PMSGAN: Parallel Multistage GANs for Face Image Translation

Changcheng Liang, Mingrui Zhu, Nannan Wang, *Member, IEEE*, Heng Yang, and Xinbo Gao, *Senior Member, IEEE*

***Abstract*— In this article, we address the face image translation task, which aims to translate a face image of a source domain to a target domain. Although significant progress has been made by recent studies, face image translation is still a challenging task because it has more strict requirements for texture details: even a few artifacts will greatly affect the impression of generated face images. Targeting to synthesize high-quality face images with admirable visual appearance, we revisit the coarse-to-fine strategy and propose a novel parallel multistage architecture on the basis of generative adversarial networks (PMSGAN). More specifically, PMSGAN progressively learns the translation function by disintegrating the general synthesis process into multiple parallel stages that take images with gradually decreasing spatial resolution as inputs. To prompt the information exchange between various stages, a cross-stage atrous spatial pyramid (CSASP) structure is specially designed to receive and fuse the contextual information from other stages. At the end of the parallel model, we introduce a novel attention-based module that leverages multistage decoded outputs as in situ supervised attention to refine the final activations and yield the target image. Extensive experiments on several face image translation benchmarks show that PMSGAN performs considerably better than state-of-the-art approaches.**

***Index Terms*— Atrous spatial pyramid, face image translation, generative adversarial networks, parallel multistage.**

I. Introduction

IMAGE-TO-IMAGE translation is a meaningful and active field of computer vision and has achieved many surprising applications [1]. Specifically, a vital application of image-to-image translation is to generate a new style of face images that can be used for digital entertainment and animation production. However, synthesizing visually realistic and semantically plausible images while surpassing the considerable discrepancies (color, texture, and shape) barrier is highly challenging.

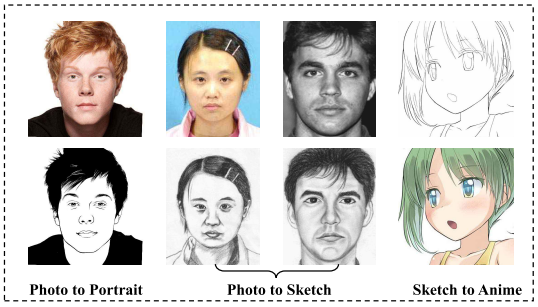

Fig. 1. Face images translation samples.

This article mainly focuses on the translation of a highly abstract style (as shown in Fig. 1). It is different from general image (such as scenery and anime) translation in that it is more sensitive to facial features and has strict standards for the consistency of facial structure, which means the flaws (missing or redundant lines) in faces will become more apparent, and small traces (e.g., around the mouth) may also be noticed. Under strict semantic constraints, the task of face image translation is challenging. Therefore, synthetic images of conventional image-to-image translation studies [2], [3] are far from satisfactory. Recently, a series of researches for specific style translation of face images is proposed that utilize the prior information or make various strategies for different facial regions. With the additional prior information and useful strategies, certain progress has been made by these methods in face image translation.

However, the loss of texture details in synthetic images still exists. We think it is caused by the following two reasons: 1) imperfect prior acquisition methods or coarse stacking strategies may lead to the loss of facial information and damage the quality of the generated image and 2) sampling operation during encoding and decoding will lead to the loss of contextual details of latent features. A large proportion of methods [4] uses skip connections to connect the activations in the encoder and decoder to preserve the contextual information. Even so, the problem is still not completely solved

Manuscript received 22 July 2022; revised 25 October 2022; accepted 24 December 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0103202; in part by the National Natural Science Foundation of China under Grant U22A2096, Grant 62106184, Grant 62036007, Grant 62176198, and Grant 62206211; in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2021JQ-198; in part by the Technology Innovation Leading Program of Shaanxi under Grant 2022QFY01-15; in part by the Open Research Projects of Zhejiang Laboratory under Grant 2021KG0AB01; and in part by the Fundamental Research Funds for the Central Universities under Grant XJS210102. *(Corresponding author: Nannan Wang.)*

Changcheng Liang, Mingrui Zhu, and Nannan Wang are with the State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China (e-mail: ccliang_xd@163.com; mrzhu@xidian.edu.cn; nnwang@xidian.edu.cn).

Heng Yang is with Shenzhen AiMall Tech, Shenzhen 518000, China (e-mail: yanghengnudt@gmail.com).

Xinbo Gao is with the Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: gaoxb@cqupt.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3233025.

Digital Object Identifier 10.1109/TNNLS.2022.3233025

because there is no strict pixel-level spatial correspondence between the encoded and decoded features which belong to different domains.

To tackle the above problems, we revisit the coarse-to-fine strategy and modify the general single-stage encoding–decoding framework to present a novel GAN-based [5] parallel multistage encoding–decoding architecture specifically for face image translation. The proposed model includes three parallel encoding–decoding stages that take three source images with high-to-low spatial resolution as inputs and generate three outputs, which can make the learning focus of each stage different and significantly reduce the training difficulty. To diminish information loss of the contextual details of the encoding–decoding process and improve multistage synergy, a cross-stage atrous spatial pyramid (CSASP) structure is specially designed for each stage to receive and fuse the contextual information from other stages. Meanwhile, a multiscale supervised attention module is introduced to gather and utilize multistage decoded outputs as in situ supervised attention to refine the final activations and improve the quality of the final output. The contributions of this work are summarized as follows.

1) We propose a novel parallel multistage architecture for face image translation that is capable of synthesizing contextually enriched and spatially accurate outputs without additional prior information.

2) The proposed CSASP structure enables the information of multiple stages to flow with each other and improves the efficiency of feature fusion.

3) A multiscale supervised attention module that utilizes multistage decoded outputs as supervision is proposed to refine the final activations and improve the quality of the final output.

4) Extensive experiments on several face image translation benchmarks demonstrate the superiority of the proposed method over state-of-the-art methods.

II. Related Work

A. General-Purpose Image-to-Image Translation

Supervised image-to-image translation methods aim to translate images across domains (e.g., summer-to-winter, day-to-night, and edge-to-image). They often utilize conditional generative adversarial networks [2] to learn a mapping from input to output images. Several works extend it to deal with superresolution [6] or video generation [7]. In terms of unsupervised image-to-image translation methods without paired databases, CycleGAN [3], DiscoGAN [8], and DualGAN [9] preserve key attribution between the input and the output by using a cycle-consistency loss. Based on CycleGAN, MUNIT [10] and DRIT++ [11] enable multimodel translations by decomposing the latent feature of images into a domain-specific style space and a domain-invariant content space to obtain diverse outputs. Another line of methods improves CycleGAN to achieve transformation across multiple domains at the same time, such as StarGAN [12]. Some studies pay attention to translation between domains with a larger difference. For example, CoupledGAN [13] and UNIT [14] utilize domain-sharing latent space, and U-GAT-IT [15] focuses on attention modules for feature choice.

B. Face Image Translation

Face image translation comprises various subtasks. Face photo-sketch synthesis [16], [17], [18], [19] is an important task that has been studied for a long time. Existing works for face photo-sketch synthesis can be mainly divided into two categories. Exemplar-based methods reconstruct target images by mining correspondences between input images (image patches) and images (image patches) in a reference set of photo-sketch pairs. Deep-learning-based methods attempt to predict the target image pixels from the source image pixels through end-to-end convolutional neural networks. Exemplar-based methods can be further grouped into three types: subspace learning-based approaches [20], sparse representation-based approaches [21], and Bayesian inference-based approaches [22]. A detailed overview of existing exemplar-based methods can be found in [1]. Recently, CNN-based and GAN-based approaches have emerged as promising paradigms for face photo-sketch synthesis. Initial effort [23] trained an end-to-end fully convolutional neural network (FCN) for directly modeling the nonlinear mapping between face photographs and face sketches. Limited by shallow layers and pixel-level loss, however, it fails to capture texture details and fails to preserve reasonable structures. Several works follow ideas from image-to-image translation and focus on improving face photo-sketch synthesis performance by adding prior information. PS2MAN [24] proposes a multiscale discriminator to provide adversarial supervision on different image resolutions. SCAGAN [25] introduces facial composition information as additional input to help the generation of sketch portraits and proposes a compositional loss based on facial composition information. To tackle the problem of insufficient paired training data, Wild [26] proposes a semisupervised learning method to augment paired training samples by synthesizing pseudo-sketch features of additional training photographs and learns the mapping function between them. Sketch-Transformer [27] proposes to learn the key elements of the Transformer architecture and adapt them to the face photo-sketch synthesis task. Although great progress has been made by the above approaches, undesirable artifacts and distorted structures, however, are still exist, especially in the results of real scenarios.

To generate artistic portrait drawings, APDrawingGAN [28] introduces an architecture that comprises hierarchical generators and discriminators that combines both global networks and local networks. APDrawingGAN++ [29] is an extended version of APDrawingGAN, which further introduces a classification-and-synthesis approach for lips and hair. U^2-Net [4] is proposed for the segmentation task, but also has a promising performance in portrait drawing generation. A popular line of research focuses on the task of face image cartoonization. AniGAN [30] proposes a double-branch discriminator to learn both domain-specific distributions and domain-shared distributions. Pixel2style2pixel [31] has been

shown to perform well in tasks such as multimodal conditional image synthesis, face frontalization, and superresolution.

C. Multistage Strategies

Among existing works, multistage strategies have achieved great success in various directions. Many algorithms [32], [33], [34], [35], [36] adopt the multistage strategy that disintegrates the general synthesis process into multiple stages by transferring the current decoding feature or reconstructed image to the next stage subnetwork as part of the input, and continuously improve the quality of the synthetic image in the form of stacked subnetworks. Such a design is effective since it decomposes challenging vision tasks into smaller, more manageable subtasks. In the unconditional image generation task, SinGAN [37] uses the strategy that stacked subnetworks and transfers the current generated image to the next subnetwork. In the image translation tasks, SCAGAN [25] uses the same multistage strategy as SinGAN. In the image deblurring task, MT-RNN [38] proposes a strategy that stacked subnetworks and transfers the feature maps from the decoder at the previous iteration to the encoder at the next iteration. In the image restoration task, MPRNet [39] uses the stacked strategy and transfers the current decoded feature to the next subnetwork. However, the common practice that connects each subnetwork in series will lead to suboptimal results and high computational costs. In this work, we rethink the multistage strategy and propose a multistage learning strategy in which each subnetwork learns image translation mapping at the same time in the training stage. Different from the previous work, the parallel multistage training strategy proposed in this work enables information to flow between multiple substages, promotes information exchange between the multistage subnetworks, and improves the quality of the synthetic image of the subnetworks at each stage. At the same time, the parallel multistage strategy proposed in this work can greatly reduce the training and testing time of the multistage strategy.

D. Attention Mechanisms

Attention mechanisms [40], [41] can model long-range dependencies, which have played a key role in many tasks in computer vision and machine learning including image classification [42], [43], image segmentation [44], [45], neural machine translation [46], image and video captioning [47], [48], and visual question answering [49]. With the use of the attention mechanism, all tasks have achieved performance improvement. Kuen et al. [50] propose a recurrent attentional convolutional-deconvolution network for saliency detection. This supervised model uses an iterative approach to attend to selected image subregions for saliency refinement in a progressive way. Wang et al. [42] propose a residual attention network for image classification with a trunk-and-mask attention mechanism. Recent studies [51], [52] show that the incorporation of attention learning in GAN-based models leads to more realistic images in image-to-image translation tasks. In this article, we design a multiscale supervised attention module, which can cooperate reasonably with the proposed parallel multistage networks and significantly improve the quality of the final output.

III. Method

In this section, details of the proposed parallel multistage framework are presented. First, we describe the application details of parallel multistage structure in face image translation. Then, the loss functions used to train the proposed model are provided.

A. Parallel Multistage GANs for Face Image Translation

Given paired training samples $\{(x_i, y_i) \in (X, Y)\}_{i=1}^{N}$, the target of face image translation is to translate source images of domain X into target images that obey the distribution of domain Y. To improve the quality of synthetic images and improve the efficiency of models, this work uses a three-stage parallel multistage strategy to improve the quality of synthetic images in the image translation tasks. The pipeline of the proposed PMSGAN is shown in Fig. 2. It consists of five closely related parts, including: 1) an ordinary encoder E and a decoder D composed of a 3×3 convolution layer and five modified residual blocks; 2) an encoded feature fusion (EFF) module; 3) a CSASP module; 4) a decoded feature fusion (DFF) module; and 5) a multiscale supervised attention (MSSA) module. The EFF, CSASP, and DFF modules cooperate closely to facilitate asymmetric information to flow efficiently between different stages. The input of multiple stages is images with gradually decreasing spatial resolution. We resize the original input x_i to obtain an input image x_i^2 that is half the height and width of x_i and an input image x_i^3 that is a quarter the length and width of x_i. We obtain y_i^2 and y_i^3 from y_i in the same way.

1) Encoded Feature Fusion: Before inputting the low-resolution input x_i^j $(j = 2, 3)$ into the encoder, the encoded features of the high- and low-resolution stages are fused across stages by the EFF module. The EFF module takes E_{j-1}^{out} and x_i^j as inputs, as shown in Fig. 3. It first extracts the dense feature by inputting the low-resolution input into the dense blocks [53]. Then, it applies a convolution layer with a stride of 2 to E_{j-1}^{out} and obtains $(E_{j-1}^{\text{out}})^{\downarrow}$, which has the same size as the dense feature. For the fusion of the dense feature and $(E_{j-1}^{\text{out}})^{\downarrow}$, the EFF module applies the feature attention structure as [35] to actively emphasize or suppress features of previous scales and learns the spatial importance of features from the dense feature. More specifically, $(E_{j-1}^{\text{out}})^{\downarrow}$ and the dense feature are element-wise multiplied. Then the multiplied features are passed through a 3×3 convolution layer, which output is expected to include complementary information for translation. The output is finally added to $(E_{j-1}^{\text{out}})^{\downarrow}$ to be further refined through the following encoder.

2) Cross-Stage Atrous Spatial Pyramid: The feature salient points extracted by the network in different resolution stages are different. The middle features in the high-resolution stage emphasize more small texture details, and the middle features in the low-resolution stage emphasize more the general contour of the face. To effectively fuse the features of different resolution stages, we design the CSASP module. The architecture

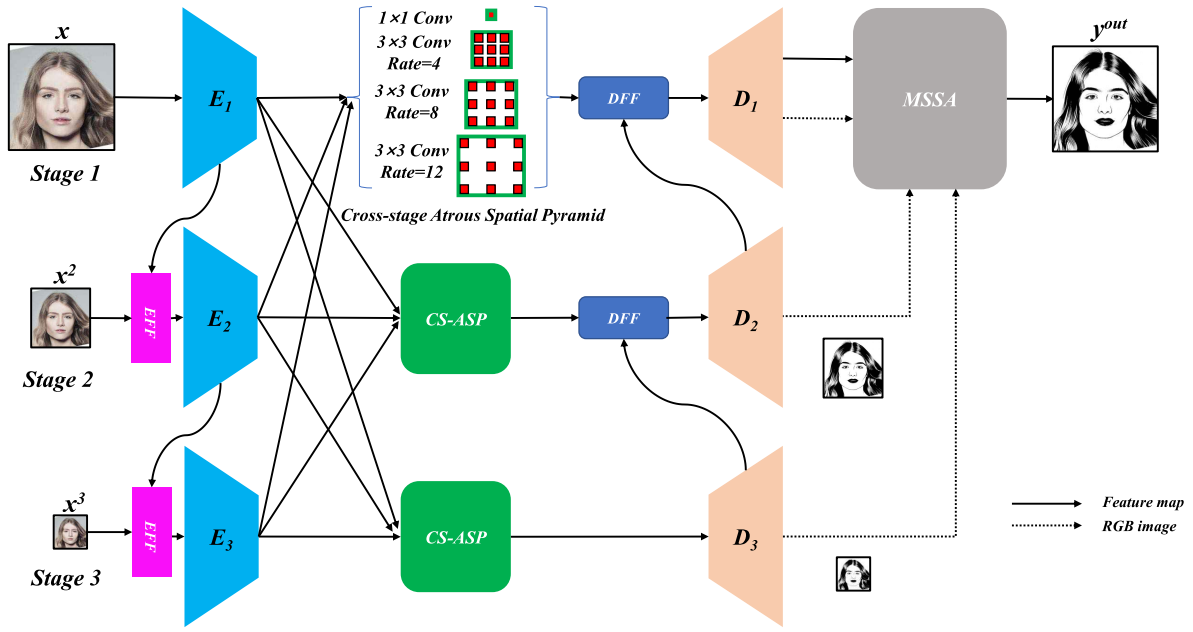

Fig. 2. Network architecture of the proposed PMSGAN framework.

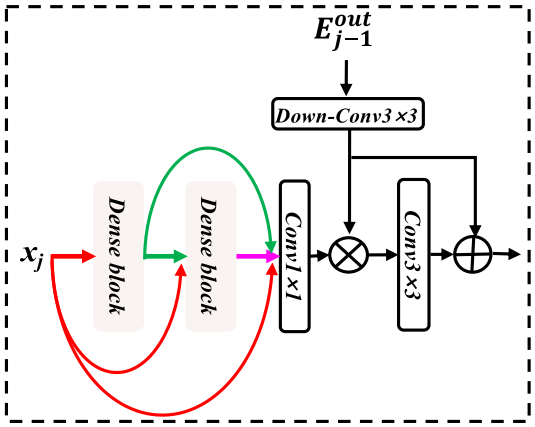

Fig. 3. Structure of the EFF module.

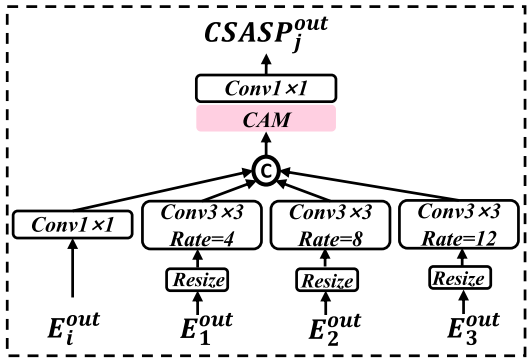

Fig. 4. Structure of the CSASP module.

of the CSASP module is shown in Fig. 4. It receives the encoded features $E_j^{\text{out}}(j = 1, 2, 3)$ of all three stages in an efficient manner. For the features of the high-resolution stage, it uses low-rate convolution to process and captures more texture details through small receptive fields. The difference is that the features of the low-resolution stage utilize high-rate convolution to process and capture a more complete facial structure through a large receptive field. At the same time, a 1×1 convolution branch is used to replace the convolution of a large enough rate for the features of the current stage to obtain the global information. The recalibrated features are concatenated and then passed through a channel attention module to reevaluate the importance of each channel of the concatenated features from all three stages. Finally, it utilizes a 1×1 convolution layer to reduce the number of channels of the cross-stage features.

3) Decoded Feature Fusion: To promote the two-way flow of information between stages, in the decoding stage, we design the DFF module to fuse the decoded features in the low- and high-resolution stages. The architecture of the DFF module is shown in Fig. 5. We apply the feature attention structure to actively emphasize features of low-resolution stages. More specifically, $(E_{j-1}^{\text{out}})^{\downarrow}$ and the dense feature are element-wise multiplied. Then the multiplied features are passed through a 3×3 convolution layer, which output is expected to include global structure information.

4) Multiscale Supervised Attention: After the parallel multistage decoding process as described above, we initially obtained multiscale translated images $D_{j(\text{RGB})}^{\text{out}}$, $(j = 1, 2, 3)$. To further strengthen the connection between each stage and improve the translation performance, we introduce a multiscale supervised attention module following the decoder. The illustration of the proposed MSSA module is shown in Fig. 6. With the help of supervised prediction of multiscale translated images, we generate attention maps to suppress the less informative features, retain useful features, and improve the quality of translated images. Specifically, the MSSA module

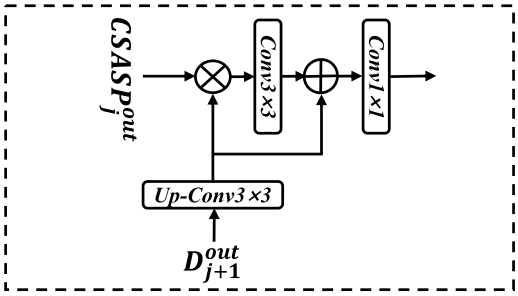

Fig. 5. Structure of the decode feature fusion module.

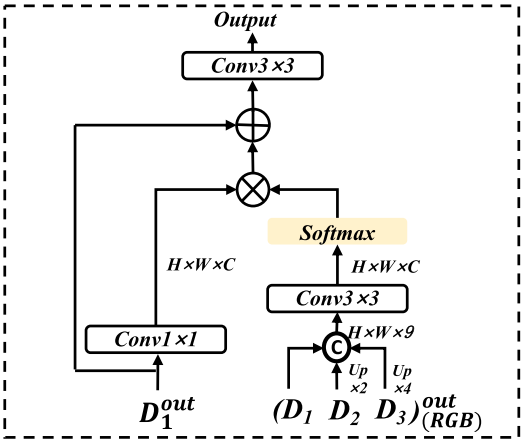

Fig. 6. Structure of the multiscale supervised attention module.

takes the translated images $D_{j(\mathrm{RGB})}^{\mathrm{out}}$ of all three stages as input and concatenates them. Then, per-pixel attention masks $M \in \mathbb{R}^{H\times W\times C}$ are generated from the concatenated feature using a 3×3 convolution followed by the Softmax activation. These masks are then used to reweight the decoded feature D_1^{out}, resulting in attention-guided features with multiscale information. The attention-guided features are added to the identity mapping path. Finally, a 3×3 convolution layer is applied to obtain the final synthetic image y^{out}.

B. Loss Function

The total loss of our model consists of two loss functions: adversarial loss and perceptual loss.

1) Adversarial Loss: To constrain the distribution of the generated images to be close to the real target domain distribution, we apply three independent discriminators D_{Y_j} $(j=1,2,3)$ to the outputs of the three scales of the generator (y^{out}, $D_{2(\mathrm{RGB})}^{\mathrm{out}}$, and $D_{3(\mathrm{RGB})}^{\mathrm{out}}$). L_{adv} is formulated as

$$\mathcal{L}_{\mathrm{adv}} = \mathbb{E}_y\left[\left(D_{Y_1}(y)\right)^2\right] + \mathbb{E}_x\left[\left(1 - D_{Y_1}\left(y^{\mathrm{out}}\right)\right)^2\right] + \sum_{j=2}^{3}\left[\mathbb{E}_y\left[\left(D_{Y_j}\left(y^j\right)\right)^2\right] + \mathbb{E}_x\left[\left(1 - D_{Y_j}\left(D_{j(\mathrm{RGB})}^{\mathrm{out}}\right)\right)^2\right]\right]. \tag{1}$$

2) Perceptual Loss: We introduce the multiscale perceptual loss [54] for the outputs of the three scales of the generator (y^{out}, $D_{2(\mathrm{RGB})}$, and $D_{3(\mathrm{RGB})}$) to ensure that the generated images and their ground truth are similar in semantic feature level

$$\mathcal{L}_p = \mathbb{E}_x\left[\frac{1}{C_k H_k W_k}\left|\left|\phi_k\left(y^{\mathrm{out}}\right) - \phi_k(y)\right|\right|_1\right] + \sum_{j=2}^{3}\mathbb{E}_x\left[\frac{1}{C_k H_k W_k}\left|\left|\phi_k\left(D_{j(\mathrm{RGB})}^{\mathrm{out}}\right) - \phi_k\left(y^j\right)\right|\right|_1\right] \tag{2}$$

where ϕ_k indicates feature maps of the kth layer of a pretrained VGG-19 model [54], and C_k, H_k, and W_k indicate the channel numbers as well as the height and width of the feature maps, respectively.

3) Full Loss: By combining the above losses, we can achieve our full loss

$$\mathcal{L}_{\mathrm{full}} = \lambda_1\mathcal{L}_{\mathrm{adv}} + \lambda_2\mathcal{L}_p. \tag{3}$$

Referring to previous work, we empirically set $\lambda_1 = 1$ and $\lambda_2 = 5$ to keep corresponding losses in the same order of magnitude. The influence of coefficient change within a certain range on the experimental results is not particularly significant.

IV. Experiments

In this section, we first discuss the experimental settings. Then, we conduct an ablation study to quantify the contribution of different configurations to the overall effectiveness. Then, we qualitatively and quantitatively compare our results with state-of-the-art methods. Finally, we analyzed the runtime, robustness, and failure cases, respectively.

A. Experimental Settings

1) Implementation Details: All models are trained on an NVIDIA GeForce RTX3090 GPU using the Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.99$. The learning rate was fixed at 0.0002. The batch size was set to 1 for all experiments. Weights were initialized from a Gaussian distribution with a mean of 0 and a standard deviation of 0.02. For the CUFS and CUFSF databases, we scaled the size of the input images to 256×256 and normalized the pixel value to the interval $[-1,1]$ before putting them into the model. For the APDrawing and Sketch2Anime databases, we scaled the size of the input images to 512×512 and normalized the pixel value to the interval $[-1,1]$. During training, we update the generator and discriminator alternatively at every iteration.

2) Database: The experiments are conducted on four databases: 1) the CUFS database [55]; 2) the CUFSF database [56]; 3) the APDrwaing database [28]; and 4) a newly collected Sketch2Anime database. The CUFS database consists of 188 identities from the Chinese University of Hong Kong (CUHK) student database [57], 123 identities from the AR database [58], and 295 identities from the XM2VTS database [59]. Each identity has a photo-sketch pair under normal light conditions and neutral expression. The CUFSF database has 1194 identities from the FERET database [60]. There is a photograph with illumination variation and a sketch with an exaggerated structure for each identity. Therefore, face image translation in the CUFSF database is more challenging than in the CUFS dataset. All images of the CUFS database and the CUFSF database are processed by aligning the center of

TABLE I
PARTITION SETTINGS OF THE DATABASES

Database		Training Pairs	Testing Pairs
CUFS	CUHK Student	88	100
	AR	80	43
	XM2VTS	100	195
CUFSF		250	944
APDrawing		70	70
Sketch2Anime		130000	5509

two eyes to the fixed position and cropping to the size of 200 × 250. We divide the training set and the test set in the same way as [22]. The APDrawing database includes 140 pairs of face photographs and corresponding portrait drawings. To make the training set distribution more consistent, all portrait drawings were drawn by a single professional artist. All images and drawings of the APDrawing database are aligned and cropped to the size of 512 × 512. We divide the training set and the test set in the same way as [28]. We further collect a Sketch2Anime database in which the anime images are from Danbooru2018 [61] and the corresponding sketches are generated by [62]. All anime images and sketches are aligned and cropped to the size of 512 × 512. Finally, there are 135 509 pairs of images in total. We apply the database to the more difficult sketch to anime task. 130 000 pairs of images are randomly chosen for training and the remaining 5509 pairs are used for testing. The configurations of the experimental databases are shown in Table I.

3) Baselines: On the CUFS and CUFSF databases, we compare our method with five state-of-the-art methods: pix2pix [2], CycleGAN [3], PS2MAN [24], Wild [26], SCA-GAN [25], Pixel2style2pixel [31], and Sketch-Transformer [27]. On the APDrawing database, we compare our method with five state-of-the-art methods: pix2pix [2], CycleGAN [3], APDrawingGAN [28], APDrawingGAN++ [29], U^2-Net [4], and Pixel2style2pixel [31]. On the Sketch2Anime database, we compare our method with four state-of-the-art methods: pix2pix [2], CycleGAN [3], UGATIT [15], and DRIT++ [11]. In particular, to ensure the fairness of the comparative experiment, we use paired data to train all supervised and unsupervised methods.

4) Evaluation Metrics: Evaluating the quality of synthetic images is an open and difficult task [63]. Classic measures, such as the pixel-level L2 Euclidean distance, cannot assess structured outputs such as images, as they assume pixelwise independence; therefore, their evaluation conclusions are often inconsistent with human visual perception. The structural similarity index metric (SSIM) [64] and peak signal-to-noise ratio (PSNR) were frequently used to evaluate the performance of exemplar-based methods. However, we find that they are not suitable for evaluating deep-learning models. One phenomenon is that blurry and smooth synthetic images tend to get a higher SSIM or PSNR score, which is contrary to the human visual perception that is biased toward sharper images.

A perceptual metric that measures the similarity of two images in a way that coincides with human judgment is challenging. This crux has been fully studied by Zhang et al. [65]. They collected a large-scale database of human judgments and evaluated key questions about image quality evaluation metrics. The most important conclusion is that deep network activations work surprisingly well as a perceptual similarity metric. Based on this discovery, they proposed a learned perceptual image patch similarity (LPIPS) metric by adding a linear layer on top of off-the-shelf classification networks (SqueezeNet [66], AlexNet [67], and VGG [54]). The LPIPS takes two images (image patches) as the input, calculates the L2 distance between their normalized deep feature embeddings, and predicts the perceptual judgment score through the linear layer. We utilize three variants [LPIPS(alex), LPIPS(squeeze), and LPIPS(vgg)] provided by the authors (version 0.1 in [68]) to evaluate the perceptual similarity between synthetic and real images. A lower score indicates better quality of synthetic images.

TABLE II
ABLATION STUDY: FID, FSIM, AND LPIPS SCORES FOR DIFFERENT VARIANTS OF CONFIGURATIONS, EVALUATED ON THE APDRAWING DATABASE

Configurations	FID ↓	FSIM ↑	LPIPS(vgg) ↓
(a)	61.13	0.7441	0.2522
(b)	59.17	0.7505	0.2483
(c)	56.11	0.7526	0.2462
(d)	50.42	0.7641	0.2238
(e)	**46.71**	**0.7754**	**0.2137**

The Fréchet inception distance (FID) [69] is representative sample-based evaluation metrics for GANs. FID is designed to capture the Fréchet difference between two Gaussians (synthetic and real-world images). We compute the FID score between the synthetic images and real ones. Lower FID scores indicate better-quality synthetic images. Notably, although FID can well evaluate the quality of natural images and has become a commonly used metric in face image translation tasks, it is not very suitable for the task of face sketch synthesis, which does not pay attention to the diversity of generated images.

The feature similarity index (FSIM) [70] is a commonly used metric for full-reference image quality assessment, which captures the similarity between low-level features of images. It shows higher consistency with human visual perception compared with SSIM [65]. We calculated the average FSIM score between synthetic images and real ones. A higher FSIM score indicates better-quality synthetic images.

B. Ablation Study

Under different experimental configurations, we compute the average LPIPS score between the translated images and ground truth on the APDrawing test database. We conduct the ablation study on five configurations: 1) a single-stage encoding–decoding baseline model that does not use any components during training; 2) using the proposed parallel multistage encoding–decoding model with the EFF module; 3) adding a DFF module based on (2); 4) adding CSASP structure based on (3); and (5) adding MSSA module based on (4). The comparison results are shown in Table II and discussed below.

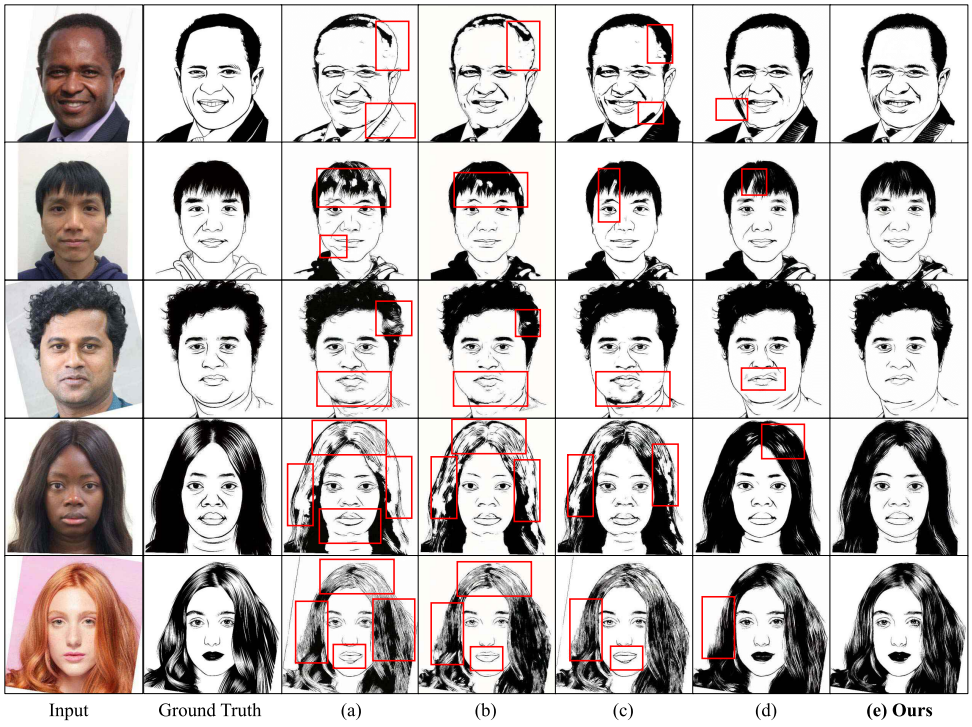

Fig. 7. Some synthetic portrait drawings synthesized by different configurations in the ablation study.

Fig. 7 shows some synthetic portrait drawings synthesized by different configurations in the ablation study. We can get the following conclusions by comparing and analyzing the results. The single-stage architecture will produce distorted patches on the hair and black flecks on the face. In the case of configuration (b), distorted patches on the hair are smaller than (a). It means that the EFF module can actively emphasize or suppress features of previous stages and progressively learn the translation function by disintegrating the general synthesis process into multiple parallel stages, thus showing a higher performance. In the case of configuration (c), distorted patches are greatly reduced, but there are still unreasonable patches on the face. It means that the DFF module facilitates the flow of information between stages. Configuration (d) achieves a great improvement in LPIPS scores. The distorted patches on the face have been improved, the lips have the correct color, and the distorted patches in the hair can also be ignored, but the white lines on the hair are not realistic enough. It means that the CSASP structure can gather contextual information about all stages and promote linkages between them. In the case of configuration (e), unreasonable patches no longer appear on the face, and the lines of hair are more realistic. By adding the MSSA module to further strengthen the connection between each stage and improve the translation performance, the full model can achieve the best performance.

Overall, we can get the conclusion that each module of the overall framework plays a significant role and jointly promotes the excellent performance of the overall model.

C. Comparison With Baselines

1) Qualitative Comparison: Fig. 8 shows some synthetic face photographs from different methods on the CUFS database and the CUFSF database. The results of pix2pix, CycleGAN, PS2MAN, and SCAGAN have noticeable artifacts and noise. Pixel2style2pixel is unable to reconstruct facial details and has poor visual quality. Sketch-Transformer has made great improvements in the details of the reconstructed image, but there are still some defects in the beard and hair that cannot be ignored. By comparison, the proposed method can generate photographs with the most reasonable texture distribution and considerable structure and therefore has the best quality. Fig. 9 shows some synthetic face sketches from different methods on the CUFS database and the CUFSF database. The results of FCN and DGFL are too blurry. The Pixel2style2pixel cannot synthesize sketches with facial details. The GAN-based methods (pix2pix, CycleGAN, PS2MAN, SCAGAN, and Sketch-Transformer) can synthesize sketches with certain sketch styles. However, some unacceptable textures are generated in the critical region (e.g., eye, mouth, and hair). Wild has strong robustness against environmental noise but tends to produce over-smooth results. The proposed PMSGAN can generate the most sketch-like texture while maintaining reasonable semantics.

Fig. 10 shows the qualitative comparison of PMSGAN with other state-of-the-art methods on the APDrawing database. pix2pix has a large proportion of artifacts and undesirable messy lines. CycleGAN fails to mimic the artistic portrait style well and cannot generate detailed textures of the mouth area. APDrawingGAN and APDrawingGAN++ can generate reasonable results that capture different texture details in different face regions and have delicate white lines in the hair. Pixel2style2pixel cannot learn the correct style. However, their results still have many artifacts. Synthetic images of U^2-Net can preserve the structure of the facial region and have smooth hair lines. However, unreasonable background artifacts and

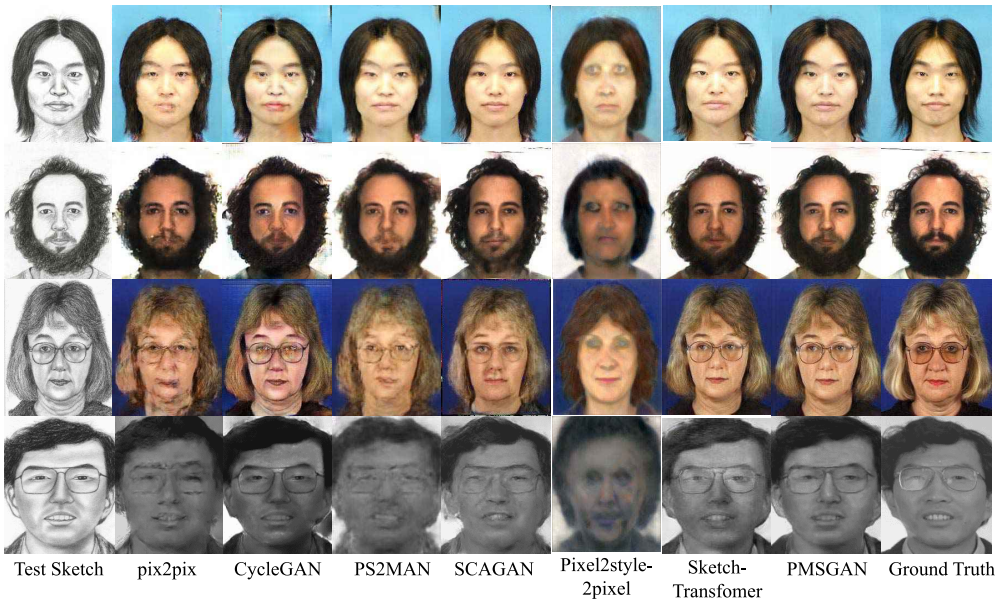

Fig. 8. Examples of synthetic face photographs on the CUFS dataset and the CUFSF dataset. From top to bottom: the examples are selected from the CUHK student database, the AR database, the XM2VTS database, and the CUFSF database, respectively.

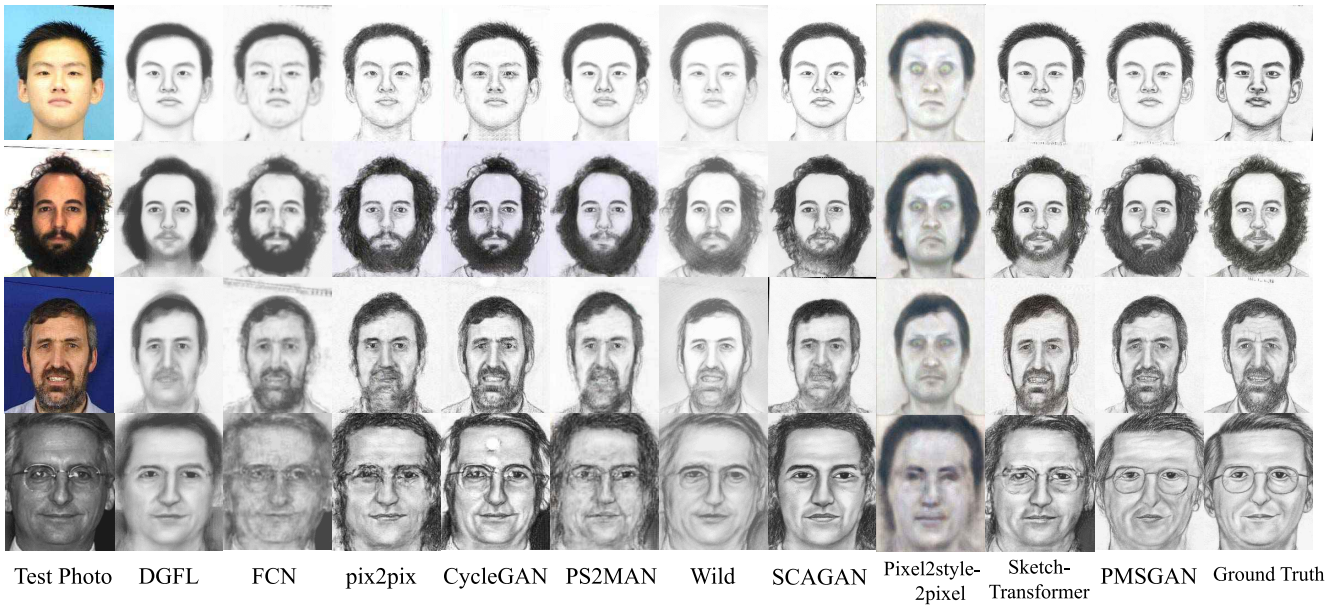

Fig. 9. Examples of synthetic face sketches on the CUFS dataset and the CUFSF dataset. From top to bottom: the examples are selected from the CUHK student database, the AR database, the XM2VTS database, and the CUFSF database, respectively.

some minor incoherent textures affect the overall appearance of the results. The proposed PMSGAN can well maintain the delicate structure of each facial region and learn the realistic artist portrait style.

Fig. 11 shows the qualitative comparison of PMSGAN with other state-of-the-art methods on the Sketch2Anime dataset that we collected. pix2pix has obvious artifacts and cannot fill the hand with reasonable colors. CycleGAN fails to preserve reasonable structures in the hair area and cannot obtain the right color distribution in the facial region. The results of UGATIT have deformation in the hair area and cannot distinguish the color of hair and clothes. DRIT++ is unable to generate reasonably colored anime images. The proposed PMSGAN protects the integrity of texture and structure and can generate anime images with impressive visual appearance. PMSGAN can easily color even subtle facial lines.

2) Quantitative Comparison: We introduce five metrics to quantitatively evaluate the quality of synthetic images. Among these metrics, LPIPS (alex), LPIPS (squeeze), and LPIPS (vgg) were verified to be more consistent with human perception in [65]. Therefore, we suggest paying more attention to the evaluation results of these three metrics. FSIM is a commonly used full-reference image quality assessment metric. In this

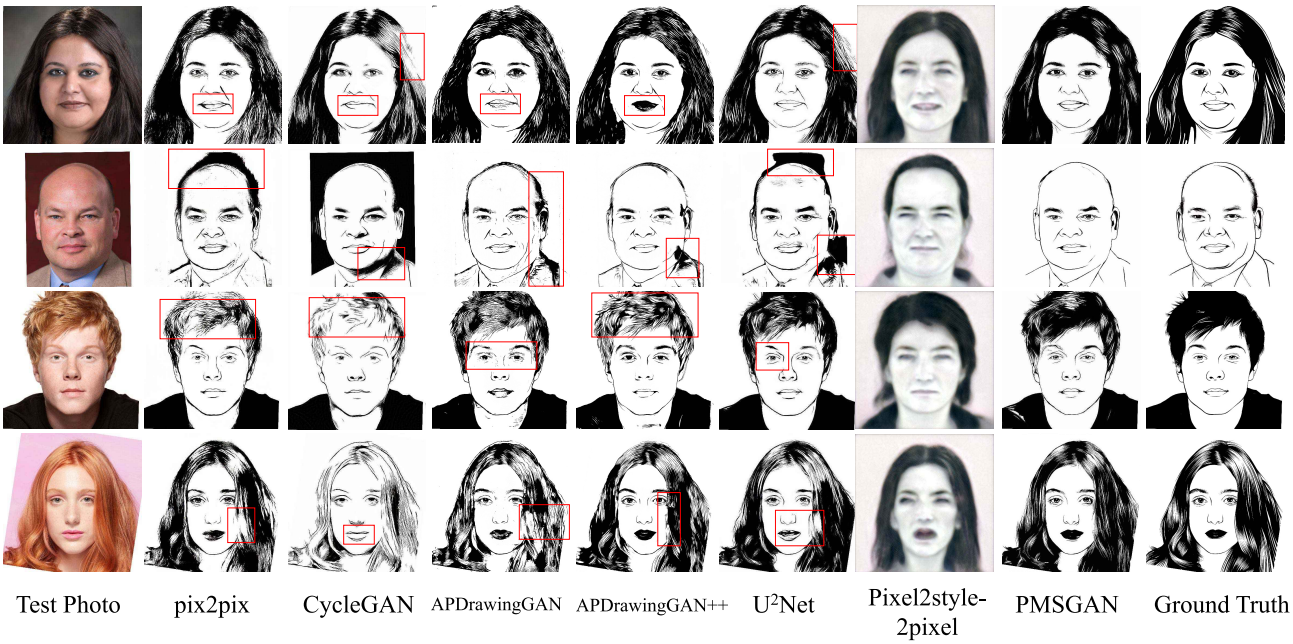

Fig. 10. Examples of synthetic artistic portrait drawings on the APDrawing dataset.

Fig. 11. Examples of synthetic anime images on the Sketch2Anime dataset.

work, we use it as one of the evaluation metrics. FID is designed to capture the Fréchet difference between two Gaussians (synthetic and real-world images). We also use it as one of the evaluation metrics.

The quantitative results of the comparison with state-of-the-art methods on face image translation of the CUFS database and the CUFSF database are shown in Table III. The analysis of Table III shows that all three variants of LPIPS show a preference for GAN-based methods. The reason is that GAN-based methods tend to produce more realistic synthetic images with high perceptual quality. Among these GAN-based methods, our PMSGAN model surpasses the other models in almost all LPIPS scores. The CUFSF database is more challenging, and the photographs inside have light changes. Compared with SCAGAN, our method does not use the prior information of face photographs, which is the reason why PMSGAN does not achieve the best performance in the photograph face translation task of the CUFSF database. For the evaluation metric of FSIM, although PMSGAN does not get the best score, PMSGAN still gets a score close to the best score.

The quantitative results of the comparison with state-of-the-art methods on face image translation of the APDrawing database are shown in Table IV. The analysis of Table IV shows that the PMSGAN model achieves the best score in all the evaluation metrics. The reason is that the PMSGAN

TABLE III

QUANTITATIVE RESULTS OF THE COMPARISON WITH STATE-OF-THE-ART METHODS ON SYNTHETIC FACE PHOTOGRAPHS/SKETCHES OF THE CUFS DATABASE AND CUFSF DATABASE

			DGFL	FCN	pix2pix	CycleGAN	PS2MAN	Wild	SCAGAN	Pixel2style2pixel	Sketch-Transformer	PMSGAN
CUFS	Photo	LPIPS (alex) ↓	-	-	0.1993	0.2096	0.2464	-	0.1727	0.3214	0.1538	**0.1507**
		LPIPS (squeeze) ↓	-	-	0.1830	0.2094	0.2158		0.1643	0.3076	0.1310	**0.1277**
		LPIPS (vgg) ↓	-	-	0.3525	0.3882	0.3254	-	0.3053	0.4136	0.2738	**0.2707**
		FSIM ↑	-	-	0.7726	0.7450	0.7819	-	**0.7937**	0.7427	0.7851	0.7816
		FID ↓	-	-	73.56	80.44	65.04	-	80.53	92.59	**27.88**	32.24
	Sketch	LPIPS (alex) ↓	0.3316	0.4517	0.2263	0.2139	0.2961	0.2807	0.2408	0.5092	0.1807	**0.1759**
		LPIPS (squeeze) ↓	0.2635	0.3596	0.1552	0.1529	0.2265	0.2210	0.1722	0.4159	0.1233	**0.1193**
		LPIPS (vgg) ↓	0.3654	0.4350	0.3734	0.3598	0.3707	0.3639	0.3627	0.4833	0.3019	**0.3002**
		FSIM ↑	0.7079	0.6936	**0.7363**	0.7219	0.7230	0.7114	0.7086	0.6469	0.7350	0.7340
		FID ↓	70.81	69.93	44.91	**23.76**	48.95	59.26	38.61	105.97	20.92	27.28
CUFSF	Photo	LPIPS (alex) ↓	-	-	0.2463	0.2557	0.3145	-	**0.1735**	0.5314	0.2199	0.2097
		LPIPS (squeeze) ↓	-	-	0.2005	0.2002	0.2853	-	**0.1469**	0.4518	0.1714	0.1642
		LPIPS (vgg) ↓	-	-	0.4019	0.3791	0.4237	-	**0.3128**	0.5430	0.3474	0.3338
		FSIM ↑	-	-	0.7777	0.7645	0.7812	-	**0.8395**	0.7529	0.7861	0.7858
		FID ↓	-	-	39.82	**14.46**	78.03	-	18.84	136.34	15.22	16.57
	Sketch	LPIPS (alex) ↓	0.3524	0.4793	0.2408	0.2371	0.3288	0.3288	0.2188	0.5716	0.1971	**0.1857**
		LPIPS (squeeze) ↓	0.2794	0.3895	0.1628	0.1589	0.2397	0.2473	0.1500	0.4793	0.1349	**0.1152**
		LPIPS (vgg) ↓	0.3972	0.5305	0.3824	0.3744	0.4170	0.4053	0.3536	0.5531	0.3400	**0.2844**
		FSIM ↑	0.6957	0.6624	0.7283	0.7088	0.7233	0.6821	0.7270	0.6509	0.7259	**0.7580**
		FID ↓	57.33	124.40	35.52	14.62	64.42	59.76	18.32	123.20	**9.39**	16.72

TABLE IV

QUANTITATIVE RESULTS OF THE COMPARISON WITH STATE-OF-THE-ART METHODS ON SYNTHETIC ARTISTIC PORTRAIT DRAWINGS OF THE APDRAWING DATABASE

		pix2pix	CycleGAN	APDrawing	APDrawing++	U^2Net	Pixel2style2pixel	PMSGAN
APDrawing	LPIPS (alex) ↓	0.3072	0.3249	0.2902	0.2547	0.2484	0.6333	**0.2373**
	LPIPS (squeeze) ↓	0.1826	0.1976	0.1686	0.1468	0.1411	0.5409	**0.1321**
	LPIPS (vgg) ↓	0.2758	0.2886	0.2645	0.2382	0.2284	0.5743	**0.2137**
	FSIM ↑	0.7329	0.7135	0.7395	0.7557	0.7666	0.6055	**0.7754**
	FID ↓	75.30	80.44	64.32	57.11	49.44	159.27	**46.71**

TABLE V

QUANTITATIVE RESULTS OF THE COMPARISON WITH STATE-OF-THE-ART METHODS ON FACE SKETCH COLORIZATION OF THE SKETCH2ANIME DATABASE

		pix2pix	CycleGAN	UGATIT	DRIT++	PMSGAN
Sketch2Anime	FSIM ↑	0.8358	0.7140	0.6810	0.6453	**0.8619**
	FID ↓	53.02	61.00	32.12	38.42	**18.69**

model can minimize the loss of information in the process of encoding and decoding.

The quantitative results of the comparison with state-of-the-art methods on face image translation of the Anime2sketch database are shown in Table V. The analysis of Table V shows that the PMSGAN model achieves the best score in the FID and the FSIM metric. The best performance in multiple databases further proves the advantages of our parallel multistage network and the effectiveness of our proposed feature fusion collaboration module.

Since image quality evaluation itself is a topic to be further studied, it is not sufficient to evaluate the model only by

Fig. 12. Left column: Examples of synthetic sketches on the APDrawing database. Middle column: Examples of synthetic portrait drawings on a mixed dataset. Right column: Examples of synthetic anime drawings on wild data.

TABLE VI
USER STUDY ON THE CUFS AND CUFSF DATABASES: PERCENTAGE OF VOTES OBTAINED BY EACH METHOD IN FIVE SATISFACTION LEVELS

Satisfaction Level	1	2	3	4	5
SCAGAN	5.21%	22.17%	40.00%	25.65%	6.96%
Pixel2style2pixel	97.39%	1.30%	0.87%	0.43%	0%
Sketch-Transformer	0.87%	14.35%	39.57%	33.04%	12.17%
PMSGAN	2.17%	10.87%	28.70%	**42.17%**	**16.09%**

TABLE VII
USER STUDY ON THE APDRAWING DATABASE: PERCENTAGE OF VOTES OBTAINED BY EACH METHOD IN FIVE SATISFACTION LEVELS

Satisfaction Level	1	2	3	4	5
APDrawingGAN	0%	49.00%	38.00%	12.00%	1.00%
APDrawingGAN++	2.00%	13.00%	13.00%	63.00%	9.00%
U^2Net	3.00%	20.00%	34.00%	32.00%	11.00%
Pixel2style2pixel	91.00%	4.00%	2.00%	2.00%	1.00%
PMSGAN	2.00%	2.00%	23.00%	38.00%	**35.00%**

the above evaluation metrics. Hence, we conduct three user studies to compare our models (PMSGAN) with state-of-the-art methods on the CUFS database, the CUFSF database, the APDrawing database, and the Sketch2Anime database. For the method of the user study, we adopt mean opinion score (MOS) testing, where participants are asked to assign perceptual quality scores to tested images. Typically, the scores are from 1 (very unsatisfactory) to 5 (very satisfactory) and the final MOS is calculated as the arithmetic mean overall ratings. For the CUFS database and the CUFSF database, there are ten and four synthetic images of the four methods listed in Table VI. For the APDrawing database, there are five groups of satisfaction rating requirements in the study. Each group includes one real image and five synthetic images of the five methods listed in Table VII. For the Sketch2Anime database, there are four groups of satisfaction rating requirements in the study. Each group includes one real image, one sketch image, and five synthetic images of the five methods listed in Table VIII. The real image is given in the title, and the synthetic images are given in the answer area. The placement order of the synthetic images is shuffled. According to the satisfaction with the synthetic images with respect to the real image, the respondents rate each synthetic image with a score between 1 and 5. For each user study, we collect a total of 20 returned questionnaires and calculate the percentage of votes obtained by each method in the five satisfaction levels, as shown in Tables VI–VIII.

TABLE VIII
USER STUDY ON THE SKETCH2ANIME DATABASE: PERCENTAGE OF VOTES OBTAINED BY EACH METHOD IN FIVE SATISFACTION LEVELS

Satisfaction Level	1	2	3	4	5
pix2pix	1.20%	1.20%	15.48%	48.81%	11.96%
CycleGAN	10.71%	45.24%	21.43%	17.86%	4.76%
UGATIT	10.71%	30.95%	35.71%	17.86%	4.76%
DRIT++	26.19%	30.95%	25.00%	14.29%	0.00%
PMSGAN	0.00%	3.57%	3.57%	32.14%	**60.71%**

TABLE IX
COMPARISON OF THE RUNTIME OF THE STATE-OF-THE-ART METHODS ON SYNTHETIC ARTISTIC PORTRAIT DRAWINGS OF THE APDRAWING DATABASE

		pix2pix	CycleGAN	APDrawing	APDrawing++	U^2Net	Pixel2style2pixel	PMSGAN
APDrawing	runtime(s) ↓	**0.050**	0.152	0.083	0.098	0.075	0.109	<u>0.058</u>

Table VI shows that our PMSGAN model gets the most votes (16.09%) on the "very satisfactory" level and a total of 86.96% of the votes above the "average" judgment. This result indicates that our PMSGAN model can produce quite satisfactory synthetic images. Sketch-Transformer gets a total of (84.78%) of the votes above the "average" judgment, second only to the PMSGAN model. Pixel2style2pixel gets the most votes on the "very unsatisfactory" level, which indicates that its results have serious defects, which we think are unreasonable texture, artifacts, and noise. SCAGAN gets the most votes (40.00%) on the "average" level, but also many "very satisfactory" and "unsatisfactory" votes, which indicates that the model produces inconsistent and relatively poor results.

Table VII shows that our PMSGAN model gets the most votes (35.00%) on the "very satisfactory" level and a total of 96.00% of the votes above the "average" judgment. This result indicates that our PMSGAN model can produce very satisfactory synthetic images. Pixel2style2pixel gets the most votes at 91.00% on the "very unsatisfactory" level, which indicates that its results have poor quality. Both U^2Net and APDrawingGAN get the most votes on the "unsatisfactory" level, which indicates that their results have some defects, which we think are unreasonable texture, artifacts, and noise. APDrawingGAN++ gets a total of (85.00%) of the votes above the "average" judgment, second only to the PMSGAN model.

Table VIII shows that our PMSGAN model gets the most votes (60.71%) on the "very satisfactory" level and a total of 96.42% of the votes above the "average" judgment. This result indicates that our PMSGAN model can produce very satisfactory synthetic images. pix2pix gets the most votes 48.81% on the "satisfactory" level, which indicates that its results have good quality. DRIT++, UGATIT, and CycleGAN get the most votes on the "unsatisfactory" level, which indicates that their results have some defects, which we think are unreasonable texture, artifacts, and noise.

3) Runtime Analysis: We compare the runtime of the proposed method with the state-of-the-art methods in the APDrawing dataset and further analyze the efficiency advantages of the proposed model. A comparison of the runtime of the state-of-the-art methods on synthetic artistic portrait drawings of the APDrawing database is shown in Table IX. The analysis of Table IX shows that the runtime of PMSGAN is second only to pix2pix and very close to pix2pix. Benefiting from the parallel multistage strategy, the time spent in the inference process of PMSGAN is not affected by the introduction of the multistage strategy.

4) Robustness Analysis: We test the robust performance of the model on sketch and portrait synthesis tasks by using wild data. Specifically, we use the PSMGAN model trained on the CUFS dataset to test the test data in the APDrawing dataset, use the PMSGAN model trained on the APDrawing dataset to test the test set of the mixed dataset composed of the CUHK dataset and the CelebA dataset, and finally use the PMSGAN model trained on the Sketch2Anime dataset to test the wild data. Fig. 12 indicates that PMSGAN has good robustness, which further verifies the effectiveness of PMSGAN.

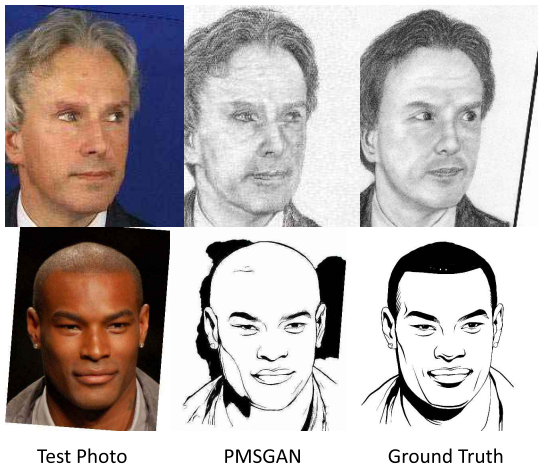

Fig. 13. Examples of failed results.

5) Analysis of Failure Cases: Although the PMSGAN model proposed in this article has strong learning ability and generalization ability and can also achieve excellent results in wild data, in some extreme cases (such as the side face, the ambient brightness is too bright or too dark), PMSGAN will also produce some failure cases. Fig. 13 shows examples of failed results. Failure cases are usually caused by insufficient data, which makes the model unable to learn the cross-domain translation of face images in extreme cases.

V. CONCLUSION

In this article, we revisit the coarse-to-fine strategy and propose a parallel multistage model for face image translation tasks. The proposed PMSGAN is driven by three insights: parallel multistage strategy, CSASP, and multiscale supervised attention module. Together, they promote highly efficient information exchange and effective function learning. Qualitative and quantitative results demonstrate that the proposed method achieves significant improvements in retaining structural information and generating detailed textures. We will explore a more effective information fusion mechanism in future work.

REFERENCES

[1] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "A comprehensive survey to face hallucination," *Int. J. Comput. Vis.*, vol. 106, no. 1, pp. 9–30, 2014.

[2] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[3] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[4] 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 Recognit.*, vol. 106, Oct. 2020, Art. no. 107404.

[5] I. J. Goodfellow et al., "Generative adversarial nets," in *Proc. NeurIPS*, 2014, pp. 2672–2680.

[6] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[7] T.-C. Wang et al., "Video-to-video synthesis," 2018, *arXiv:1808.06601*.

[8] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1857–1865.

[9] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2849–2857.

[10] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 172–189.

[11] H.-Y. Lee et al., "DRIT++: Diverse image-to-image translation via disentangled representations," *Int. J. Comput. Vis.*, vol. 128, no. 10, pp. 2402–2417, 2020.

[12] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[13] Q. Li et al., "Coupled GAN with relativistic discriminators for infrared and visible images fusion," *IEEE Sensors J.*, vol. 21, no. 6, pp. 7458–7467, Mar. 2021.

[14] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 700–708.

[15] J. Kim, M. Kim, H. Kang, and K. Lee, "U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," 2019, *arXiv:1907.10830*.

[16] M. Zhu, J. Li, N. Wang, and X. Gao, "A deep collaborative framework for face photo–sketch synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 3096–3108, Oct. 2019.

[17] C. Peng, X. Gao, N. Wang, D. Tao, X. Li, and J. Li, "Multiple representations-based face sketch–photo synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2201–2215, Nov. 2016.

[18] M. Zhang, N. Wang, Y. Li, and X. Gao, "Neural probabilistic graphical model for face sketch synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 7, pp. 2623–2637, Jul. 2020.

[19] M. Zhu, J. Li, N. Wang, and X. Gao, "Knowledge distillation for face photo–sketch synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 3, no. 2, pp. 893–906, Feb. 2022.

[20] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma, "A nonlinear approach for face sketch synthesis and recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 1005–1010.

[21] L. Chang, M. Zhou, Y. Han, and X. Deng, "Face sketch synthesis via sparse representation," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2146–2149.

[22] M. Zhu, N. Wang, X. Gao, and J. Li, "Deep graphical feature learning for face sketch synthesis," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 3574–3580.

[23] L. Zhang, L. Lin, X. Wu, S. Ding, and L. Zhang, "End-to-end photo-sketch generation via fully convolutional representation learning," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, Jun. 2015, pp. 627–634.

[24] L. Wang, V. Sindagi, and V. Patel, "High-quality facial photo-sketch synthesis using multi-adversarial networks," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 83–90.

[25] J. Yu et al., "Toward realistic face photo–sketch synthesis via composition-aided GANs," *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4350–4362, Sep. 2021.

[26] C. Chen, W. Liu, X. Tan, and K.-Y. K. Wong, "Semi-supervised learning for face sketch synthesis in the wild," in *Proc. ACCV*, 2018, pp. 216–231.

[27] M. Zhu, C. Liang, N. Wang, X. Wang, Z. Li, and X. Gao, "A sketch-transformer network for face photo-sketch synthesis," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1352–1358.

[28] R. Yi, Y.-J. Liu, Y.-K. Lai, and P. L. Rosin, "APDrawingGAN: Generating artistic portrait drawings from face photos with hierarchical GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10743–10752.

[29] R. Yi, M. Xia, Y.-J. Liu, Y.-K. Lai, and P. L. Rosin, "Line drawings for face portraits from photos using global and local structure based GANs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3462–3475, Oct. 2021.

[30] B. Li, Y. Zhu, Y. Wang, C.-W. Lin, B. Ghanem, and L. Shen, "AniGAN: Style-guided generative adversarial networks for unsupervised anime face generation," 2021, *arXiv:2102.12593*.

[31] E. Richardson et al., "Encoding in style: A styleGAN encoder for image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 2287–2296.

[32] M. Li, H. Huang, L. Ma, W. Liu, T. Zhang, and Y. Jiang, "Unsupervised image-to-image translation with stacked cycle-consistent adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 184–199.

[33] J. Liu, C. Wang, H. Su, B. Du, and D. Tao, "Multistage GAN for fabric defect detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3388–3400, 2019.

[34] D. Peng, W. Yang, C. Liu, and S. Lü, "SAM-GAN: Self-attention supporting multi-stage generative adversarial networks for text-to-image synthesis," *Neural Netw.*, vol. 138, pp. 57–67, Jun. 2021.

[35] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4641–4650.

[36] M. Liu, W. Chen, C. Wang, and H. Peng, "A multiscale ray-shooting model for termination detection of tree-like structures in biomedical images," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1923–1934, Aug. 2019.

[37] T. R. Shaham, T. Dekel, and T. Michaeli, "SinGAN: Learning a generative model from a single natural image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4570–4580.

[38] D. Park, D. U. Kang, J. Kim, and S. Y. Chun, "Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 327–343.

[39] S. W. Zamir et al., "Multi-stage progressive image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 14821–14831.

[40] R. A. Rensink, "The dynamic representation of scenes," *Vis. Cognition*, vol. 7, nos. 1–3, pp. 17–42, Oct. 2000.

[41] H. Du, J. Wang, M. Liu, Y. Wang, and E. Meijering, "SwinPA-Net: Swin transformer-based multiscale feature pyramid aggregation network for medical image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–12, 2022.

[42] F. Wang et al., "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3156–3164.

[43] L. Wang, L. Zhang, X. Qi, and Z. Yi, "Deep attention-based imbalanced image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 3320–3330, Aug. 2022.

[44] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3640–3649.

[45] J. Fu, J. Liu, J. Jiang, Y. Li, Y. Bao, and H. Lu, "Scene segmentation with dual relation-aware attention network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2547–2560, Jun. 2021.

[46] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[47] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[48] L. Yao et al., "Describing videos by exploiting temporal structure," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4507–4515.

[49] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 21–29.

[50] J. Kuen, Z. Wang, and G. Wang, "Recurrent attentional networks for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3668–3677.
[51] J. Yang, A. Kannan, D. Batra, and D. Parikh, "LR-GAN: Layered recursive generative adversarial networks for image generation," 2017, *arXiv:1703.01560*.
[52] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.
[53] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*, Jul. 2017, pp. 4700–4708.
[54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[55] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1955–1967, Nov. 2009.
[56] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proc. CVPR*, Jun. 2011, pp. 513–520.
[57] X. Tang and X. Wang, "Face sketch synthesis and recognition," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 687–694.
[58] A. M. Martinez and R. Benavente, "The ar face database," CVC, New Delhi, India, Tech. Rep. #24, 1998.
[59] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. AVBPA*, 1999, pp. 72–77.
[60] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.
[61] G. B. Anonymous, T. D. Community, and A. Gokaslan. *Danbooru2018: A Large-Scale Crowdsourced and Tagged Anime Illustration Dataset*. Accessed: Jan. 2019. [Online]. Available: https://www.gwern.net/Danbooru2018
[62] X. Xiang, D. Liu, X. Yang, Y. Zhu, and X. Shen. (2021). *Anime2sketch: A Sketch Extractor for Anime Arts With Deep Networks*. [Online]. Available: https://github.com/Mukosame/Anime2Sketch
[63] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," 2016, *arXiv:1606.03498*.
[64] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[65] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[66] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," 2016, *arXiv:1602.07360*.
[67] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.
[68] R. Zhang. *Perceptualsimilarity*. Accessed: May 1, 2022. [Online]. Available: https://github.com/richzhang/PerceptualSimilarity
[69] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Conf. Workshop Neural Inf. Process. Syst.*, 2017, pp. 1–12.
[70] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

Changcheng Liang received the B.Eng. degree in electronic science and technology from Xidian University, Xi'an, China, in 2020, where he is currently pursuing the M.Sc. degree.

His current research interests include computer vision and machine learning.

Mingrui Zhu received the B.Eng. degree in electronic and information engineering from Guangxi University, Nanning, China, in 2014, and the Ph.D. degree in circuits and systems from Xidian University, Xi'an, China, in 2020.

He is currently a Lecturer with the State Key Laboratory of Integrated Services Networks, Xidian University. His current research interest includes computer vision and machine learning.

Nannan Wang (Member, IEEE) received the B.Sc. degree in information and computation science from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2009, and the Ph.D. degree in information and telecommunications engineering from Xidian University, Xi'an, in 2015.

He is currently a Professor with the State Key Laboratory of Integrated Services Networks, Xidian University. He has published over 150 articles in refereed journals and proceedings, including IEEE T-PAMI, IJCV, CVPR, ICCV, and so on. His current research interests include computer vision and machine learning.

Heng Yang received the B.Sc. degree in simulation engineering and the M.Sc. degree in pattern recognition and intelligent system from the National University of Defense Technology (NUDT), Changsha, China, in 2009 and 2011, respectively, and the Ph.D. degree from the Department of Electrical Engineering and Computer Science, Queen Mary University of London, London, U.K., in 2015.

He worked as a Research Associate with the Computer Laboratory, University of Cambridge, Cambridge, U.K., in 2015. He is currently a Senior Leader with Shenzhen AiMall Tech, China. His research interests include computer vision, pattern recognition, and applied machine learning.

Xinbo Gao (Senior Member, IEEE) received the B.Eng., M.Sc., and Ph.D. degrees in electronic engineering, signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively.

From 1997 to 1998, he was a Research Fellow with the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a Post-Doctoral Research Fellow with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong. Since 2001, he has been with the School of Electronic Engineering, Xidian University. He is also a Cheung Kong Professor of the Ministry of Education of China, a Professor of pattern recognition and intelligent system with Xidian University, and a Professor of computer science and technology with the Chongqing University of Posts and Telecommunications. He has published six books and around 300 technical articles in refereed journals and proceedings. His current research interests include image processing, computer vision, multimedia analysis, machine learning, and pattern recognition. He is a fellow of the Institute of Engineering and Technology and the Chinese Institute of Electronics.

Dr. Gao served as the General Chair/Co-Chair, the Program Committee Chair/Co-Chair, or a PC Member for around 30 major international conferences. He is on the Editorial Boards of several journals, including *Signal Processing* (Elsevier) and *Neurocomputing* (Elsevier).