image generation



(a) local occlusion face images

(b) non-frontal face images

complete and accurate facial edges and hairstyles. This demonstrates the advantages of our method in the domain of face attribute transfer.

The first column is the source image, and the second is the latent code. The source image is transferred to the target domain using a random sample of the latent code. And the third, fourth, fifth and the sixth columns are the transferred images generated by the MUNIT, StarGANv2, SAVI2I, and FES-StarGANv2(ours), respectively.

5.2. Reference Image-Guided Generation

Table 3 presents a quantitative comparison of reference image-guided image generation on the CelebA-HQ datasets. FES-StarGANv2 methods (ours) achieves better FID and LPIPS values. With an FID value of 20.2, our method outperforms the baseline network by 23%. Additionally, our LPIPS score reaches 0.389, marking a 1.6% improvement over the baseline.

Furthermore, we visualize the experimental results of the MUNIT, StarGANv2, SAVI2I and FES-StarGANv2 methods on local occlusion and non-frontal face images in Figure 6. Figure 6(a) and 6(b) depict the visualization results for local occlusion and non-frontal face images, respectively.

Comparing and observing the visualizations in Figure 6, the MUNIT and SAVI2I method struggle to learn the target image style, resulting in transferred images that exhibit little attribute change and substantial face distortion. Whereas the StarGANv2 method correctly learns the target image style, it still faces issues such as missing face edges and ghosting in the generated images.

Table 3

Quantitative comparison of reference image-guided generation

Datasets	Methods	FID↓	LPIPS↑
CelebA-HQ	MUNIT	107.1	0.176
	DRIT	53.3	0.311
	MSGAB	39.6	0.312
	StarGANv2	26.2	0.383
	SAVI2I	40.1	0.287
	ours	**20.2**	**0.389**

In contrast, our FES-StarGANv2 method excels both in terms of visual quality and functionality. It effectively learns the target domain's style to achieve attribute transfer while preserving the distinct characteristics of the source image. The generated images are comprehensive, devoid of ghosting or missing elements, and boast lifelike features. These results strongly affirm the substantial advantages of FES-StarGANv2 in the field of face attribute transfer.

The source image is transferred to the target domain by reference images that belong to the target domain.

5.3. Robustness Analysis

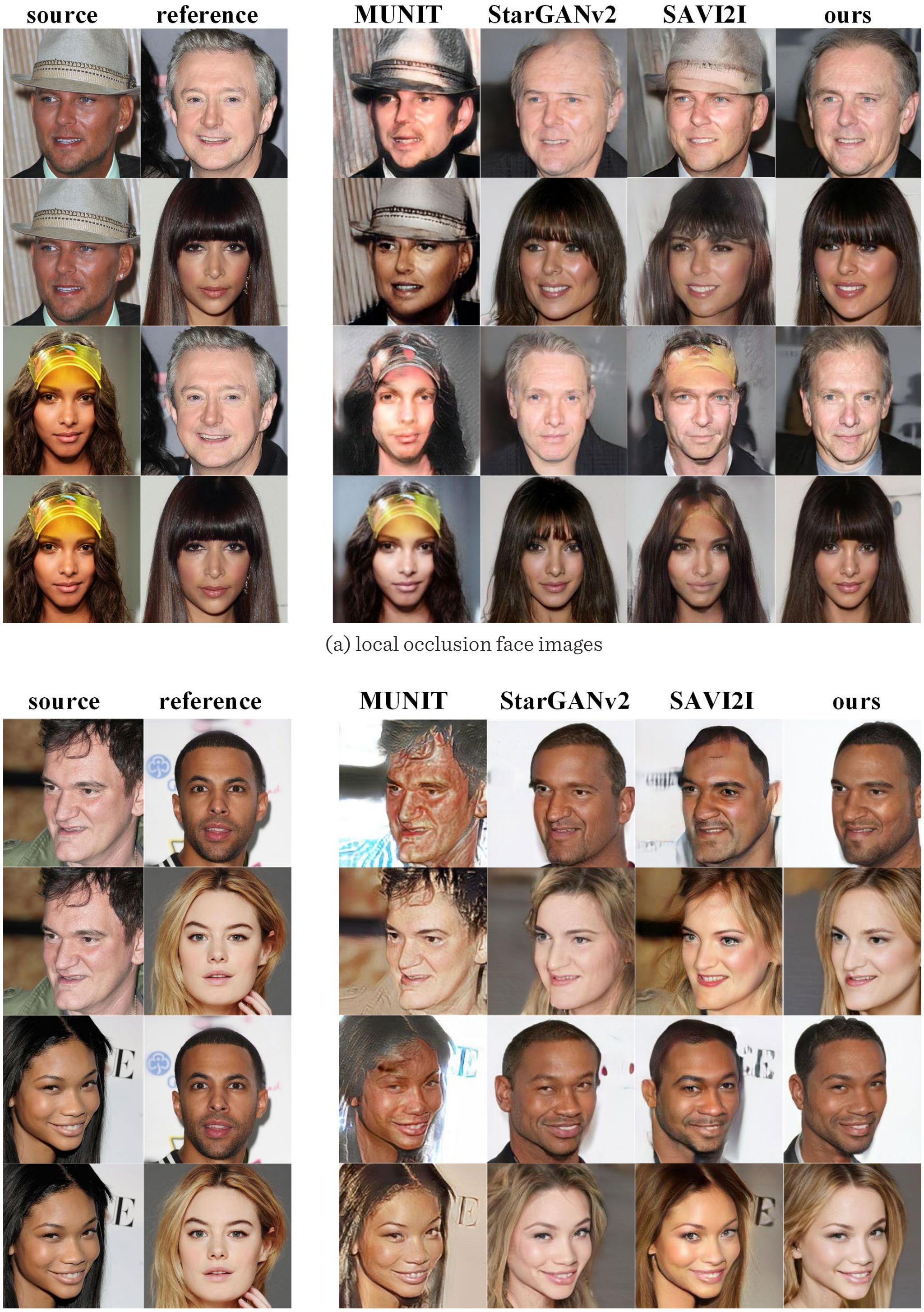
To thoroughly validate the robustness and generalization capabilities of our model, we conducted a comprehensive elasticity test of the FES-StarGANv2 method.

This involved extending the application of the method beyond the CelebA-HQ dataset to include the FFHQ dataset, specifically focusing on challenging scenarios. We deliberately selected non-frontal face images and facial occlusions (such as such as faces with hats) from the FFHQ dataset, which included images of both men and women. The generated images exhibited clear outlines and no missing facial features, demonstrating strong robustness of the method in handling complex conditions, as seen in Figure 7.

In addition to this, we applied five different types of distortions, including Gaussian blur, image saturation changes, image contrast changes, etc., to test the ability of the model to cope with real-world perturbations. These perturbations were designed to mimic real-world scenarios more accurately, as shown in Figure 8. Figure 8(a) shows images processed with real-world perturbations that were fed into the generator as source images. Figure 8(b) shows the image generated after the processed image is fed into the model. As can be seen from Figure 8(b), most of the images generated after processing are good and complete, and the most important thing is that the generated image effect is not significantly different from the original image. However, the image generated by the internal occlusion perturbation processing does not perform as well at the nose, where the original black spot is still present. Nevertheless, the excellent performance of other images also proves the adaptability and effectiveness of the FES-StarGANv2 method in handling various data types.

Figure 6

Qualitative comparison of reference image-guided image generation



(a) local occlusion face images

(b) non-frontal face images

Figure 7

FES-StarGANv2 method on the FFHQ dataset

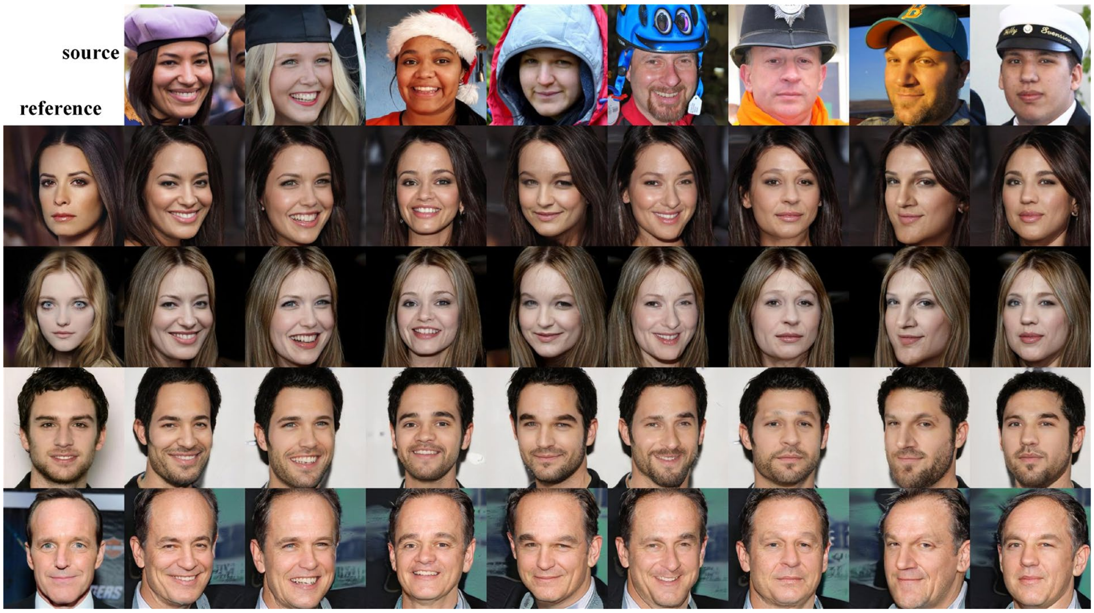
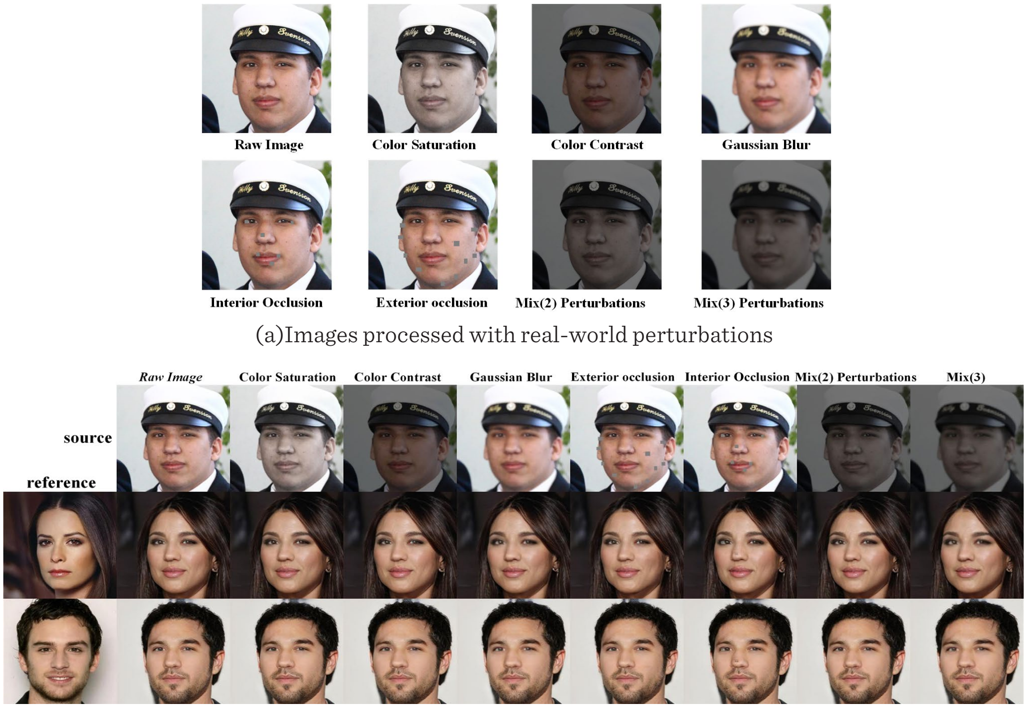


Figure 8

Real-World Perturbation



(a)Images processed with real-world perturbations

(b)The generated image of the processed with real-world perturbations

In summary, the FES-StarGANv2 method exhibits strong robustness and generalization capabilities, performing well on challenging datasets, complex scenarios, and real-world disturbances. This suggests its potential to be applied in various applications, including face recognition, facial expression analysis, and facial attribute editing, with high accuracy and reliability.

The top row is the source image, the first column is the reference image.

5.4. User Survey Evaluation

The qualitative experiment weas enriched through user evaluation to ensure the fairness and scientific nature. 30 students were randomly selected to rate the images generated using different methods. These shared images consist of 20 sets of evaluation samples, and each collection contains three transferred images generated by 3 methods based on the same input. In this setting, we can obtain 30×20=600 sets of subjective evaluation results.

Calculate the proportion of data obtained from 600 questionnaire evaluation data groups; Tables 4-5 show the results. The user criteria for selecting the best image are as follows (in choosing the best image generated by the latent code-guided, not to consider the target style expression):

1. Image Quality: The best image should have precise details such as complete contour edges, hair, and facial features.
2. Visual Perception: The best image should look like a human.
3. Target Style Expression: The best image should be consistent with the target image domain features to the greatest extent, without other unrelated region changes.

As indicated in Table 4, our method excels in both image quality and visual perception. Specifically:

1. Image Quality: Our approach produces images with exceptional image quality.
2. Visual Perception: The performance of our method is closely with the baseline method. The striking precision in the edge contours of our generated images can sometimes lead observers to believe that these images are the creations of AI.

These observations underscore the effectiveness and precision of our method in generating high-quality images that are visually compelling and closely resemble real human subjects.

Table 4
User survey evaluation results for latent code-guided generated images

Methods	proportion		
	Quality	Visual	Overall
MUNIT	0.01	0.02	0.015
StarGANv2	0.37	0.55	0.460
ours	0.62	0.43	0.525

As depicted in Table 5, our method excels in three crucial aspects when it comes to reference image-guided image generation:

1. Image Quality: Our method consistently delivers high-quality images. These images exhibit clarity in terms of contour edges, hair, and facial features.
2. Visual Perception: The images generated by our method are highly visually appealing. They closely resemble real human subjects.
3. Target Style Expression: Our method effectively captures and embodies the texture features of the target domain image. This results in images that not only maintain high quality but also exhibit a strong adherence to the desired style of the target domain.

Table 5
User survey evaluation results for latent code-guided generated images

Methods	proportion			
	Quality	Visual	Style	Overall
MUNIT	0.005	0.002	0	0.0023
StarGANv2	0.375	0.306	0.388	0.3563
ours	0.620	0.692	0.612	0.6413

Overall, our method demonstrates remarkable capabilities in producing images that are not only aesthetically pleasing but also represent the style and details of the target domain, including contour edges, hair, and facial features.

5.5. Ablation Study

We conducted separate assessments to validate the effects of different modules (AFF, SCE FIO) and the structural diversity loss on image quality and diversity. We still analyze experimental data from two viewpoints: latent code-guided image generation and reference image-guided image generation.

In our experimental setup, we conducted three control trials to assess the effectiveness of the AFF and the SCE. The results are presented in Table 6.

Table 6
Ablation study on CelebA-HQ

Methods	Latent		Reference	
	FID↓	LPIPS↑	FID↓	LPIPS↑
Baseline [10]	14.6	0.436	26.2	0.383
+ AFF	14.3	0.432	21.3	0.370
+ SCE	13.8	0.442	21.0	0.382
+ AFF + SCE	14.0	0.450	19.7	0.380

When the AFF module was added, the FID value decreased, indicating an improvement in image quality. Lower FID values suggest better image quality. However, AFF primarily focuses on image quality and does not impact image diversity, as indicated by the unchanged LPIPS score. When the SCE module was added, there was a significant decrease in FID values, indicating an improvement in image quality. The LPIPS score either improved or remained unchanged. This suggests that SCE's style code extraction effectively captures the style of the target domain, contributing to higher-quality images. When both AFF and SCE modules were added simultaneously, they collectively addressed the limitations of only adding AFF in terms of image diversity. The FID value was superior to the baseline, indicating improved image quality and diversity. This combination enhanced both image quality (due to AFF) and style accuracy (due to SCE).

The above result confirms that AFF enhances image quality, while SCE accurately captures the style of the target domain. When combined, AFF and SCE contribute to improved image quality and diversity compared to the baseline method StarGANv2.

Both FIO and the structural diversity loss function share a common goal, which is to enhance image diversity. This means they are designed to make sure that the generated images are not repetitive or too similar to each other. Two sets of experiments were conducted to assess the effectiveness of FIO and the structural diversity loss function. The details of these experiments are displayed in Table 7.

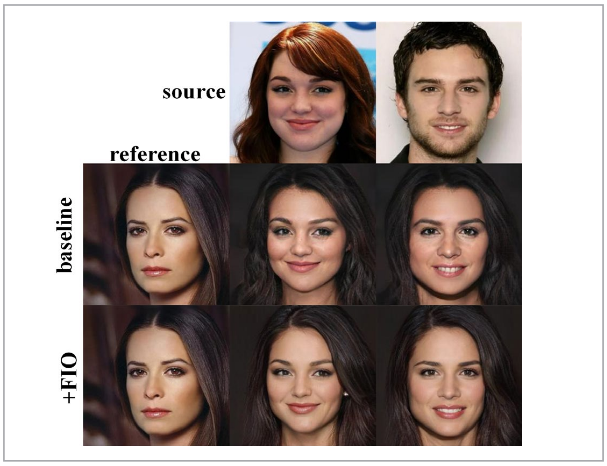
Table 7
Ablation study on CelebA-HQ

Methods	Latent		Reference	
	FID↓	LPIPS↑	FID↓	LPIPS↑
Baseline [10]	14.6	0.436	26.2	0.383
+FIO, L_{sd}	13.2	0.451	22.0	0.396
+ AFF + SCE +FIO, L_{sd}	13.0	0.453	20.2	0.389

Upon adding both FIO and the structural diversity loss function to the system, there was a significant improvement in the LPIPS value. LPIPS is often used to measure the similarity between two images from a perceptual perspective, so an improvement suggests that the generated images became more diverse. That indicates that FIO and the structural diversity loss function effectively contribute to increasing the expressive diversity of the generated images.

To further demonstrate the influence of FIO on image agility, Figure 9 shows the influence of the FIO on image authenticity. As can be seen from Figure 9, the

Figure 9
The influence of the FIO on image authenticity



facial details of the image generated by the addition of the FIO module are more realistic and more like real people compared to the baseline. This suggests that FIO plays a role in determining how authentic or genuine the generated images appear.

Overall, these findings suggest that integrating FIO and the structural diversity loss function into the system has a positive impact on the diversity and authenticity of the generated images, as evidenced by the improvement in LPIPS values and the observed correlation between image diversity and LPIPS.

Ultimately, the best overall performance was achieved by combining AFF, SCE, FIO, and L_{sd}. In this configuration, both FID and LPIPS reached their respective optima, surpassing the performance of the baseline method. This demonstrates the synergistic effect of integrating these components in improving both image quality and diversity.

6. Limitations

Unsupervised facial attribute transfer typically relies on existing datasets, and its ability to achieve attribute transfer depends on the attributes included in that dataset. Due to dataset limitations, facial attribute transfer cannot achieve complete freedom in transferring various attributes. If the dataset lacks substantial information about certain attributes, such as glasses or freckles, attribute transfer for these attributes may not perform well because the model lacks sufficient training data to learn and understand the details of these attributes.

The diversity and coverage of attributes in the dataset are crucial for the success of unsupervised facial attribute transfer. If the dataset contains samples with various attributes, and the range of variations in these attributes is sufficiently wide, unsupervised facial attribute transfer models will be better able to learn and achieve the transfer of these attributes. However, if the dataset lacks samples of certain attributes or if the variations in these attributes are limited, attribute transfer for those attributes may be constrained.

Therefore, for future endeavors, it is advisable to choose or create a dataset that encompasses diversity and offers a wide range of attribute coverage when conducting unsupervised facial attribute transfer. Additionally, when dealing with uncommon or rare attributes, it is recommended to consider utilizing more data to train the model and enhance its performance.

7. Conclusion

This paper addresses the common challenges faced in face attribute transfer for non-frontal and obstructed facial images, such as low-quality images, artifacts, missing facial details, and a lack of diversity. We introduce several key modules designed to enhance the overall quality and performance of the generated images.

The AFF module is integrated into the generator to tackle the aforementioned problems. It effectively enhances the quality of generated images by guiding the fusion of important features, resulting in clearer and more realistic images. The SCE module optimizes the style coding of the target domain. Improving the way the generator learns from target domain images, contributes to generating more style-consistent images. The FIO module and L_{sd} are designed to improve the overall flexibility and diversity of the generated images. It works in conjunction with the structural diversity function to ensure that the generated faces exhibit a wider range of attributes and characteristics.

In experiments on the CelebA-HQ dataset, the FES-StarGANv2 method outperforms other algorithms by enhancing image quality, reducing artifacts, and increasing diversity, effectively addressing common face attribute transfer challenges.

There are also limitations to our approach. Due to limitations in our dataset, there were several attributes, such as glasses, freckles, etc., that we could not adequately address during training, inevitably leading to certain oversights. In the future, we plan to select or create more extensive datasets for training, focusing specifically on uncommon or rare attributes. However, there are not too many datasets for specific domains, the FES-StarGANv2 method may not be entirely fair or representative when generalized to entirely new domains. In the future, we plan to explore the application of the FES-StarGANv2 method be generalized to other fields such as animals and automobiles.

Acknowledgement

This work was supported by Nature Science Foundation of Jilin Province, No. 20230101179JC, China.

References

1. Choi, Y., M.-J. Choi, M. S. Kim, J.-W. Ha, S. Kim, Choo, J. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, 8789-8797. https://doi.org/10.1109/CVPR.2018.00916
2. Choi, Y., Y. Uh, J. Yoo, Ha, J.-W. StarGAN v2: Diverse Image Synthesis for Multiple Domains. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 8185-8194. https://doi.org/10.1109/CVPR42600.2020.00821
3. Cao, S., Gao, P. LSTM-Gate CNN Network for Aspect Sentiment Analysis. 2020 5th International Conference on Information Science, Computer Technology and Transportation (ISCTT), 2022, 443-447. https://doi.org/10.1109/ISCTT51595.2020.00084
4. Fang, Z., Liu, Z., Liu, T., C. H., Xiao, J., Feng, G. Facial Expression GAN for Voice-Driven Face Generation. The Visual Computer, 2022, 38, 1151-1164. https://doi.org/10.1007/s00371-021-02074-w
5. Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, Bengio, Y. Generative Adversarial Nets. Advances in Neural Information Processing Systems, 2014, 139-144. https://doi.org/10.1145/3422622
6. Gatys, L. A., A. S. Ecker, Bethge, M. Image Style Transfer Using Convolutional Neural Networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 2414-2423. https://doi.org/10.1109/CVPR.2016.265
7. Huang, X., M.-Y. Liu, S. J. Belongie, Kautz, J. Multimodal Unsupervised Image-to-Image Translation. European Conference on Computer Vision (ECCV), 2018, 172-189. https://doi.org/10.1007/978-3-030-01219-9_11
8. He, Z., W. Zuo, M. Kan, S. Shan, Chen, X. AttGAN: Facial Attribute Editing by Only Changing What You Want. 2017 IEEE Transactions on Image Processing (TIP), 2017, 28(11), 5464-5478. https://doi.org/10.1109/TIP.2019.2916751
9. Hertzmann, A., C. E. Jacobs, N. Oliver, B. Curless, Salesin, D. Image Analogies. Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, 2001, 327-340. https://doi.org/10.1145/383259.383295
10. Huang, X., Belongie, S. J. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. 2017 IEEE International Conference on Computer Vision (ICCV), 2017, 1510-1519. https://doi.org/10.1109/ICCV.2017.167
11. Huang, G., Z. Liu, Weinberger, K. Q. Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 2261-2269. https://doi.org/10.1109/CVPR.2017.243
12. He, K., X. Zhang, S. Ren, Sun, J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90
13. Hochreiter, S., Schmidhuber, J. Long Short-Term Memory. Neural Computation 9, 1997, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735
14. Hopfield, J. J. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. Proceedings of the National Academy of Sciences of the United States of America, 1982, 79(8), 2554-8. https://doi.org/10.1073/pnas.79.8.2554
15. Huang, X., Belongie, S. J. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. 2017 IEEE International Conference on Computer Vision (ICCV), 2017, 1510-1519. https://doi.org/10.1109/ICCV.2017.167
16. Heusel, M., H. Ramsauer, T. Unterthiner, B. Nessler, Hochreiter, S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. Advances in Neural Information Processing Systems (NIPS), 2017, 30. https://doi.org/10.48550/arXiv.1706.08500
17. Isola, P., J.-Y. Zhu, T. Zhou, Efros, A. A. Image-to-Image Translation with Conditional Adversarial Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 5967-5976. https://doi.org/10.1109/CVPR.2017.632
18. Johnson, J., A. Alahi, Fei-Fei, L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. European Conference on Computer Vision, Springer, Cham, 2016, 694-711. https://doi.org/10.1007/978-3-319-46475-6_43
19. Karras, T., S. Laine, Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 4396-4405. https://doi.org/10.1109/CVPR.2019.00453
20. Lee, H.-Y., H.-Y. Tseng, J.-B. Huang, M. K. Singh, Yang, M. Diverse Image-to-Image Translation via Disentan-

gled Representations. Proceedings of the European Conference on Computer Vision (ECCV), 2018, 35-51. https://doi.org/10.1007/978-3-030-01246-5_3

21. Liu, M., Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, Wen, S. STGAN: A Unified Selective Transfer Network for Arbitrary Image Attribute Editing. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 3668-3677. https://doi.org/10.1109/CVPR.2019.00379
22. Laurinavičius, D., R. Maskeliūnas, Damaševičius, R. Improvement of Facial Beauty Prediction Using Artificial Human Faces Generated by Generative Adversarial Network. Cognitive Computation, 2023, 15, 998-1015. https://doi.org/10.1007/s12559-023-10117-8
23. Lin, T.-Y., P. Dollár, R. B. Girshick, K. He, B. Hariharan, Belongie, S. J. Feature Pyramid Networks for Object Detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 936-944. https://doi.org/10.1109/CVPR.2017.106
24. Liang, X., X. Shen, J. Feng, L. Lin, Yan, S. Semantic Object Parsing with Graph LSTM. European Conference on Computer Vision, Springer, Cham, 2016, 125-143. https://doi.org/10.1007/978-3-319-46448-0_8
25. Mirza, M., Osindero, S. Conditional Generative Adversarial Nets. Computer Science. arXiv Preprint arXiv:1411.1784, 2014. https://doi.org/10.48550/arXiv.1411.1784
26. Mao, Q., H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, S. Ma, Yang, M.-H. Continuous and Diverse Image-to-Image Translation via Signed Attribute Vectors. International Journal of Computer Vision, 2020, 130, 517-549. https://doi.org/10.1007/s11263-021-01557-6
27. Moritz, N., T. Hori, Le Roux, J. Unidirectional Neural Network Architectures for End-to-End Automatic Speech Recognition. Interspeech, 2019, 2019-2837. https://doi.org/10.21437/Interspeech.2019-2837
28. Diamant, N., D. Zadok, C. Baskin, E. Schwartz, Bronstein, A. M. Beholder-GAN: Generation and Beautification of Facial Images with Conditioning on Their Beauty Level. 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019, 739-743. https://doi.org/10.1109/ICIP.2019.8803807
29. Perarnau, G., J. van de Weijer, B. Raducanu, Álvarez, J. M. Invertible Conditional GANs for Image Editing. arXiv Preprint arXiv:1611.06355, 2016. https://doi.org/10.48550/arXiv.1611.06355
30. Ronneberger, O., P. Fischer, Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, 3(18), 234-241, Springer. https://doi.org/10.1007/978-3-319-24574-4_28
31. Saeed, V. A. A Framework for Recognition of Facial Expression Using HOG Features. International Journal of Mathematics, Statistics, and Computer Science, 2023, 2, 1-8. https://doi.org/10.59543/ijmscs.v2i.7815
32. Ulyanov, D., A. Vedaldi, Lempitsky, V. S. Improved Texture Networks: Maximizing Quality and Diversity in Feed-Forward Stylization and Texture Synthesis. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 4105-4113. https://doi.org/10.1109/CVPR.2017.437
33. Wei, W., Ho, E. S., McCay, K. D., Damaševičius, R., Maskeliūnas, R., Esposito, A. Assessing Facial Symmetry and Attractiveness using Augmented Reality. Pattern Analysis and Applications, 2022, 1-17. https://doi.org/10.1007/s10044-021-00975-z
34. Woo, S., J. Park, J.-Y. Lee, Kweon, I.-S. CBAM: Convolutional Block Attention Module. Proceedings of the European conference on computer vision (ECCV), 2018, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1
35. Wang, Z., A. C. Bovik, H. R. Sheikh, Simoncelli, E. P. Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Transactions on Image Processing 13, 2004, 600-612. https://doi.org/10.1109/TIP.2003.819861
36. Yang, G., N. Fei, M. Ding, G. Liu, Z. Lu, Xiang, T. L2M-GAN: Learning to Manipulate Latent Space Semantics for Facial Attribute Editing. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 2950-2959. https://doi.org/10.1109/CVPR46437.2021.00297
37. Yi, Z., H. Zhang, P. Tan, Gong, M. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. 2017 IEEE International Conference on Computer Vision (ICCV), 2017, 2868-2876. https://doi.org/10.1109/ICCV.2017.310
38. Yang, G., J. M. Manela, M. Happold, Ramanan, D. Hierarchical Deep Stereo Matching on High-Resolution Images. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 5510-5519. https://doi.org/10.1109/CVPR.2019.00566
39. Zhu, J.-Y., T. Park, P. Isola, Efros, A. A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. 2017 IEEE International Conference on Computer Vision (ICCV), 2017, 2242-2251. https://doi.org/10.1109/ICCV.2017.244

40. Zeebaree, I. M., Kareem, O. S. Face Mask Detection Using Haar Cascades Classifier to Reduce the Risk of COVID-19. International Journal of Mathematics, Statistics, and Computer Science, 2023, 2, 19-27. https://doi.org/10.59543/ijmscs.v2i.7845

41. Zhang, R., P. Isola, A. A. Efros, E. Shechtman, Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, 586-595. https://doi.org/10.1109/CVPR.2018.00068