Facial Image Regeneration

Haolin Ye, Haipeng Liu, Jiayu Wang, Peixu Xin, Runchuan Feng
{haolinye, haipengl, jiayuw, brucexin, fengrc}@bu.edu

1 Task

Facial Image Manipulation has become a highly intriguing and rapidly evolving area of research in computer vision and image processing. With the advent of advanced deep learning models and image synthesis algorithms, researchers have been able to develop highly sophisticated methods for generating natural-looking facial images by manipulating various facial features such as expressions, poses, and identities.

The primary objective of this task is to manipulate facial features while preserving the spatial relationships between different facial components to generate new and realistic images. However, this is a highly challenging task due to the intricate interdependencies between different facial features. For instance, changing the expression of a face without altering its overall appearance is a daunting task that requires sophisticated techniques to accurately reproduce the natural variations in facial expressions and poses.

The ability to generate highly realistic facial images holds tremendous potential in several applications, including entertainment, virtual reality, and forensics. For instance, in the entertainment industry, facial image manipulation can be used to create highly realistic CGI characters and special effects, while in virtual reality, it can be used to create highly immersive environments. Similarly, in forensics, facial image manipulation can be used to generate age progression images of missing persons or create accurate facial composites of suspects in criminal investigations.

Overall, the development of advanced facial image manipulation techniques has opened up several new avenues for research and has the potential to revolutionize several industries.

2 Related Work

2.1 MSG-GAN: Multi-Scale Gradients for Generative Adversarial Networks [1]

In recent years, Generative Adversarial Networks (GANs) have become increasingly popular for their ability to generate high-quality images. However, one of the primary challenges in training GANs is to maintain stability during the training process, particularly for high-resolution images. In this work, the researchers propose a simple yet effective approach called Multi-Scale Gradient Generation Countermeasure Network (MSG-GAN) to address this challenge. The MSG-GAN provides gradient from discriminator to generator on multiple scales, facilitating stable training for high-resolution image synthesis.

Limitations: Despite its effectiveness, GANs are still challenging to adapt to different datasets due to their instability and sensitivity to superparameters during training. The selection of appropriate hyperparameters remains an active area of research in GANs.

2.2 StyleRig: Rigging StyleGAN for 3D Control Over Portrait Images[2]

StyleGAN is a state-of-the-art GAN architecture that can generate high-quality portraits with a high degree of variability in terms of facial features such as expression, pose, and identity. However, controlling the facial semantic parameters of StyleGAN can be challenging. In this paper, the researchers propose a method called StyleRig that allows users to manipulate the semantic parameters of pre-trained and fixed StyleGAN through 3D Morphable Model (3DMM), without the need for manual labeling. The approach is based on self-supervised training, where the model learns to control the parameters by minimizing a set of reconstruction and regularization losses.

Limitations: Although 3DMM provides seman-

tic parameters to control, it lacks realism in rendering and only models the face, not other parts of the portrait, such as hair, mouth interior, and background. This limits the degree of control users have over the overall appearance of the portrait.

2.3 Deep 3D Portrait From a Single Image[3]

Reconstructing a 3D portrait from a single 2D image is a challenging task that has received significant attention in recent years. In this paper, the researchers propose a two-stage approach to address this problem. The first stage involves reconstructing the face shape using a 3D Morphable Model (3DMM) by minimizing the difference between the rendered face image and the input image. The second stage involves learning the geometry of other parts of the portrait, such as hair and ears, using a depth map and stereo visual matching, in an unsupervised manner. The entire process is trained without any supervision, making it highly scalable.

Limitations: Although the method is unsupervised and does not require any real 3D data, it still has limitations in terms of accurately representing the geometry of non-face parts and handling complex lighting conditions. Additionally, the accuracy of the depth map estimation can be affected by factors such as image resolution and noise.

3 Approach

3.1 Overall Structure

Our goal is to implement a conditional generative adversarial network (cGAN) [4] which can regenerate high-resolution facial images with modified features. The overall structure of our network is shown in Fig. 1, consisting of two separate part. The The top half part is a spatial-aware style network [5] which extracts style information from the reference image and corresponding labeled mask. The bottom half part is a backbone of Pix2PixHD model [6] which generates images from the input labeled mask (modified). Our model can generate an image based on the modified labeled mask as well as reserving the basic feature information of the reference image. Most of the GAN part is based on the implementation of Pix2PixHD model, which has proved to be powerful in synthesizing high-resolution photo-realistic images from semantic label maps.

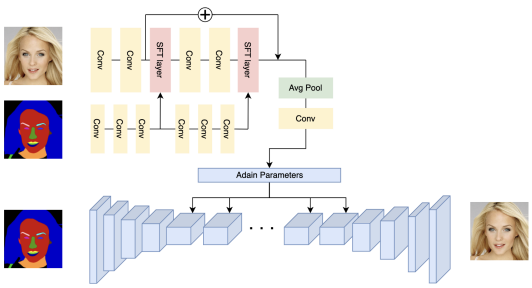

Figure 1: Overall Structure of the purposed model

3.2 Spatial-Aware Style Network

Apart form just regenerating a random facial image according to a given semantic labeled masks, our project requires more in keeping the origin feature information of the reference images such as the complexion, hair color or style, etc. Therefore, feature information needed to be extracted in advance and send to the generator as a kind of key parameter. To meet such requirement, we purposed a spatial-aware style network which takes in a reference image and corresponding semantic labeled masks and outputs an array of parameter whose size is determined by the number of residual blocks in the generator. To maintain as more feature information from the reference image as possible, we utilized the spatial feature transformer layer and adaptive normalization layer as well as setting the second-convolved of the reference image as an parameter directly added to the output of the spatial feature transformer layer.

3.2.1 Spatial Feature Transform Layer

Spatial feature transformation layer (SFT layer) [7] is a layer used for deep neural networks to realize spatial transformation of feature mapping. It learns to apply affine transformations to feature maps based on learned parameters such as scaling, rotation, and shift. The SFT layer is usually used for image processing tasks such as image recognition and segmentation to enhance the robustness of the model to the spatial transformation of input data. The SFT layer can be inserted into various parts of the neural network architecture, such as the encoder or decoder of the convolutional neural network, to allow spatial transformation of feature maps at different stages of the network. By applying spatial transformations, the different domains of style information obtained from reference image and that of its corresponding spatial information can be fused together.

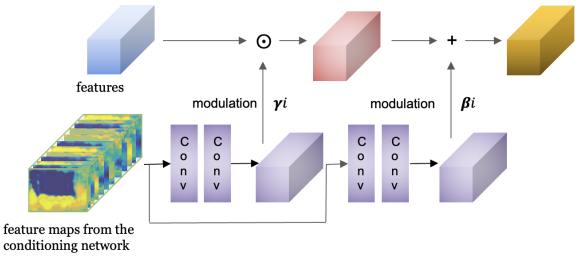

Figure 2: Structure of the Spatial Feature Transform Layer

Through spatial affine transformation of each intermediate feature map in the super resolution network, the learned parameters will influence the output adaptively. During the test, a high resolution image can be generated with only one forward pass given a low resolution input and a segmentation probability graph. As shown in Fig. 2, the prior condition Ψ is modeled by a pair of affine transformation parameters (γ, β) through a mapping function, thus the process can be concluded in the following two formulas:

$$(\gamma, \beta) = \mathcal{M}(\Psi)$$

$$SFT(\mathbf{F} \mid \gamma, \beta) = \gamma \odot \mathbf{F} + \beta$$

Here, a convolution neural network is applied for the mapping function $\mathcal{M}$ which can be optimized end-to-end during the training process.

3.2.2 Adaptive Instance Normalization Layer

Adaptive instance normalization (AdaIN) [8] is a normalization method to align the mean and variance of content features with that of style features. Instance normalization normalizes the input to a single style specified by the affine parameter. Adaptive instance normalization is an extension. In AdaIN, a content input and a style input is received, then the mean and variance of the channels are aligned to match each other. Unlike batch normalization, instance normalization, or conditional instance normalization, AdaIN has no learnable affine parameters. Instead, it adaptively computes affine parameters from style input as:

$$AdaIN(x, y) = \sigma(y) \left(\frac{x - \mu(x)}{\sigma(x)} \right) + \mu(y)$$

During inference, given the content image and style image, the mean and variance of each feature channel in the content image are adjusted according to the learning parameters of the style image, and the style of the style image is effectively applied to the content image. This allows the creation of highly stylized images, such as transferring the style of a painting to a photograph. Overall, AdaIN is a powerful technique for style conversion, image processing, and other tasks that require transferring style information between images.

3.3 Image Generation Backbone

The image generation backbone of our model is based on the pix2pixHD architecture. [6]

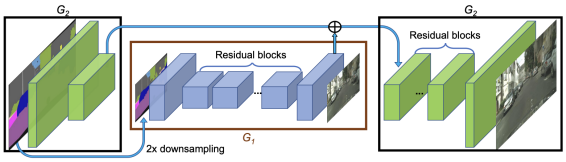

Figure 3: pix2pixHD $G_1 : GlobalGeneratornetwork$ $G_2 : LocalEnhancernetwork$

Pix2pixHD is a high-resolution image-to-image translation model based on the original pix2pix framework, introduced by NVIDIA researchers in 2017. The main purpose of the pix2pixHD model is to generate high-quality and visually coherent images, given a corresponding input image. It is particularly effective for translating semantic label maps into photorealistic images, which makes it suitable for various applications such as facial image manipulation, style transfer, and image synthesis.

The pix2pixHD model incorporates several key components and improvements over the original pix2pix framework:

Generator: The generator in pix2pixHD is based on a deep convolutional architecture that leverages residual blocks and skip connections to capture both high-level semantic information and low-level image details. It also uses a global generator with a coarse-to-fine structure, combined with a local enhancer network to refine the generated images.

Discriminator: Pix2pixHD employs a multi-scale discriminator, which is composed of multiple discriminators operating at different scales. This allows the model to capture both global and local features, improving the overall realism of the generated images.

These components and improvements make pix2pixHD a powerful and versatile tool for image-to-image translation tasks, particularly when high-resolution and visually coherent outputs are desired, such as in facial image manipulation applications.

This architecture uses a U-Net generator and a multi-Scale discriminator to generate high-resolution images from semantic label maps. The generator consists of an encoder, a bottleneck layer, and a decoder. The encoder reduces the spatial resolution of the input semantic label map and extracts feature maps. The bottleneck layer processes the feature maps to produce an array of parameters that are used to condition the decoder. The decoder then uses the parameters and the feature maps to generate a high-resolution image.

3.3.1 U-Net Generator

The U-Net generator [9] is a popular architecture for image-to-image translation tasks, and it consists of an encoder, a bottleneck layer, and a decoder. In our model, we modified the U-Net architecture [10] without cross-layer connections in the decoder, which helps to reduce the number of parameters and make the model more efficient.

The encoder uses convolutional layers to extract feature maps from the input semantic label map, while the decoder uses transpose convolutional layers to generate the output image. The bottleneck layer processes the feature maps to produce an array of parameters that are used to condition the decoder.

The U-Net architecture is effective for preserving the spatial structure of the input semantic label map and generating high-resolution images with detailed textures. The modified U-Net architecture without cross-layer connections is more efficient and suitable for high-resolution image generation tasks.

Overall, our model combines the spatial-aware style network and the modified U-Net architecture to generate high-resolution facial images with modified features while preserving the basic feature information of the reference image. The spatial-aware style network extracts the style information from the reference image and semantic label map and produces an array of parameters that condition the generator to generate images with the desired features. The modified U-Net architecture generates high-quality images from the semantic label maps conditioned by the style parameters using adversarial loss, perceptual loss, and feature matching loss.

3.4 Multi-scale Discriminator

A multi-scale discriminator [11] is an architecture used in Generative Adversarial Networks (GANs) to improve the quality of the generated images. It consists of multiple discriminators operating at different scales, allowing the model to capture both global and local features in the generated images, resulting in higher visual quality and realism.

In a GAN, the generator creates images while the discriminator evaluates the generated images' quality and authenticity. The generator's goal is to produce images that the discriminator cannot distinguish from real images. In the case of the multi-scale discriminator, several discriminators are employed, each focusing on a specific scale or resolution of the generated images. This enables the model to assess and capture a wide range of features and details, from coarse structures to fine textures.

The idea behind using multiple discriminators at different scales is to ensure that the generator learns to create images that look realistic at various levels of detail. This is achieved by training the generator with feedback from each of the discriminators. As the generator improves, it becomes better at generating images that can fool all the discriminators.

Multi-scale discriminators are especially beneficial in high-resolution image synthesis tasks, such as those in the pix2pixHD model, because they address some of the challenges associated with generating visually coherent and high-quality images. In particular, multi-scale discriminators help tackle the following issues:

1. Global structure and local details: High-resolution image synthesis demands the preservation of both global structure and local details. By using multiple discriminators operating at different scales, the model can focus on both coarse and fine features, thus maintaining the balance between global and local information. This results in more visually coherent images.

2. Gradient vanishing and instability: In high-resolution image generation, the generator must generate images across a wide range of scales, making training more challenging. The use of multiple discriminators provides diverse and rich feedback, which helps stabilize the training process and avoid gradient vanishing problems.

3. Computational efficiency: Training a single discriminator to capture all levels of detail in

high-resolution images can be computationally expensive. Multi-scale discriminators distribute the workload across several smaller discriminators, making the overall process more efficient.

4. Improved image quality: The multi-scale approach encourages the generator to produce images that look realistic at multiple levels of detail, improving the overall image quality. This is particularly important for applications like pix2pixHD, where the goal is to generate photorealistic images with high fidelity.

By addressing these challenges, multi-scale discriminators enhance the performance of high-resolution image synthesis models like pix2pixHD. They enable the generation of images with better visual quality, coherence, and realism, making them well-suited for the applications of our project.

3.5 Objective Function

Our objective function consists of three components: the adversarial loss, the perceptual loss, and the discriminator gradient penalty loss. We use the Adam optimizer to update the parameters of the generator and discriminator networks.

The adversarial loss measures the difference between the generated images and the target images in terms of realism. We use the GANLoss module, which implements either the mean squared error (MSE) loss or the binary cross-entropy (BCE) loss.

The perceptual loss [12] measures the difference between the generated images and the target images in terms of high-level content features. We use the VGGLoss module, which computes the L1 loss between the feature maps of the generated images and the target images at multiple scales using a pre-trained VGG19 network.

The discriminator gradient penalty loss [13] helps to enforce the Lipschitz constraint on the discriminator by penalizing the norm of the gradient of the discriminator's output with respect to the interpolated images between the real and fake images. We use the DiscriminatorGradientPenaltyLoss module, which computes the gradient penalty for a batch of real and fake images using the Wasserstein GAN with gradient penalty (WGAN-GP) algorithm.

$$L = L_{adv} + \lambda_{vgg} * L_{vgg} + \lambda_{gp} * L_{gp}$$

where L_{adv} is the adversarial loss, L_{vgg} is the perceptual loss, L_{gp} is the discriminator gradient penalty loss, and λ_{vgg} and λ_{gp} are the weighting factors for the perceptual loss and the gradient penalty loss, respectively.We update the generator and discriminator alternatively. For the generator, we compute the adversarial loss and the perceptual loss, and update the generator parameters using the sum of these two losses. For the discriminator, we compute the adversarial loss and the gradient penalty loss, and update the discriminator parameters using the sum of these two losses.

Overall, our model combines the spatial-aware style network and the pix2pixHD architecture to generate high-resolution facial images with modified features while preserving the basic feature information of the reference image. The spatial-aware style network extracts the style information from the reference image and semantic label map and produces an array of parameters that condition the generator to generate images with the desired features. The pix2pixHD architecture generates high-quality images from the semantic label maps conditioned by the style parameters using adversarial loss, perceptual loss, and feature matching loss. The multi-scale discriminator provides detailed feedback to the generator to improve the quality of the generated images.

4 Dataset

We use the CelebAMask-HQ dataset [14] provided by the Chinese University of Hong Kong to conduct our experiment. CelebAMask-HQ is a large-scale face image dataset, which selects 30,000 high-resolution face images from the CelebA dataset, captured in diverse poses and expressions, with a high resolution of 1024×1024 pixels.In addition to face images, each image includes pixel-level annotations that identify semantic masks for different facial attributes, such as hair, eyes, nose and mouth.

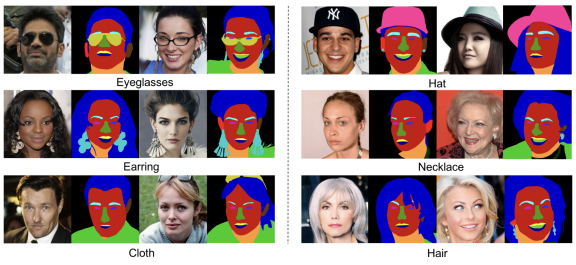

Figure 4: Sample Images of CelebAMask-HQ dataset

4.1 Data Preprocessing

Since CelebAMask-HQ is precisely marked by U-Net in size 512 x 512 and comes in 19 categories including all facial ingredients and accessories, such as "skin", "nose", "eyes", "eyebrows", "ears", "mouth", "lip", "hair" and "hat", "glasses", "earrings", "necklace" and the "neck" and "material", we first re-scale those labeled masks back to 1024×1024 to unify the size of reference images and masks. During the experiment, in order to save training time and server memory, we only choose 400 images to operate the training, 50 for validation and 50 for test.

5 Evaluation Metrics

Attribute Classification Accuracy, Segmentation Accuracy, and Fréchet Inception Distance (FID) Score are metrics commonly used to evaluate the performance of various image generation and manipulation models. In our project, we used these three metrics to evaluate over model;

5.1 Attribute Classification Accuracy [15]

This metric measures the performance of a model in predicting the correct features or attributes present in an image. It is often used for models that classify or recognize facial attributes, such as emotions, age, or gender. A higher attribute classification accuracy indicates that the model can accurately predict the features present in the image. Compare the predicted labels to the ground truth labels for the images in the testing set. For each correctly predicted attribute, mark it as a true positive (TP) or true negative (TN), depending on whether the attribute is present or absent in the image. For each incorrectly predicted attribute, mark it as a false positive (FP) if the model predicted the attribute to be present when it was not, or a false negative (FN) if the model predicted the attribute to be absent when it was actually present. And then, Compute the attribute classification accuracy using the following formula:

$$Accuracy = (TP+TN)/(TP+TN+FP+FN)$$

5.2 Segmentation Accuracy

This metric evaluates the performance of a model in identifying the boundaries of different attributes or objects within an image. It is particularly relevant for image segmentation tasks, where the goal is to separate and identify different regions of an image. A higher segmentation accuracy indicates that the model can accurately delineate the boundaries of different attributes or objects within the image. Segmentation accuracy is the ratio of Number of Correct Pixels and Total Number of Pixels.

5.3 Fréchet Inception Distance (FID) Score [16]

Fréchet Inception Distance (FID) Score is a metric used to evaluate the quality and diversity of generated images compared to a set of real images. It measures the similarity between the two sets of images by comparing their feature representations obtained from an Inception network (a deep learning model for image classification). Lower FID scores indicate that the generated images are more similar to the real images, suggesting better image quality and realism. For both the real and generated image sets, pass the images through the pre-trained Inception network to obtain their feature representations. Typically, the activations from the penultimate layer are used for this purpose.

After that, for each set of features (real and generated), calculate the mean and covariance. You will have a mean vector and a covariance matrix for each set.

Finally, we can compute the Fréchet distance between the two Gaussian distributions (one for real images and one for generated images) using the following formula:

$$FID = ||\mu_1 - \mu_2||_2^2 + \mathcal{T}_r(\sum\nolimits_1 + \sum\nolimits_2 - 2(\sum\nolimits_1 \sum\nolimits_2)^{1/2})$$

where μ_1 and μ_2 are the mean vectors, $\sum_1$ and $\sum_2$ are the covariance matrices, |||| denotes the Euclidean norm, and $\mathcal{T}_r$ denotes the trace of a matrix.

The resulting FID score represents the similarity between the feature distributions of the real and generated images. Lower FID scores indicate that the generated images are closer to the real images in terms of quality and diversity.

6 Results

We have re-implemented the Globalgenerator network with only four residual blocks as our generation backbone. The three evaluation metrics Attribute Classification Accuracy, Segmentation Accuracy and Fréchet Inception Distance Score are tested both directly on the generator backbone network and our improved network, the result are shown in Table.1 and Fig. 5. Compared with

Table 1: Evaluation on geometry-level facial attribute transfer between pix2pixHD(only Globalgeberator) and our improved model

	Attribute Classification Accuracy (%)	Segmentation Accuracy (%)	FID Score
Pix2PixHD	63.54	83.82	64.19
StarGAN	68.78	91.56	44.27
Our Model	72.46	91.27	48.85
GT	92.3	92.11	-

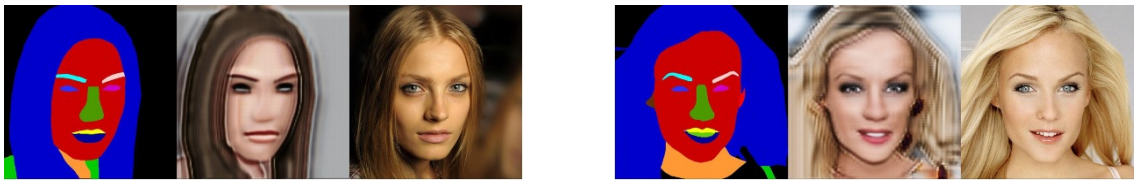

Figure 5: Visual comparison between pix2pixHD(only Globalgeberator) trained for 40 epochs and our improved model trained for 28 epochs

Figure 6: Experiment result of generating random images using the trained model, left is the changed labeled mask and its corresponding image, right is the reference image and the generated image based on the changed labeled mask

the Pix2PixHD model with only a Global generator, our purposed model has an comprehensive improvement, showing that the modification of add a spatial style network is positive. However, StarGAN, another powerful tool for image synthesis applied a technique called "domain classification", has a better FID score than our model. We may consider adding the domain classification in the future work to imporve the model performance.

We also had an experiment on using our trained model to generate a facial image based on a changed labeled mask and a reference image. To make the result more understandable, we provided a result based on an existing labled mask belong to a different face image in the dataset, the result is shown in Fig. 6. Compared with the reference image, we can find that our generated image do maintain lots of style information such as complexion and hair color.

7 Conclusion

In this paper, we proposed a conditional generative adversarial network (GAN) for facial image manipulation. Facial image manipulation is a rapidly growing research area in computer vision and image processing. The goal of this project was to generate high-resolution facial images by manipulating various facial features. The main challenges in our task is to preserve the spatial relationships between different facial components while generating a new image.

Our proposed network incorporates a spatial-aware style network and a Pix2pixHD backbone to generate high-quality facial images while preserving the spatial relationships between different facial components. The pix2pixHD model was used as the basis for this task, with a global generator and multi-scale discriminator. To enhance the model's performance, the spatial feature transform layer and adaptive instance normalization were integrated. The model has a three-input overall design structure, with the top half extracting style information from the reference image and labeled mask, and the bottom half generating images from the input labeled mask. The CelebAMask-HQ dataset was used, and three evaluation metrics, Attribute Classification Accuracy, Segmentation Accuracy, and Fréchet Inception Distance Score, were applied.

The results show that our model has higher ac-

curacy and lower FID score, indicating that the generated images are closer to the original ones. The experiment also demonstrated that our model can effectively preserve the style information of the reference image when generating images given different paired masks and images. Therefore, our proposed method shows potential for practical applications in the field of facial image manipulation.

From the perspective of application, this work has potential applications in entertainment, virtual reality, and forensics, and we believe that our proposed method can contribute to the further advancement of facial image manipulation research.

A Detailed Roles

Task	File names	Who
Implementation of data process	data_preprocess.ipynb	Haolin Ye
Reimplementation and modification of spatial aware network and generator	function.py generator.py model.ipynb	Haolin Ye
Reimplementation of multi-scale discriminator	discriminator.py	Peixu Xin, Jiayu Wang
Implementation of loss functions	Loss_Function.py	Haipeng Liu
Training files and validation	train.ipynb	Runchuan Feng
Designing PPT	-	Haolin Ye , Runchuan Feng, Peixu Xin , Jiayu Wang , Haipeng Liu
Presentation	-	Haolin Ye , Haipeng Liu
Sections 1, 2.1, 2.2, 2.3 of report	-	Runchuan Feng
Sections 3.1, 3.2, 3.2.1, 3.2.2, 4, 4.1 6, Appendix of report	-	Haolin Ye
Sections 3.3, 3.3.1, 3.5 of report	-	Haipeng Liu
Sections 3.3, 3.4, 5.1, 5.2, 5.3, 7 of report	-	Peixu Xin, Jiayu Wang

B Code repository

https://github.com/Leaf-hl/EC-523-Project

References

[1] A. Karnewar and O. Wang, "Msg-gan: Multi-scale gradients for generative adversarial networks," 2020.

[2] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H.-P. Seidel, P. Pérez, M. Zollhöfer, and C. Theobalt, "Stylerig: Rigging stylegan for 3d control over portrait images," 2020.

[3] S. Xu, J. Yang, D. Chen, F. Wen, Y. Deng, Y. Jia, and X. Tong, "Deep 3d portrait from a single image," 2020.

[4] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014.

[5] F. Zhan and C. Zhang, "Spatial-aware gan for unsupervised person re- identification," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 6889–6896.

[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," 2018.

[7] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," 2018.

[8] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," 2017.

[9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2018.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[11] W. Tang, G. Li, X. Bao, and T. Li, "Mscgan: Multi-scale conditional generative adversarial networks for person image generation," 2020.

[12] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," 2016.

[13] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," 2017.

[14] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] S. Park, J. Lee, P. Lee, S. Hwang, D. Kim, and H. Byun, "Fair contrastive learning for facial attribute classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 389–10 398.

[16] A. Obukhov and M. Krasnyanskiy, "Quality assessment method for gan based on modified metrics inception score and fréchet inception distance," in *Software Engineering Perspectives in Intelligent Systems*, R. Silhavy, P. Silhavy, and Z. Prokopova, Eds. Cham: Springer International Publishing, 2020, pp. 102–114.