Biphasic Face Photo-Sketch Synthesis via Face Semantic-Aware CycleGAN

Xudong Wang , Feng Ye , Yaojuan Li , Yiwei Dai
School of Artificial Intelligence, Jianghan University, Wuhan, China
729531163@qq.com, 89666538@qq.com, nyuman9@163.com, 167998906@qq.com
Corresponding Author: Feng Ye Email: 89666538@qq.com

2023 IEEE 11th Joint International Information Technology and Artificial Intelligence Conference (ITAIC) | 979-8-3503-3366-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ITAIC58329.2023.10409099

***Abstract*—With the introduction of CycleGAN, there have been significant advancements in the task of biphasic face photo-sketch synthesis (FPSS). To further enhance the results in this task, most methods have focused on optimizing the model without attempting to exploit the additional information inherent in the face itself. To address this issue, this paper proposes a novel Face Semantic-Aware CycleGAN (FSACycleGAN) for bidirectional face photo-to-sketch synthesis. Firstly, a method is introduced to inject facial semantic distribution images into the generator, providing additional spatial and identity supervision for the synthesized face photos and sketches. Furthermore, facial semantic information is incorporated into the loss function along with identity differences in generated images, aiding in preserving facial details and identity information in the generated face photos and sketches. Lastly, a novel cyclic adversarial training strategy is proposed. It involves introducing an additional recognition model, enabling a cycle of adversarial training between the recognition and generation models to optimize the generated results. Through extensive training on the CUFS dataset, the results demonstrate that the proposed method has achieved significant improvements compared to CycleGAN. Specifically, it has shown a 16.3% improvement in SSIM and a 5.5% improvement in FSIM. The accuracy in sketch matching and photo matching tasks reached 99.42% and 98.64%, respectively, showcasing excellent performance.**

Keywords—generative adversarial network; face photo-sketch synthesis; iterative cycle training; Face Semantic Aware;

I. INTRODUCTION

Biphasic face photo-sketch synthesis (FPSS) is a task in image-to-image translation problem that finds wide applications in areas such as digital entertainment, law enforcement, and criminal case investigation [1]. Bidirectional FPSS comprises two sub-problems: synthesizing sketches from facial photos and synthesizing facial photos from sketches.

In the past research on FPSS has followed two fundamental technical approaches: sample-based methods [1,2,4] and subspace learning-based methods [2,3]. Sample-based methods involve modeling using image patches. These methods heavily rely on the types and quantity of patches, and they sometimes result in over-smoothing or excessive blurriness in generated images at patch boundaries. Subspace learning-based methods require less computational power but struggle to ensure consistency in identity features and texture details between generated and original images [5].

Goodfellow's GAN model [6], introduced in 2017, possesses impressive capabilities in image-to-image translation tasks. In the original GAN framework, there are two crucial components: the discriminator and the generator. The discriminator's role is to determine whether a given input is real or fake, while the generator continuously learns to produce clearer and more realistic sample images to deceive the discriminator's judgment. Currently, many improved models based on the GAN framework have been proposed. For instance, Pix2Pix [7] employs conditional GANs and supervised training to address image style transformation while retaining crucial information from the original images. However, training Pix2Pix requires paired databases with various image styles, and collecting such data can be challenging due to its scarcity. To overcome the difficulty of obtaining paired datasets, CycleGAN [8] was introduced. It simultaneously trains two GAN networks, transforming images into another style and then back, thus eliminating the need for paired images. Additionally, it utilizes cycle-consistency loss to preserve essential attributes between input and generated images.

However, CycleGAN still has room for improvement. It attempts to find the relationship between the source and target domains, with its generator and discriminator primarily focusing on style differences between these two domains. In the context of bidirectional FPSS, it overlooks the local characteristics and positional information of different regions within the face.

To address this issue, this paper introduces a novel cycle generative adversarial network (FSACycleGAN) based on semantic perception of portraits. Firstly, it leverages a pre-trained semantic segmentation network, U2-Net [9], to acquire semantic distribution images corresponding to the portrait dataset. Secondly, FSACycleGAN incorporates FSA residual blocks within the generator to inject the facial semantic information into the generation process. Lastly, by introducing an identity-aware loss between the facial semantic images and the generated images, the model's loss function is optimized to enhance the preservation of identity and texture information in the generated images. Additionally, in order

to improve the precision of image generation, this paper introduces a practical training approach: fine-tuning the recognition network with a fake dataset generated by FSACycleGAN. This iterative optimization process ultimately results in more realistic image generation by the model.

The structure is organized as follows: Section 2 presents the architecture, objective function, and training strategies of FSACycleGAN. Section 3 establishes the implementation details and an overview of the facial datasets and evaluation criteria. Subsequently, presents both quantitative and qualitative results, showcasing the effectiveness of the proposed method. Finally, Section 4 concludes the paper and points out further work.

II. METOHDS

A. *Face-Semantic-Aware CycleGAN Network Architecture*

The goal of the biphasic FPSS task is to find a mapping function between face photos and sketches, corresponding to paired samples in the image spaces X and Y. The objective of face sketch synthesis is to construct a mapping from the source photo domain X to the target sketch domain Y, and the overall mapping can be represented as X→Y. On the other hand, the objective of face photo synthesis is to build a mapping from the source sketch domain Y to the target photo domain X, and the overall mapping can be represented as Y→X. This paper aims to incorporate facial semantic information into this mapping process. To achieve this, a pre-trained semantic segmentation network, U2-Net [9], is utilized to obtain facial semantic distribution images, which form the domain of facial semantic information images, denoted as S. Therefore, in this paper, the overall mapping for face sketch synthesis can be represented as {X, S}→Y, while the overall mapping for face photo synthesis can be represented as {Y, S}→X.

Given a facial photo x, the corresponding facial semantic image s, and the corresponding facial sketch y, this model necessitates training two generators, G_X and G_Y. G_X transforms facial photos into sketches, while G_Y transforms facial sketches into photos. Additionally, this generator pair receives the facial semantic distribution image s and utilizes this additional information during the image generation process. The overall network framework is depicted in Fig. 1. The short dash lines in the diagram represent the photo-sketch-photo training process: Generator G_X takes facial photo x and the corresponding facial semantic image s, generating the corresponding fake sketch F_y. Then, Generator G_Y takes the fake sketch F_y and the corresponding facial semantic image s, producing the corresponding fake photo Cyc_x. The long dashed denote the sketch-photo-sketch training process, which is the mirror image of the former. The fake sketch generated by G_X is represented as F_y, and the fake photo generated by G_Y is denoted as F_x. The relationships are as follows:

$$F_x = G_Y(y, s), F_y = G_X(x, s). \quad (1)$$

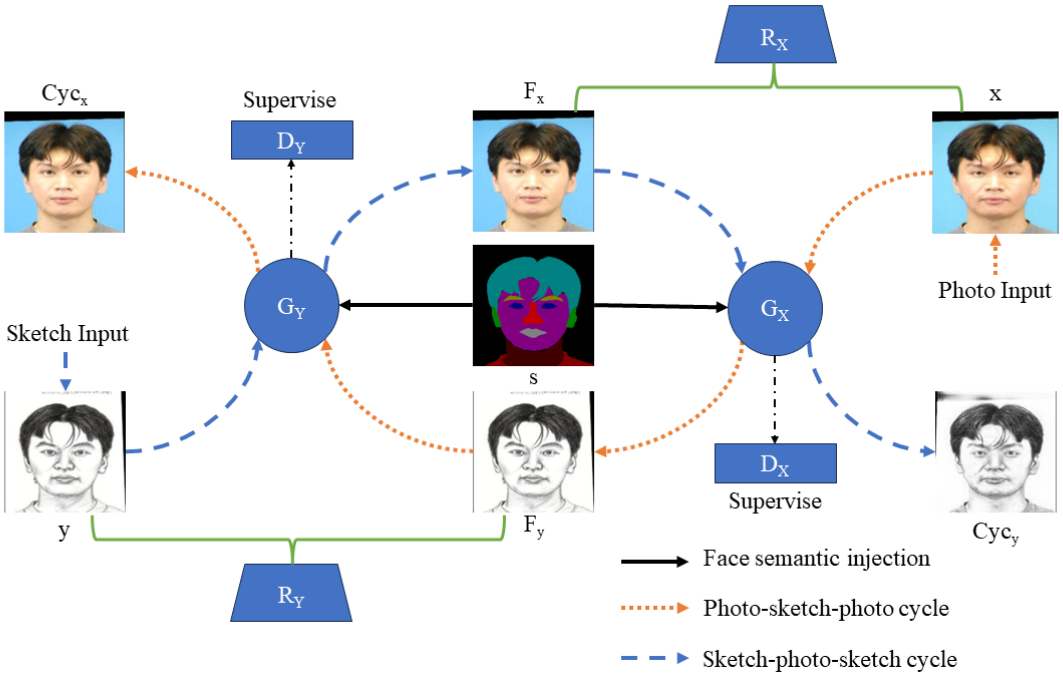

Fig. 1. Framework of the FSACycleGAN synthesis network.

To achieve a complete cycle in the image transformation process, the generated fake images are input into another generator:

$$Cyc_x = G_Y(F_y, s), Cyc_y = G_X(F_x, s). \quad (2)$$

The two recognition networks, R_X and R_Y, are employed to extract identity information from each individual's face. In this paper, the VGGFace model [13] is employed to extract high-level features from facial images. These features are then utilized by two fundamental recognition networks, R_X and R_Y, which are responsible for recognizing photos and sketches, respectively. These networks capture the most prominent facial characteristics and structures for identity recognition. This process directly regularizes the distance between real and fake images. The introduction of these recognition networks serves the purpose of recognizing the authenticity of facial images using the VGGFace model and determining whether two images depict the same individual. This provides additional perceptual supervision for the training of FSACycleGAN, ultimately enhancing its ability to generate facial images.

The generator of FSACycleGAN is based on an encoder-transformer-decoder architecture, as depicted in Fig. 2. It consists of three convolutional layers with a stride of 2 and padding mode set to "reflect," two FSA (Facial Semantic-Aware) residual blocks proposed in this paper (as shown in Fig. 3), and two deconvolutional layers with a stride of 2. Finally, a fake image is generated through a single convolutional layer.

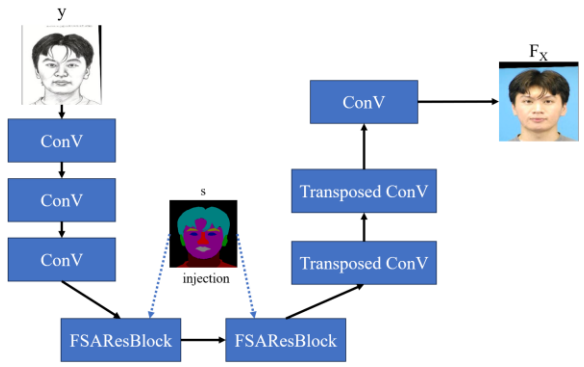

Fig. 2. The generator of FSACycleGAN.

FSACycleGAN employs a discriminator based on the Patch-GAN architecture [6], which is known to yield better results for high-resolution images. To strike a balance between image quality and training speed, the model utilizes patches of size 70x70 for discrimination. This choice effectively addresses issues such as image artifacts and blurring, while also requiring fewer parameters compared to larger-scale models [7].

Building upon the work of Shaosheng [14], this paper introduces the FSA residual block, as illustrated in Fig. 3(a). It comprises two FSA modules for information perception and injection, a 3x3 convolutional layer, a 1x1 convolutional layer, and two ReLU layers. In contrast to Qi et al. [12] and others, this paper includes a 1x1 convolutional layer at the end to enable full-channel information interaction. Furthermore, it restores the number of image channels to the same as before entering the FSA residual block. This enhancement empowers the FSA block with improved information interaction capabilities and greater adaptability.

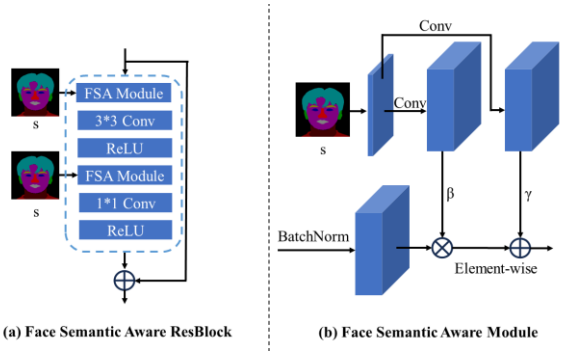

Fig. 3. (a) Face Semantic Aware ResBlock. (b) Face Semantic Aware Module.

The structure of the FSA (Face Semantic-Aware) module is illustrated in Fig. 3(b). Each FSA module has two inputs: data passed through a batch normalization layer from the preceding module and the forward activation features of the facial semantic image set S. This paper segments the facial semantic information into 10 classes, which include: background, hair, face, a pair of eyebrows, a pair of eyes, a pair of ears, nose, mouth, neck, and clothing. The relationships are as follows:

$$S = \{s(1), s(2), \cdots, s(c)\} \in Rh \times w \times c,$$
$$c \in [1,2,\cdots,10], s(c) \in [0,1]. \quad (3)$$

Here, c represents the ten classes of facial semantic segmentation, while h and w denote the height and width of the facial semantic information image, *s(c)* represents the classification label corresponding to a specific pixel. The module employs two convolutional layers to process facial semantic information and injects it into the generator through two hyperparameters, β and γ, using element-wise addition and multiplication. This injection method exhibits robustness across different sketch styles and input low-quality images [12].

B. Objective Function

The loss function in FSACycleGAN consists of four different components: adversarial loss, cycle consistency loss, identity-aware loss and identity mapping loss.

(1) **Adversarial Loss:** The adversarial loss is simultaneously applied to both the generator G_X and the discriminator D_Y to make the generated images convincingly resemble real images from the target domain. The generator G_X takes the input photo x and generates a fake sketch $G_X(x, s)$, which is then fed into the discriminator. The discriminator's task is to distinguish between the fake sketch $D_Y(G_X(x, s))$ and the real sketch

$D_Y(y)$. $\mathbb{E}_y[\log D_Y(y)]$ encourages the discriminator to assign a high probability to real images, aiming to make it as close to 1 as possible. And $\mathbb{E}_x\left[\log\left(1-D_Y\left(G_X(x,s)\right)\right)\right]$ encourages the discriminator to assign a low probability to fake images, aiming to make it as close to 0 as possible. Since CycleGAN is trained bidirectionally, a similar adversarial $\mathcal{L}_{GAN_Y}$ can be defined for the generator G_Y and discriminator D_X. The adversarial loss is represented as follows:

$$\mathcal{L}_{GAN_X}(G_X, D_Y) = \mathbb{E}_y[\log D_Y(y)] + \mathbb{E}_x\left[\log\left(1-D_Y\left(G_X(x,s)\right)\right)\right]. \quad (4)$$

(2) **Cycle Consistency Loss:** CycleGAN employs a supervision mechanism called cycle-consistency loss to prevent information loss. It assumes that the generated images can be transformed back to the source domain. For images in domain X, the expectation is that $G_Y(G_X(x,s),s) \approx x$, and for images in domain Y, the expectation is symmetric. The generator G_X takes the input photo x and generates a fake sketch $G_X(x,s)$. This fake sketch is then fed into the generator G_Y, producing $G_Y(G_X(x,s),s)$, which was previously referred to as Cyc_X. The training objective is to minimize the discrepancy between Cyc_X and x, aiming to make $\mathbb{E}_x[\|G_Y(G_X(x,s),s)-x\|_1]$ as close to 0 as possible. The same principle applies to the generator G_Y and discriminator D_X. This loss function is represented as follows:

$$\mathcal{L}_{cyc}(G_X, G_Y) = \mathbb{E}_x[\|G_Y(G_X(x,s),s)-x\|_1] + \mathbb{E}_Y[\|G_X(G_Y(y,s),s)-y\|_1]. \quad (5)$$

(3) **Identity-Aware Loss:** To ensure that the generated sketches and the target sketches exhibit a more similar identity specificity, this paper employs a pretrained VGGFace network [13] as a recognition network and feature extractor to obtain high-level representations of facial identity information. Here, l represents the pooling layer index, and ω^l represents the corresponding output features. For the sketch images, the process involves taking the difference between the output features of the fake sketch $\omega^l\left(G_X(x,s)\right)$ and the corresponding real sketch $\omega^l(y)$ from the same layer. Then, the squared L2 norm of the obtained result is computed, and the results from both layers are added together. A similar process is applied to the photos. It is represented as follows:

$$\mathcal{L}_{ip}(G_X, G_Y) = \sum_{l=1}^{2}\left\|\omega^l\left(G_X(x,s)\right)-\omega^l(y)\right\|_2^2 + \sum_{l=1}^{2}\left\|\omega^l\left(G_Y(y,s)\right)-\omega^l(x)\right\|_2^2 \quad (6)$$

(4) **Identity Mapping Loss:** By introducing pixel-level consistency between input images and generated images, the generator aims to make the generated images closely resemble the original images. Pixel-level consistency is introduced between the input image and the generated image, ensuring that the images generated by the generator are close to the original images. This is achieved by taking the difference between the generated image $G_X(x,s)$ and the input image x and then computing the L1 norm. The objective is to minimize the output of $\mathbb{E}_x[\|G_X(x,s)-x\|_1]$, aiming for it to be as close to 0 as possible. The same calculation is applied to the generator G_Y. The identity mapping loss is expressed as follows:

$$\mathcal{L}_{im}(G_X, G_Y) = \mathbb{E}_x[\|G_X(x,s)-x\|_1] + \mathbb{E}_Y[\|G_Y(y,s)-y\|_1]. \quad (7)$$

Summarizing the above, the overall objective function can be expressed as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{GAN_X} + \mathcal{L}_{GAN_y} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{ip}\mathcal{L}_{ip} + \lambda_{im}\mathcal{L}_{im}. \quad (8)$$

Where λ_{cyc}, λ_{ip}, and λ_{im} are hyperparameters that control the relative importance of each loss. The final objective can be formulated as a min-max problem:

$$G_X, G_y, D_X, D_Y = \arg\min_G \max_D \mathcal{L}_{total}. \quad (9)$$

C. *Training Strategy between the Synthesis and the Recognition Network*

Given that cyclic learning [21,22] and adversarial learning have proven to be effective model optimization techniques, and considering the research objective of bidirectional generation of face photos and sketches, this paper attempts to engage in adversarial training once again between the generative model FSACycleGAN and the recognition model VGGFace. This approach aims to further enhance the performance of the trained generative model.

The initial generative model, denoted as M_0, and the recognition model, VGGFace, denoted as R_0, are given. The original dataset CUFS is referred to as D_{real}, and the training epochs are denoted as I. For the first training iteration of M_0, we can begin by mitigating the interference from the recognition model R_0. In other words, we can train M_0 on D_{real} using a loss function that excludes $\mathcal{L}_{ip}$. This is because the pre-training of the recognition model R_0 is based on the VGGFace Dataset, and its ability to distinguish between real and fake data in the CUFS dataset is not robust. Train until stability is reached to obtain a well-trained model M_1 and generate a fake face photo-sketch dataset D_1. Set $i = 1$ to enter the loop. Use D_{real} and D_i to fine-tune VGGFace, resulting in R_{Xi} and R_{Yi} with enhanced recognition capabilities for photos and sketches, respectively on the CUFS dataset. Subsequently, Use the fine-tuned R_{Xi} and R_{Yi} to supervise the training of M_i, then obtain a more improved generative model M_{i+1} and a fake dataset D_{i+1}. Increment the i counter and repeat the loop until the specified number of training rounds is reached, and select the generative model and fake dataset with the best performance as the final result. Please refer to Algorithm 1.

Algorithm 1 Training Strategy between the Synthesis and the Recognition Network

Input:

Base synthesis network FSACycleGAN M_0;

Pretrained recognition network VGGFace R0;

Real photo-sketch dataset D_{real};

The number of iterations: I;

Output:

Optimized synthesis network M_I;

Optimized synthesis network M_I;

Recognition networks R_{XI} and R_{YI};

Fake photo-sketch dataset D_I;

1: Train M_0 on D_{real} according to FSACycleGAN's objective function without $\mathcal{L}_{ip}$, then generate fake photo-sketch dataset D_1 and well-trained M1.

2: **for** i = 1 to I **do**

3: Finetune R_0 using D_{real} and D_i and obtain two new recognition models R_{Xi} for photo and R_{Yi} for sketch.

4: Finetune FSACycleGAN M_i on D_{real} with R_{Xi} and R_{Yi} according to Eq. (8), then generate fake photo-sketch dataset D_{i+1} and well-trained M_{i+1}.

5: $i = i + 1$.

6: end for

III. EXPERIMENT

A. Experimental Environment and Model Parameters

For the sake of experimental rigor, both the FSACycleGAN and the baseline model should be trained on the same server. The model proposed in this paper was trained from scratch. The model was deployed on a Linux operating system with a CPU model of AMD Ryzen 5 5600X 6-Core Processor, 32GB of RAM, and an NVIDIA GeForce RTX 3070 graphics card. The experiments utilized the Adam optimizer with β1=0.5 and β2=0.999. The training lasted for 200 epochs, with the initial learning rate set to 0.0002 for the first 100 epochs and linearly decayed to 0 in subsequent training. Additionally, the batch size was set to 2, and the hyperparameters were configured as follows: λ_{cyc} =10 , λ_{ip} =10 , λ_{im} =5 , β=2 , γ=0.1. Finally, the number of iterations in Algorithm 1 was set to 2.

B. Datasets and Evaluation Criteria

The dataset used for the model is CUFS [1]. The CUFS dataset consists of 606 facial images, including:188 facial images from the CUHK student database, with 88 pairs of face-photo and face-sketch images used for training. 123 facial images from the AR database [16]. 295 facial images from the XM2VTS database [17]. For each facial sample in the dataset, there is a corresponding pair of a photograph and a sketch. These sketches are created based on photographs taken under 18 different lighting conditions and with neutral expressions by artists. To align the photographs and sketches geometrically, a strategy for photo-sketch alignment is employed, and the two images are concatenated in the order of photo-sketch. Finally, the input image size is resized to 256 × 256 using the reshaping and padding conventions mentioned in reference [18].

In terms of model performance, the experiments use Structural Similarity Index Measure (SSIM) [19] and Feature Similarity Index Measure (FSIM) [20] to evaluate the feature quality of the synthesized sketches. SSIM measures the structural similarity between images and aligns with human visual perception. It can demonstrate the perceptual similarity between synthesized results and real reference images. FSIM quantifies low-level similarity between paired images. It extracts features such as the phase congruency (PC) and gradient magnitude (GM) to assess image quality. Therefore, FSIM can be used to evaluate the blurriness and noise in the generated images.

Additionally, the experiments also validate the recognition accuracy of the generated photos and sketches using the VGGFace model. This metric can demonstrate the magnitude of differences between the generated photos and the original photos.

C. Results on Face Sketch Synthesis Task

The experiments showcase the performance of the model on the task of synthesizing face sketches using the CUFS dataset, along with some sample synthesized images. The experimental results are presented in Table I. Compared to other methods, the proposed model achieved the best performance in terms of SSIM and FSIM metrics on the CUFS dataset. The proposed method improved the SSIM metric to 0.6784 and the FSIM metric to 0.7732, indicating that the sketches generated by this method exhibit higher similarity and realism.

TABLE I. *COMPARISON OF SSIM AND FSIM SCORES FOR FOUR METHODS ON FACE SKETCH SYNTHESIS TASK*

	CycleGAN	***CNN***	***Pix2Pix***	***FSACycleGAN***
SSIM↑	0.5834	0.4958	0.6103	0.6784
FSIM↑	0.7329	0.6689	0.7132	0.7732

The experiments randomly selected some sketches generated by benchmark methods and FSACycleGAN, as shown in Fig. 4. It can be observed that due to the inclusion of FSA residual blocks, the sketches generated by FSACycleGAN exhibit significantly less blurriness and artifacts compared to those generated by benchmark methods. The texture of the hair is better preserved, and the sketch style is closer to real sketches.

D. Results on Face Photo Synthesis Task

Compared to the task of synthesizing face sketches, the task of synthesizing face photos is generally more challenging. The experimental results, as shown in Table II, demonstrate that FSACycleGAN achieves the best performance on the CUFS dataset in terms of SSIM and FSIM metrics compared to other benchmark methods. The proposed approach improves the SSIM metric to 0.7240 and the FSIM metric to 0.7934, indicating that the photos generated by FSACycleGAN exhibit higher similarity and realism.

TABLE II. *COMPARISON OF SSIM AND FSIM SCORES FOR FIVE METHODS ON FACE PHOTO SYNTHESIS TASK*

	CycleGAN	***CNN***	***Pix2Pix***	***FSACycleGAN***
***SSIM*↑**	0.6208	0.5285	0.6493	0.7240
***FSIM*↑**	0.7578	0.6939	0.7309	0.7934

The experiments randomly selected some baseline methods and sketches generated by FSACycleGAN, as shown in Fig. 5. It can be observed that the style of the sketches generated by FSACycleGAN is closer to real sketches, with reduced blurriness and artifacts compared to the baseline methods. Thanks to this method retaining more facial semantic information, the texture and shadows of the hair appear more natural, the facial features are generated more clearly, and most of the facial details are preserved.

E. Recognition for Different Methods

Table III shows the recognition accuracy of the four methods in generating images using the VGGFace model. In previous experiments, we have trained recognition models R_X and R_Y on the CUFS dataset. They can respectively serve as recognition models for the photo-matching task and the sketch-matching task. The sketch-matching task involves matching the fake sketches generated by FSACycleGAN with real photos in the dataset, while the photo-matching task is the opposite. There are a total of 518 image pairs in the test set. To ensure the reliability of the results, these methods were trained on the same computer to obtain accuracy. It can be observed that FSACycleGAN outperforms other baseline models in both photo and sketch recognition performance, achieving an accuracy of 99.42% in the sketch matching task and 98.64% in the photo matching task.

TABLE III. *THE SKETCH AND PHOTO MATCHING ACCURACY(%) OF THE IMAGES SYNTHESIZED BY THE FOUR METHODS ON VGGFACE*

Synthesis model	***CycleGAN***	***CNN***	***Pix2Pix***	***FSACycleGAN***
Recognition model	VGGFace			
Sketch matching	98.84	88.42	97.68	99.42 (515/518)
Photo matching	97.49	85.91	96.91	98.64 (511/518)

The comparison results from Tables III suggest that FSACycleGAN, due to its preservation of more facial semantic information through the FSA residual blocks in the image generation phase and the inclusion of identity-awareness loss and identity-mapping loss in the loss functions, attempts to maximize the preservation of both identity and texture features in the generated images. As a result, it outperforms all other competitors under the same recognition algorithms.

IV. CONCLUSIONS

This paper introduces a model called FSACycleGAN, based on CycleGAN, which incorporates facial semantic perception. By injecting additional facial semantic information into the image generation process, the model achieves improved performance on the problem of biphasic face photo-sketch synthesis (FPSS). In summary, this paper makes the following key contributions:

(1) Introduction of Face Semantic-Aware GAN (FSACycleGAN): The paper proposes a novel Generative Adversarial Network (GAN) that incorporates prior information about the semantic distribution of facial features. Unlike traditional CycleGAN models that rely solely on pairs of face photos and sketches as inputs, this work introduces pre-trained facial semantic distribution maps as additional input content, providing extra supervision to the generator. This enhances the generation of realistic face photos and sketches.

(2) Improvement of Loss Functions: The paper enhances the loss function of the CycleGAN model by introducing a novel perceptual loss term that measures the difference between the facial semantic distribution map and the generated images. This additional loss term contributes to improved training and better image quality.

(3) Effective Cycle Adversarial Training Strategy: The paper presents an effective training strategy that leverages a recognition network to retrain on photos and sketches generated by FSACycleGAN. This fine-tuned recognition model is then used to further train FSACycleGAN through adversarial training, resulting in a more effective model and more realistic images.

FSACycleGAN was trained and evaluated on the widely-used CUFS dataset. The proposed approach has shown significant improvements in terms of image quality and photo-sketch matching accuracy. Compared to CNN, CycleGAN, Pix2Pix, and DR-GAN, the images generated by this synthesis model achieved visual observations and quantitative metrics with an SSIM of 0.6784 and an FSIM of 0.7732. This represents a 16.3% improvement in SSIM and a 5.5% improvement in FSIM compared to the results obtained with CycleGAN. Furthermore, the recognition accuracy for generated images reached 99.42% and 98.64%, demonstrating better performance. While the proposed method has achieved significant performance improvements, this work still has its limitations. For instance, it has not been validated on complex datasets like CUFSF, and there is room for improvement in handling details such as scars and hair accessories in generated photos. Additionally, overfitting may occur during training

with simple datasets. Future work will continue to evaluate the performance of FSACycleGAN on other facial datasets and work on improving the FSA residual block architecture to better preserve scars and other facial decorations.

ACKNOWLEDGEMENT

This research was funded by the Research Fund of Jianghan University (Grant No. 2021kjzx005).

REFERENCES

[1] Xiaogang Wang and Xiaoou Tang, "Face Photo-Sketch Synthesis and Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1955–1967, Nov. 2009, doi: 10.1109/tpami.2008.222.

[2] Xiaoou Tang and Xiaogang Wang, "Face sketch synthesis and recognition," in Proceedings Ninth IEEE International Conference on Computer Vision, Nice, France, Jan. 2003. doi: 10.1109/iccv.2003.1238414.

[3] Qingshan Liu, Xiaoou Tang, Hongliang Jin, Hanqing Lu, and Songde Ma, "A Nonlinear Approach for Face Sketch Synthesis and Recognition," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, Jul. 2005. doi: 10.1109/cvpr.2005.39.

[4] C. Peng, N. Wang, J. Li, and X. Gao, "Universal Face Photo-Sketch Style Transfer via Multiview Domain Translation," IEEE Transactions on Image Processing, pp. 8519–8534, Jan. 2020, doi: 10.1109/tip.2020.3016502.

[5] M. Zhang, N. Wang, Y. Li, and X. Gao, "Deep Latent Low-Rank Representation for Face Sketch Synthesis," IEEE Transactions on Neural Networks and Learning Systems, pp. 3109–3123, Oct. 2019, doi: 10.1109/tnnls.2018.2890017.

[6] I. Goodfellow et al., "GAN（Generative Adversarial Nets）," Journal of Japan Society for Fuzzy Theory and Intelligent Informatics, pp. 177 - 177, Oct. 2017, doi: 10.3156/jsoft.29.5_177_2.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, Jul. 2017. doi: 10.1109/cvpr.2017.632.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," in 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Oct. 2017. doi: 10.1109/iccv.2017.244.

[9] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-Net: Going deeper with nested U-structure for salient object detection," Pattern Recognition, p. 107404, Oct. 2020, doi: 10.1016/j.patcog.2020.107404.

[10] L. Tran, X. Yin, and X. Liu, "Disentangled Representation Learning GAN for Pose-Invariant Face Recognition," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, Jul. 2017. doi: 10.1109/cvpr.2017.141.

[11] Hao Zhou, Zhanghui Kuang, and K. K. Wong, "Markov Weight Fields for face sketch synthesis," in 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, Jun. 2012. doi: 10.1109/cvpr.2012.6247788.

[12] X. Qi, M. Sun, W. Wang, X. Dong, Q. Li, and C. Shan, "Face Sketch Synthesis via Semantic-Driven Generative Adversarial Network," in 2021 IEEE International Joint Conference on Biometrics (IJCB), Shenzhen, China, Aug. 2021. doi: 10.1109/ijcb52358.2021.9484393.

[13] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," in Procedings of the British Machine Vision Conference 2015, Swansea, Jan. 2015. doi: 10.5244/c.29.41.

[14] C. Shaosheng, W. Lu, and Q. Xu, "Deep neural networks for learning graph representations," Proceedings of the ... AAAI Conference on Artificial Intelligence,Proceedings of the ... AAAI Conference on Artificial Intelligence, Feb. 2016.

[15] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," in Computer Vision – ECCV 2016,Lecture Notes in Computer Science, 2016, pp. 694–711. doi: 10.1007/978-3-319-46475-6_43.

[16] A. Nez and R. Benavente, "The AR Face Database".

[17] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maître, "XM2VTSDB: The Extended M2VTS Database," Jan. 1999.

[18] J. Yu et al., "Towards Realistic Face Photo-Sketch Synthesis via Composition-Aided GANs," Cornell University - arXiv,Cornell University - arXiv, Dec. 2017.

[19] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," IEEE Transactions on Image Processing, pp. 600–612, Apr. 2004, doi: 10.1109/tip.2003.819861.

[20] Lin Zhang, Lei Zhang, Xuanqin Mou, and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment," IEEE Transactions on Image Processing, pp. 2378–2386, Aug. 2011, doi: 10.1109/tip.2011.2109730.

[21] Jian Wang, Jie Yang, Wei Wu. Convergence of cyclic and almost-cyclic learning with momentum for feedforward neural networks. IEEE Transactions on Neural Networks, 22(8): 1297-1306, 2011.

[22] Jian Wang, Zhenyun Ye, Weifeng Gao, Jacek M. Zurada. Boundedness and Convergence Analysis of Weight Elimination for Cyclic Training of Neural Networks. Neural Networks, 82: 49-61, 2016.

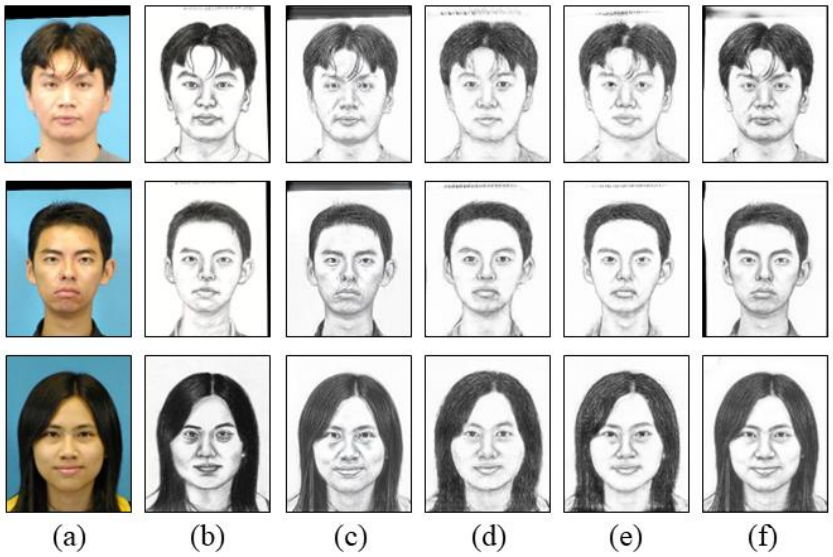

Fig. 4. Real photo(a) and Ground truth(b) and Sample results for photo-to-sketch synthesis on the CUFS generated by CycleGAN(c), CNN(d), Pix2Pix(e) the proposed FSACycleGAN(f).

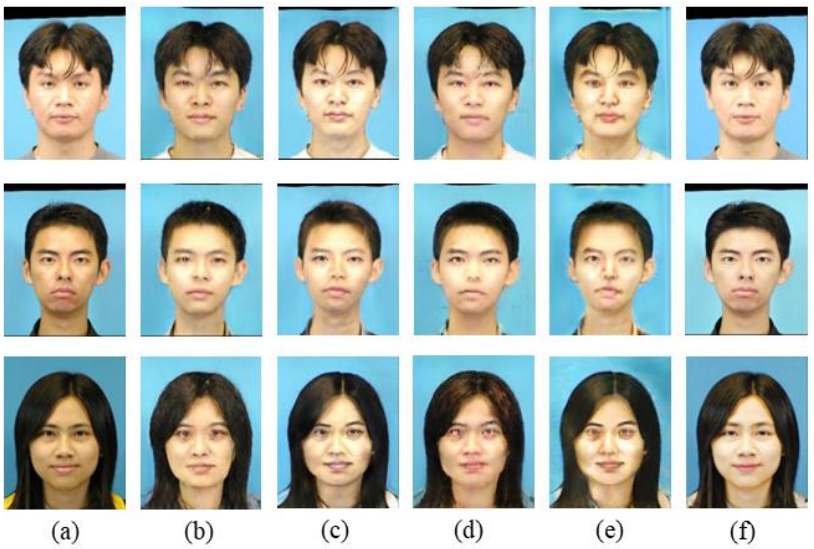

Fig. 5. Ground truth(a) and Sample results for sketch-to-photo synthesis on the CUFS generated by DR-GAN(b), CycleGAN(c), CNN(d), Pix2Pix(e) the proposed FSACycleGAN(f).