High-Quality Facial Pose Generation with Latent Space Processing

Wing-Ho Cheng*†, Wan-Chi Siu*†, *Life-FIEEE*, H. Anthony Chan*, *FIEEE*
* St. Francis University Hong Kong and
† The Hong Kong Polytechnic University

***Abstract*—Manipulating facial poses is challenging, especially when addressing significant pose variations. While extensive research has been dedicated to address large poses and manipulate various facial expressions, this frequently results in compromised image quality. The challenge may arise from non-linearity of the latent space. We must navigate a complex path along the high-quality image manifold and determine the optimal direction for the face rotation task, which may secure the most effective disentanglement. Moreover, the regularity of the latent space also affects directly the quality of the resulting image. In this paper, we have made a careful study of the latent space, and deliberately crafted our model to identify the complicated trajectory of rotating facial manipulation with exceptional disentanglement. Our facial pose generative model, aims at enhancing the quality of generated images while preserving the identity and fidelity and achieving better disentanglement. Data acquisition is another challenging aspect, requiring extensive preparation and meticulous setup. To address this, we suggest a flipping technique to mitigate dataset limitations. Ultimately, we strive to strike a balance between image quality and pose generation, ensuring that our results are both visually pleasing and accurately representing the desired facial pose.**

I. Introduction

Fig. 1 Illustration of our model to synthesize various poses from the input images (leftmost image in red).

Facial pose manipulation from a static image involves adjusting the orientation, rotation, and perspective of the human face while retaining its inherent features such as facial expression, hairstyle, and other attributes.

The proposed method entails altering the facial pose of a reference image without fundamentally changing the face's appearance, ensuring that the modified image appears as if captured from a different angle. Fig. 1 illustrate our model to synthesize various poses from an input image (i.e., leftmost image in red). Faces at the right-hand-side are the synthesized faces at different directions (i.e., different poses) ranging from -45, -30, 0, 30 and 45 degrees and their pose values are indicated above the faces.

Manipulating facial poses based on a static image is challenging. The intricacies inherent to the human face and the non-linear characteristics of pose alterations necessitate innovative approaches. Recent developments, particularly in generative models driven by deep learning, have paved the way for novel methodologies to address this difficult task. Unlike supervised tasks in machine learning, facial pose manipulation lacks a clear ground truth dataset for training and evaluation.

StyleGAN [1, 2] stands out as an innovative generative model that leverages the concept of the Generative Adversarial Network (GAN) [3]. This model can generate hyper-realistic human images. Beyond its ability to faithfully replicate facial features, StyleGAN exhibits considerable capabilities in various tasks. By manipulating the latent space, facial images can be transformed into different viewpoints or have their facial expressions altered.

Some models utilize StyleGAN as a generator to generate images [4, 5]. pSp [4] trains an encoder to identify the trajectory of semantic changes within the image domain and uses the StyleGAN generator to generate an image. InterfaceGAN [5] identifies a linear path in the latent space of StyleGAN for specific semantic attributes in the image domain. However, this approach produces artifacts, as the optimal path is typically not linear. Rotate and Render (RR) [6] obtains a 3D fit from the source image, rotates the 3D model to a different pose, and then reconstructs the image by applying texture using a trained model. CFR-GAN [7] employs several techniques to achieve face transformations for various tasks, including 3D face reconstruction and the Swap-R&R strategy. Although many methods are employed to handle pose manipulations; they frequently result in compromised image quality. BPRN [8] employs a Back-Projection method to enhance image quality, which is effective for super-resolution and various other applications. EFGPN [9] utilizes an edge-guided feature extraction encoder and achieved promising results.

A. Contributions

The primary contributions of this work can be outlined as follows. (1) We propose an unsupervised learning approach for pose manipulation through latent space manipulation. This approach eliminates the need to prepare ground truth images in

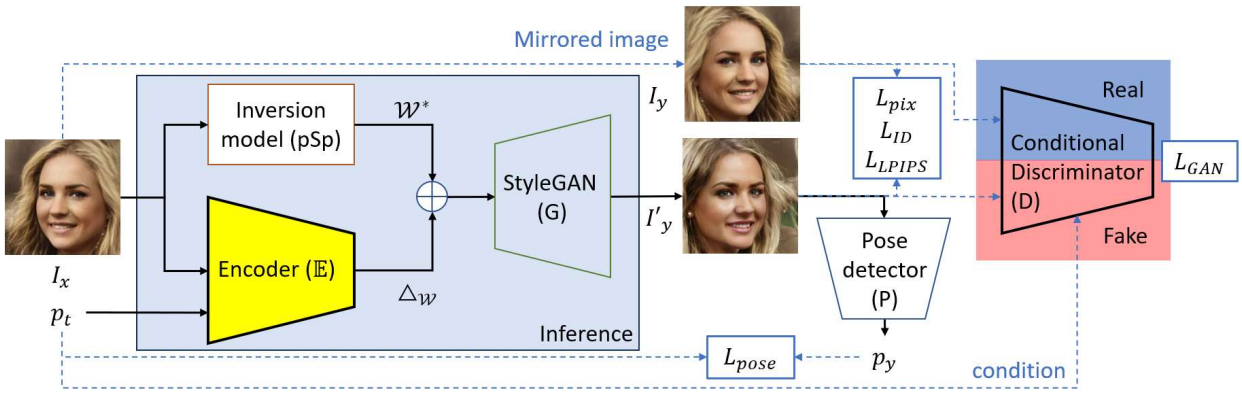

Fig. 2 Architecture of our system. The blue square at the right-hand side is the inference model. The whole model will be used during training.

training by utilizing a face-flipping strategy. (2) The model explicitly separates the learning of identity representation from the features of facial pose, enabling it to produce high-quality image of desired pose. (3) Our architecture utilizes an encoder and the StyleGAN [2] generator to synthesize the image. We also use a conditional discriminator [10] to restrict the generated images conforming it within the facial domain. In order to ensure that the generated face features align with the original image, the back projection technique [8] or the Back projection loss [8, 11] can be used. This architecture improves the quality of the generated image while preserving its identity. (4) The proposed model demonstrates outstanding image quality when generating facial poses on CelebA-HQ [12] and LFW [13] datasets. The high-quality images generated can effectively be utilized in various forms of creative entertainment and enhance the recognition rate of facial recognition systems. As shown below, the experimental results have verified that our approach give remarkable results compared with other approaches in the literature.

II. PROPOSED APPROACH

This paper is on pose generation of a face. We need face generation because we want to generate a face to have various pose appearances or want to make a real face photo to talk with various poses. The method focuses on manipulating the latent vector of an object in StyleGAN [2]. By incorporating a dedicatedly designed encoder to navigate pathways within the latent space, which includes the functionality of facial pose adjustments. The resultant images benefit from the high-quality synthesis capabilities of the StyleGAN model, ensuring realistic and detailed outcomes. Fig.2 shows the overall architecture of our approach to manipulating facial poses. In the diagram, the solid black lines illustrate the main data flow, while the dotted blue lines denote the path of data comparison with loss functions.

Inference: In Fig.2, I_x is the input face. Initially, the input image (I_x) undergoes inversion [4, 9] to produce a latent vector ($\mathcal{W}^*$) in the latent space. The Encoder (E) takes input from both the input image (I_x) and the target pose (p_t). The target pose (p_t) gives the angle to be turned. During the inference stage, we explicitly specify the desired output pose value, p_t. The pose layer is concatenated to the input image (3-layer of RGB) to form a 4-layer tensor and fed into the encoder to produce delta change ($\triangle_W$). This delta change will be added to the latent vector ($\mathcal{W}^*$) produced from the inversion model and fed into the StyleGAN generator (G) to generate the output image. It is worth noting that the pose detector and discriminator are not needed during the inference stage.

Training: Our objective is to train the system to change the pose of the input image I_x such that it matches closely the pose of the target image I_y. During training, the pose is determined by the target image, I_y using our novel image mirroring technique (to be explained shortly in the Training Strategy section) or a real number generated randomly. The input image undergoes the same processing as described in the inference stage to generate the output image. The generated image is compared with the target image to evaluate $\mathcal{L}_{pix}$, $\mathcal{L}_{ID}$ and $\mathcal{L}_{LPIPS}$ losses. Moreover, the pose of the generated image is compared with the target pose to calculate $\mathcal{L}_{pose}$ loss. Finally, the conditional discriminator (D) discriminates the difference between the mirrored image and the generated one, classifying them as real or fake under a condition based on the pose ($\mathcal{L}_{GAN}$). The solid-colored components in the model represent the modules to be trained (i.e., Encoder and Conditional Discriminator), while other components like the version model, StyleGAN (G), and the Pose Detector (P) remain fixed during training. Note that we have to train the encoder and the discriminator model such that they can be improved progressively. The process can be summarized as follows:

$$I'_y = \mathrm{G}(E(I_x, p_t) + \mathcal{W}^*) \quad (1)$$

where I_x is the input image, p_t is the target pose. E is the encoder. G represents the StyleGAN generator network that generates the output image I'_y (Fig. 2).

A. Architecture

The architecture of the encoder uses a Feature Pyramid Network (FPN) [14] built on top of a ResNet [15] backbone, which is similar to those in pSp [4, 9]. Initially, the encoder takes input from both the input image and the pose. The image and pose will be combined to create a 4-channel input. The FPN then extracts features from the input. These features are subsequently mapped to the 18 styles and inputted into StyleGAN generator.

The conditional discriminator used in our model base on the architecture of StyleGAN's discriminator while condition on the input pose (Conditional Discriminator (D) in Fig.2).

B. Training strategy

Novel image mirroring for training: Our methodology leverages the inherent symmetry of the human face to streamline the process of generating training pairs for pose manipulation. We intentionally mirror the image for training as shown in Fig. 2. To explain by an example, if the input image is facing +23 degrees, the target image will be laterally flipped image, and the pose of the target image will be -23 degrees. We use this input image and flipped input image as training pairs to train our model. This approach is grounded in the observation that facial features often exhibit bilateral symmetry, making it feasible to generate a close approximation of the ground truth image solely through mirroring techniques. This also simplifies the data collection process significantly, as well as eliminates the requirement for extensive manual annotation.

C. Loss function

Our encoder employs a variety of loss functions.

1) Pixelwise loss

We utilize the L2 norm distance to align the generated image with its corresponding counterpart, calculated through the following equations:

$$L_{pix} = \left\| I_y - I'_y \right\|_2 \quad (2)$$

where I_y and I'_y are the target image and the generated image respectively and $\|.\|_2$ is the L2 norm distance. Our proposed method has designed that the target image, I_y, will be a mirrored version of I_x if the target pose is facing the opposite direction to the input image or else it will be I_x itself. While other losses like ID loss, perceptual loss, or GAN loss prioritize capturing specific facial features from the source image, they may overlook finer skin details such as facial lines and pores. Using the L2 norm can improve the overall quality and fidelity of the generated face images.

2) ID loss

We employ the ID loss to steer the generation process, maintaining the identity of the source image. The identity of a human face encompasses various factors, such as the shape and size of facial components like the eyes, nose, ears, and mouth, as well as their positioning and coloration. The ID loss aids in capturing these intricate details. This loss quantifies the feature differences between the target and generated images, employing a feature extraction model built on Arcface [16] with a ResNet [15] backbone. It measures the cosine similarity between the output and target images.

$$L_{ID} = 1 - \left\| \mathrm{R}(I_y), \mathrm{R}(I'_y) \right\|_{cos} \quad (3)$$

where R is a pre-trained ArcFace network [16] and $\|.\|_{cos}$ is the cosine similarity.

3) Perceptual loss

In addition to maintaining identity consistency, we preserve the perceptual similarities in the generated images. The Perceptual loss shares similarities with the ID loss but emphasizes capturing details that enhance visual perception. To achieve this, we integrate the LPIPS loss [17] into our training framework. This loss metric helps the model learn and reinforce perceptual similarities between the generated face images and their corresponding target face images.

$$L_{LPIPS} = \left\| \mathrm{K}(I_y) - \mathrm{K}(I'_y) \right\|_2 \quad (4)$$

where K denotes the perceptual feature extractor [17] and $\|.\|_2$ is the L2 norm distance.

4) Conditional GAN loss [10]

To guarantee the high quality and photorealism of the generated image, we employ a GAN discriminator for validation purposes. The GAN loss contributes in capturing the overall distribution of facial images and ensures that the output closely aligns with facial properties. By adopting the idea of conditional GAN [10], we utilize pose conditioning to separate the pose feature from the facial feature, enhancing disentanglement from the generated face image. The adversarial loss can be calculated as:

$$L_{GAN} = -log\left(\mathrm{D}(I'_y, p_t)\right) \quad (5)$$

where I'_y and p_t are the generated image and target pose respectively. D represents the discriminator network conditioned on the pose of the image that is leaned to distinguish whether the generated face appears natural and realistic.

5) Pose loss

To evaluate the accuracy of the facial pose transformation, we employ a pre-trained facial pose predictor network, Hopenet [18]. This network analyzes a facial image and determines its pose. We then compute the L2 loss by measuring the difference between the facial poses of the generated and target images as follows:

$$L_{pose} = \left\| \mathrm{P}(I_y) - \mathrm{P}(I'_y) \right\|_2 \quad (6)$$

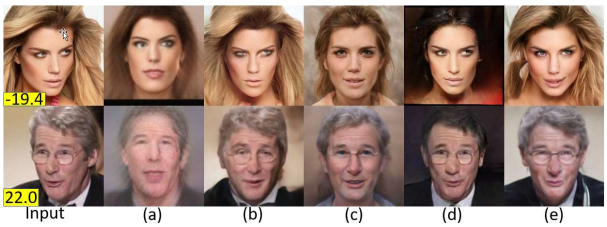

Fig. 3 Visual comparison with other SOTA models for the frontalization task using the CelebA-HQ (Upper row) and LFW (Lower row) dataset. (a) CFR-GAN [7], (b) InterfaceGAN [5], (c) pSp [4], (d) RR [6] and (e) ours.

where P is the facial pose prediction network, Hopenet [18] and $\|.\|_2$ is the L2 loss.

6) Back projection loss [8]

This loss function operates as a revert of the normal forward pass, where after generating the target pose, the generated image is then required to revert to the original pose. This method offers several advantages. Firstly, in the absence of a supervised or target facial image, we obtain a reversed target facial image that should ideally match the original input face image. This allows for precise comparison with the target image, facilitating accurate loss calculation.

Furthermore, this approach imposes constraints on the model to preserve the identity of the face image during the transformation process. If the model diverges from the correct transformation path, it becomes difficult to revert the image back to its original form. By enforcing this constraint through Back projection loss [8, 11], the model learns to produce high-quality face image with improved both fidelity and identity. We denote the Back projection loss as L_{cycle}. The same set of losses introduced in forward pass is applied to cycle pass.

7) Total loss

In summary, the loss for the forward pass is expressed as follow:

$$L_{forward} = w_{pix} \cdot L_{pix} + w_{ID} \cdot L_{ID} + w_{LPIPS} \cdot L_{LPIPS} + w_{GAN} \cdot L_{GAN} + w_{pose} \cdot L_{pose} \quad (7)$$

where w_{pix}, w_{ID}, w_{LPIPS}, w_{GAN}, w_{pose} are constants defining the loss weights.

To alleviate the impact of a lesser accurate approximation of the target image when the target pose is randomly generated, the weighting for pixel-wise loss, w_{pix}, will be reduced.

Combining both forward and cycle passes, the total loss is given by:

$$Total\ Loss = L_{forward} + w_{cycle} \cdot L_{cycle} \quad (8)$$

where w_{cycle} is the weight of the Back projection loss.

III. Experimental Works

A. Datasets

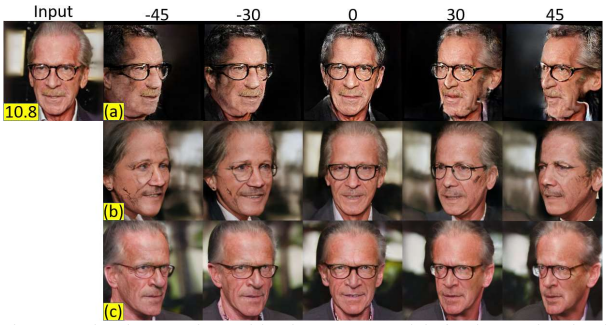

Fig. 4 Visual comparison with other SOTA models for the rotational task using the CelebA-HQ dataset. The input images are depicted in the leftmost column, while the synthesized images are presented on the right-hand side. (a) RR [6], (b) InterfaceGAN [5] and (c) ours.

We employed the unlabeled image dataset FFHQ [1] for training, consisting of 70,000 high-quality human face images sourced from the Flickr photo-sharing platform. This dataset is highly valued for its high resolution and extensive size. The images are already aligned with each image having a resolution of 1024x1024 pixels. The facial images are displayed in various poses. It proves particularly beneficial for applications requiring detailed facial features and realistic representations.

B. Implementation details

We implemented our model based on PyTorch and conducted our experimental work using training on a single GPU RTX3090 with 24G VRAM. We greatly appreciate the source code shared by Alaluf, et al. [19], which serves as a reference for our implementation. During the training phase, we adopted the Adam [20] optimizer with a batch size of 6. Throughout the training process, we maintained fixed parameters for the inversion model, StyleGAN generator [2], and the pose detection model, Hopenet [18]. The pre-trained models were obtained from official channels and loaded to these fixed models before training. We solely trained the encoder and the conditional discriminator network. We utilized FFHQ [1] as our training dataset, comprising 70,000 high-resolution faces with a resolution of 1024x1024.

In the testing phase, we employed CelebA-HQ [12] and LFW [13] to assess our trained network and compared methods. All experiments were conducted using the model output from unseen images during the training phase. Obviously, the discriminator will not be included in the testing phase.

C. Qualitative Comparison

We conducted a comparison with four state-of-the-art models with source code available. Two models, InterfaceGAN [5] and pSp [4] which utilize StyleGAN as a generator which share similarities with our architecture. The other two models: Rotate and Render (RR) [6], and CFR-GAN [7], employ different approaches for comparison. A short demonstration is included in the following webpage:

https://cis.sfu.edu.hk/2024.RotationalFace/index.html.

Our analysis in Fig. 3 indicates that the quality of our output surpasses that of the compared models in both datasets. Our model exhibits superior disentanglement across various aspects.

For instance, while the facial details and hairstyle preservation in the compared models (Column (a)-(d) in Fig.3) are inadequate, our model (Column (e) in Fig. 3) closely recovers the original input photo. Our model has produced the most photorealistic output compared to all other models. Artifacts rarely appear in the images generated by our model.

When outputs of Rotate and Render (RR)'s [6] (Column (d) in Fig.3) possess sufficient realism, they tend to lack the refinement and clarity found in higher-resolution images. This lower resolution may result in a loss of details and the overall visual quality.

InterfaceGAN [5] (Column (b) in Fig.3) adjusts the pose based solely on the original input and there is no specific input control mechanism to precisely regulate the output pose of the generated image. In order to compare this approach, we have to choose the frontal pose solely by human inspection. The output image quality by InterfaceGAN is significantly influenced by the amount of pose that deviates from the original photo. The synthesized image sometimes displays artifacts or image distortion due to diverges from the optimal image manifold (e.g. Fig.3. column (b) of 2nd row).

The pSp model [4] (Column (c) in Fig.3) can generate highly detailed images, sometimes these details may appear exaggerated, leading to images that deviate from reality. In their model, features such as shape, expression, and hairstyle were averaged and exhibited similarity across different instances. In contrast, our approach (Column (e) in Fig.3) better preserves the identity and achieves a higher level of disentanglement. Interestingly, pSp outputs often exhibit blurred and plain backgrounds. In contrast, our model offered vibrant color representation closer to the original and effectively maintained the overall appearance and atmosphere of the image, resulting in outputs that closely resemble the original photos.

Our synthesized output surpasses CFR-GAN [7] (Column (a) in Fig.3) in terms of realism and fidelity compared to the original image.

For the task of altering the pose of an image, models like pSp, CFR-GAN does not provide the feature to rotate a face to a specific view. Instead, we focus our comparison on RR [6] and InterfaceGAN [5], as they are capable of generating poses different from the frontal pose, as depicted in Fig.4. Let us showcase our model results across a range of pose angles ranging form -45, -30, 0, +30 to +45 degrees, and compare them with results from other models. Once more, our model (Fig.4 row (c)) demonstrates superior image quality compared to other models, and this distinction becomes apparent in larger pose generation.

All models exhibited varying degrees of artifacts or distortions as the magnitude of the generated pose increases. However, our model shows better disentanglement capabilities; for instance, glasses intermittently appeared or disappeared during the generation of different poses (Fig.4 row (b)). In order to reduce variations across different poses of the same image, such as stochastic fluctuations in hair generation, we have fixed the per-pixel noise input in the StyleGAN generation settings. Typically, distortions became bad at the larger poses on the left and right sides.

Although RR (Fig.4 row (a)) preserves many details and fidelity, it deviates considerably from the original image. Its image quality is realistic only around the frontal position, however, there were significant changes to the background, diverging from the original photos. Furthermore, the face undergoes more distortion at larger poses, and our model always does a better job for preserving the hairstyle and the prevention of artifact generation.

InterfaceGAN (Fig.4 row (b)) adjusts the pose based on the original input and the magnitude cannot be precisely controlled. Furthermore, generating poses with larger differences have proved to be more challenging and resulted in increased distortion and the generation of more artifacts. InterfaceGAN employs a linear interpolation of the original pose in the latent space. A larger pose difference resulted in larger discrepancies from the optimal facial latent representation, especially when the optimal path has deviated from the linear trajectory. This also suggested that the optimal path is non-linear in nature.

D. Quantitative Comparisons

It is challenging to perform a quantitative comparison of image outputs when the pose varies. The absence of an accurate ground truth for comparison also further complicates quantitative comparison. Generation of large poses is not the sole consideration; image quality is most important.

TABLE 1.
FID Scores (CelebA-HQ). The best model is highlighted in red while the 2nd best is highlighted in blue.

Method	Avg	0	-45	-30	-15	15	30	45
pSp [4]	-	67.8	-	-	-	-	-	-
CFR-GAN [7]	-	187.3	-	-	-	-	-	-
RR [6]	118.81	84.3	155.1	133.9	105.4	100.8	119.5	132.7
Ours	**23.3**	**23.2**	**30.6**	**20.7**	**20.2**	**21.0**	**21.2**	**26.2**

TABLE 2.
FID Scores (LFW). The best model is highlighted in red while the 2nd best is highlighted in blue.

Method	Avg	0	-45	-30	-15	15	30	45
pSp [4]	-	145.2	-	-	-	-	-	-
CFR-GAN [7]	-	106.2	-	-	-	-	-	-
RR [6]	58.6	50.9	75.1	51.3	49.1	50.1	54.4	79.1
Ours	**34.4**	**35.4**	**41.5**	**32.5**	**32.2**	**32.4**	**31.6**	**35.4**

Let us compare the frontalization output from different models using the Fréchet Inception Distance (FID) score [21] (lower value indicated better quality) and the Rank-1 recognition rate. To evaluate the Rank-1 recognition rate, we extracted features of the images in both datasets using LightCNN [22] for the original images and the generated pose

image. We then compared the original and generated images, selecting the pair with the top cosine similarity to compare their identities. Since CelebA-HQ lacks labeling in the datasets, each image is considered a unique identity, even if there are instances of the same identity present.

While InterfaceGAN can rotate images to different angles, it is hard to determine the output pose angle of the generated image. As a result, we did not use it for quantitative comparison. Note that only RR can rotate images into different poses, and there were not data available for other models.

Table 1 and Table 2 present the FID [21] scores of various models for different generated poses, our model gains significant superiority over other models, indicating that it achieves the highest similarity in terms of overall similarity and color distribution compared to the original datasets.

TABLE 3.
Rank-1 Accuracy (CelebA-HQ). The best model is highlighted in red while the 2nd best is highlighted in blue.

Method	Avg	0	-45	-30	-15	15	30	45
pSp [4]	-	**99.5**	-	-	-	-	-	-
CFR-GAN [7]	-	80.2	-	-	-	-	-	-
RR [6]	92.5	99.4	79.5	94.7	98.9	98.8	94.9	81.1
Ours	**97.0**	98.9	**90.5**	**97.9**	**99.3**	**99.1**	**98.6**	**94.9**

TABLE 4.
Rank-1 Accuracy (LFW). The best model is highlighted in red while the 2nd best is highlighted in blue.

Method	Avg	0	-45	-30	-15	15	30	45
pSp [4]	-	**98.3**	-	-	-	-	-	-
CFR-GAN [7]	-	53.8	-	-	-	-	-	-
RR [6]	83.4	97.4	66.5	87.4	94.9	93.7	84.3	59.5
Ours	**94.3**	97.9	**86.2**	**96.0**	**98.3**	**98.0**	**96.6**	**87.1**

Table 3 and Table 4 list the comparison of the Rank-1 Accuracy with other state-of-the-art models. For all models, the recognition rates were generally higher around frontal poses and the rate tended to decrease at larger poses. It may be because the datasets were dominated by frontal-facing images. In general, rotation to frontal image means fewer pose variations.

Our model has the highest average accuracy of 94.3% (CelebA-HQ) (Table 3. Row 4 Column 1) and 97.0 (LFW) (Table 4. Row 4 Column 1). Note that the result may be impacted by the presence of multiple images for the same identity existing within the CelebA-HQ dataset. Since the dataset lacks labeling, each image is treated as a distinct identity.

Interestingly, our model achieves its highest recognition rate for faces with small pose angle of -15 degrees, reaching 98.3% (Table 3. Row 4 Column 5) and 99.3% (Table 4. Row 4 Column 5) in CelebA-HQ datasets and LFW datasets respectively. Furthermore, our model outperforms other models notably for faces with large poses. In large poses, our model sustains exceptional image quality, demonstrating its ability to successfully disentangle facial features and maintain identity after pose rotation. Consequently, our model achieves a recognition rate significantly surpassing that of the other models.

IV. Conclusions

In conclusion, our research primarily focuses on manipulating the latent space of StyleGAN [2] to transform the original image into a different pose. By training the model in an unsupervised manner and incorporating image-flipping techniques, we aim to enhance its ability to learn diverse facial properties and enhance its image generation quality for various poses. Our model integrates Back projection loss [8, 11] to enhance output quality and leverages conditional GANs [10] to improve disentanglement during face rotation, thereby preserving identity. One future direction is the incorporation of facial expressions into the rotated faces. This not only would improve the visual fidelity of the generated images but also expand the potential use cases in fields such as virtual reality, animation, and human-computer interaction.

V. Acknowledgment

This work is partly supported by the St. Francis University (1SG200206) and UGC Grant (UGC/FDS11/E05/22) of the Hong Kong Special Administrative Region.

References

[1] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401-4410.

[2] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 8110-8119.

[3] I. Goodfellow et al., "Generative adversarial nets," Advances in neural information processing systems, vol. 27, 2014.

[4] E. Richardson et al., "Encoding in style: a stylegan encoder for image-to-image translation," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 2287-2296.

[5] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of gans for semantic face editing," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9243-9252.

[6] H. Zhou, J. Liu, Z. Liu, Y. Liu, and X. Wang, "Rotate-and-render: Unsupervised photorealistic face rotation from single-view images," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 5911-5920.

[7] Y.J. Ju, G.H. Lee, J.H. Hong, and S.W. Lee, "Complete face recovery gan: Unsupervised joint face rotation and de-occlusion from a single-view image," in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2022, pp. 3711-3721.

[8] Z.S. Liu, W.C. Siu, and Y.L. Chan, "Joint back projection and residual networks for efficient image super-resolution," in 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2018: IEEE, pp. 1054-1060.

[9] X. Cheng, W.C. Siu and J. Yang, "Large-Scale Blind Face Super-Resolution via Edge Guided Frequency Aware Generative Facial Prior Networks", Proceedings, pp. 1638-1643, Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC'2022), Chiang Mai, Thailand, 2022, doi: 10. 23919/APSIPAASC55919.2022.9980270.

[10] P. Isola, J.Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125-1134.

[11] J.Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223-2232.

[12] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," arXiv preprint arXiv:1710.10196, 2017.

[13] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition, 2008.

[14] T.Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117-2125.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.

[16] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4690-4699.

[17] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586-595.

[18] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 2074-2083.

[19] Y. Alaluf, O. Patashnik, and D. Cohen-Or, "Only a matter of style: Age transformation using a style-based regression model," ACM Transactions on Graphics (TOG), vol. 40, no. 4, pp. 1-12, 2021.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[21] 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," Advances in neural information processing systems, vol. 30, 2017.

[22] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," IEEE transactions on information forensics and security, vol. 13, no. 11, pp. 2884-2896, 2018.