One Shot Face Swapping on Megapixels

Yuhao Zhu [1], Qi Li * [1,2], Jian Wang [1,3], Chengzhong Xu [2], Zhenan Sun[1,3]
[1]Center for Research on Intelligent Perception and Computing, NLPR, CASIA
[2]State Key Laboratory of IoTSC, Faculty of Science and Technology, University of Macau
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences
{yuhao.zhu, jian.wang}@cripac.ia.ac.cn, {qli, znsun}@nlpr.ia.ac.cn, czxu@um.edu.mo

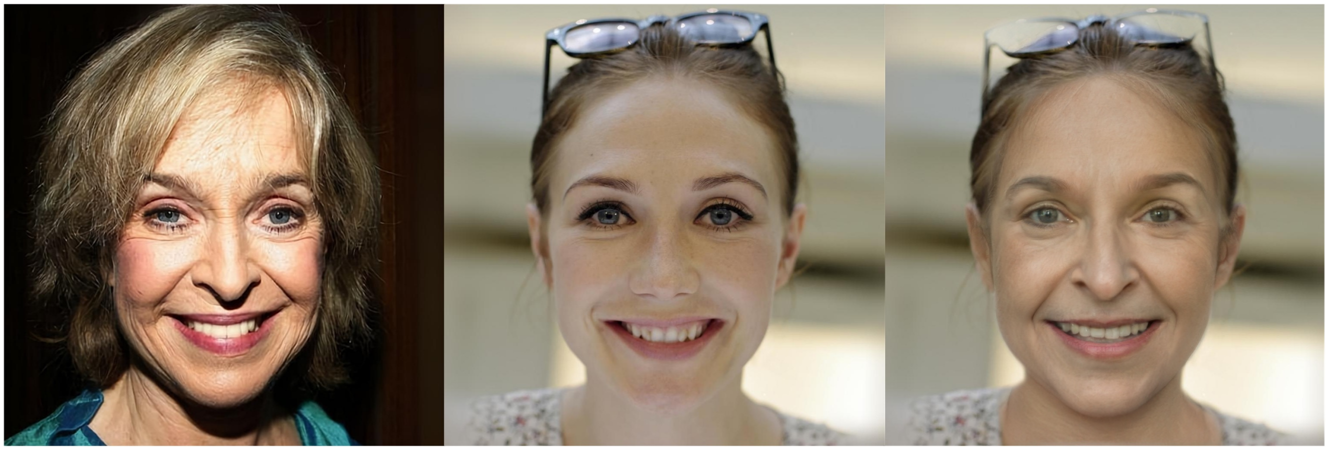

Figure 1. Example of a swapped face. Left: source image that represents the identity; Middle: target image that provides the attributes; Right: the swapped face image. All images are in 1024^2.

Abstract

Face swapping has both positive applications such as entertainment, human-computer interaction, etc., and negative applications such as DeepFake threats to politics, economics, etc. Nevertheless, it is necessary to understand the scheme of advanced methods for high-quality face swapping and generate enough and representative face swapping images to train DeepFake detection algorithms. This paper proposes the first Megapixel level method for one shot Face Swapping (or MegaFS for short). Firstly, MegaFS organizes face representation hierarchically by the proposed Hierarchical Representation Face Encoder (HieRFE) in an extended latent space to maintain more facial details, rather than compressed representation in previous face swapping methods. Secondly, a carefully designed Face Transfer Module (FTM) is proposed to transfer the identity from a source image to the target by a non-linear trajectory without explicit feature disentanglement. Finally, the swapped faces can be synthesized by StyleGAN2 with the benefits of its training stability and powerful generative capability. Each part of MegaFS can be trained separately so the requirement of our model for GPU memory can be satisfied for megapixel face swapping. In summary, complete face representation, stable training, and limited memory usage are the three novel contributions to the success of our method. Extensive experiments demonstrate the superiority of MegaFS and the first megapixel level face swapping database is released for research on DeepFake detection and face image editing in the public domain.

1. Introduction

Given two face images, face swapping refers to transferring the identity from the source image to the target image, while the facial attributes of the target image hold intact. It has attracted extensive attention in recent years for its broad application prospects in entertainment [4, 24], privacy protection [6, 33], and theatrical industry [34].

Existing face swapping methods can be roughly divided into two categories: subject-specific and subject agnostic methods. Subject-specific face swapping methods [11, 27, 34] need to be trained and tested on the same pair of subjects, which restricts their potential applications. On the contrary, subject agnostic face swapping methods

[*] Corresponding author

[35, 5, 28, 38, 36] can be applied to arbitrary identities without additional training procedures. In this paper, we focus on a more challenging topic: *one shot face swapping*, where only one image is given from the source and target identity for both training and testing.

With the rapid growth of high resolution image and video data on the web, it becomes increasingly popular to process high resolution samples. However, generating high resolution swapped faces is rather difficult because of the following problems. Firstly, information is insufficient for high-quality face generation due to the compressed representation in an end-to-end framework [5, 35, 28]. Secondly, adversarial training is unstable [8], which confines the resolution of previous methods only up to 256^2. Thirdly, the GPU memory limitation makes the training untenable, or the training batch is bounded by a small size, which aggravates the collapse of the training process.

To this end, this paper proposes the first Megapixel level one shot Face Swapping method (MegaFS) by adopting the "divide and conquer" strategy in three steps. Firstly, to overcome the information loss in the encoder, we adopt GAN Inversion methods [42, 10, 16, 1, 2, 10, 31, 11, 59] and propose a Hierarchical Representation Face Encoder (HieRFE) to find the complete face representation in an extended latent space $\mathcal{W}^{++}$. Secondly, to modify face representations and resolve the problem of previous latent code manipulation methods [43, 17, 48, 51, 50, 7, 37, 3] that only one attribute can be modified once a time, a novel swapping module, Face Transfer Module (FTM), is proposed to control multiple attributes synchronously without explicit feature disentanglement. Finally, the unstable adversarial training problem is evaded by exploiting StyleGAN2 [23] as the decoder, which is fixed and the discriminator is not used for optimization. Each part of MegaFS can be trained separately so the GPU memory requirements are satisfied for megapixel face swapping. The contributions of this paper can be summarized as:

- To the best of our knowledge, the proposed MegaFS is the first method that can conduct one shot face swappings at megapixel level.

- For encoding and manipulating the complete face representation, faces are encoded by HieRFE hierarchically in the new extended latent space $\mathcal{W}^{++}$ and a new multistep non-linear latent code manipulation module, FTM, is proposed to manage multiple attributes synchronously without explicit feature disentanglement.

- Experimental results on benchmark dataset have shown the effectiveness of the proposed MegaFS. Furthermore, the first megapixel face swapping database is released for research of DeepFake detection and face image editing in the public domain.

2. Related Works

2.1. Face Swapping

Subject-specific face swapping methods are popular in recent years, where DeepFake [11] and its variants are trained using pairwise samples. Besides, Korshunova *et al.* [27] model different source identities separately, such as a CageNet for Nicolas Cage, or a SwiftNet for Taylor Swift. Recently, Disney Research realizes high resolution face swapping [34], but it requires training decoders for different subjects, which hinders its generalization. Besides, it is time consuming and difficult for subject-specific methods to train specific models for distinct pairs of faces [14, 46, 45, 29, 20, 55, 26]. Subsequently, subject agnostic face swapping methods break the limitations of previous subject-specific face swapping methods. Realistic Neural Talking Head [39] adopts meta-learning to relieve the pain of fine-tuning on different individuals. FaceSwapNet [56] proposes a landmark swapper to handle the identity leakage problem from landmarks. In the meanwhile, other mindsets follow the attribute disentanglement heuristic to explore new high fidelity face swapping frameworks. FSNet [35] represents the face region of the source image as a vector, which is combined with a non-face target image to generate the swapped face image. IPGAN [5] disentangles identities and facial attributes as different vectorized representaions. Based on previous works, FSGAN [36] and FaceShifter [28] achieve state-of-the-art results by their outstanding performance.

2.2. GAN Inversion

Based on a well-trained GAN, GAN Inversion, or Latent Space Embedding, tries to find the latent code that can accurately reconstruct a given image synthesized. To this end, two problems need to be settled: determining a proper latent space and designing an algorithm to search for the optimal latent code within that space. As for the latent space, early methods perform image inversion into $\mathcal{W} \in \mathbb{R}^{1\times512}$ [42, 19, 17], while later works [1, 2, 11, 10] extend the latent space to $\mathcal{W}^{+} \in \mathbb{R}^{18\times512}$, which proves to have better reconstruction results. As for the inversion algorithms, they either train an encoder [42, 10, 16] to predict latent codes of images or minimize the error between predicted and given images by optimizing latent codes from random initializations [1, 2, 10, 31]. Some methods [11, 59] combine both to optimize latent codes initialized by encoders.

2.3. Latent Code Manipulation

Latent Code Manipulation, or Latent Control, is another attractive research area to manipulate latent codes based on the observation that semantic editing operations can be realized by adding high dimensional directions [43]. Several linear semantic directions, or trajectories, of $\mathcal{W}$ are found

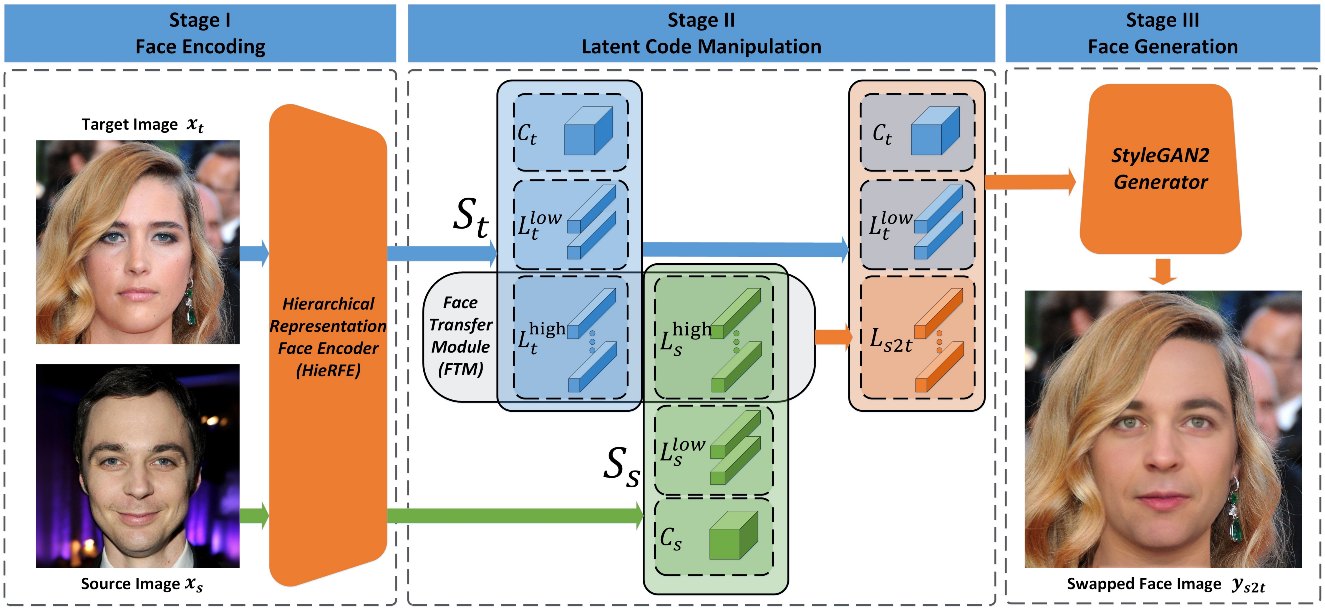

Figure 2. The proposed MegaFS consists of three stages: Face Encoding, Latent Code Manipulation, and Face Generation. Firstly, HieRFE projects two face images into latent space $\mathcal{W}^{++}$. Then FTM manipulates L_s^{high} and L_t^{high} in two hierarchical latent sets S_s and S_t to get L_{s2t}. Finally, the swapped face image y_{s2t} can be synthesized by a pre-trained StyleGAN2 generator from C_t, L_t^{low}, and L_{s2t}.

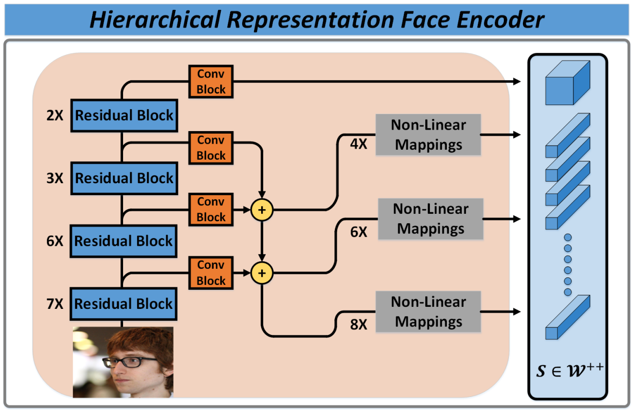

Figure 3. HieRFE consists of a ResNet50 backbone based on residual blocks, a feature pyramid structure based on FPN, and eighteen lateral non-linear mapping networks, in which $n\times$ refers to the number of the corresponding parts.

[17, 48]. StyleRig [51] and PIE [50] propose to manipulate latent space through an existing 3D model [7], which successfully control facial poses, expressions, and illuminations. Previous methods [17, 51, 37, 43] have found good controllability of StyleGAN based on the assumption that semantic directions in StyleGAN latent space are linear. Recently, StyleFlow [3] achieves better manipulation results through non-linear trajectories.

3. Method

Fig.2 demonstrates the overall pipeline and notations of the proposed MegaFS, which combines the identity information from a source image x_s and attribute information from a target image x_t to generate the final swapped face image y_{s2t}. In the following, we will present the details of our method.

3.1. Hierarchical Representation for Face Swapping

In the first stage, face images are projected into latent space $\mathcal{W}^{++}$ using Hierarchical Representation Face Encoder (HieRFE) to deposit complete face information. The structure of HieRFE is detailed in Fig.3.

Specifically, HieRFE consists of a ResNet50 backbone based on several residual blocks [18], a feature pyramid structure based on FPN [30] for feature refinement, and eighteen lateral non-linear mapping networks for latent code prediction. Please refer to the corresponding papers for details of residual blocks and FPN. As for the non-linear mapping network, it comprises repeated downsampling, convolution, batchnorm, and leakyReLU layers until the feature map can be pooled as a vector, *i.e.*, $l \in \mathbb{R}^{1\times512}$.

Then, the constant input of StyleGAN2 predicted by the backbone and four latent codes predicted by the smallest feature map, denoted as $C \in \mathbb{R}^{4\times4\times512}$ and $L^{low} \in \mathbb{R}^{4\times512}$, represent low-level topology information. Other latent codes are gathered as $L^{high} \in \mathbb{R}^{14\times512}$ to represent high-level semantic information. Finally, subscript s and t are adopted to represent the source and target images if it is necessary in the following paper.

3.2. Synchronized Control of Multiple Attributes

During the second stage, Face Transfer Module (FTM) is proposed to control multiple attributes of identity information in a synchronized manner for face swapping demands. In detail, FTM contains 14 Face Transfer Blocks, the number of which equals that of l^{high}.

As shown in Fig.4, each Face Transfer Block contains three identical transfer cells. In each transfer cell, l_s^{high} and l_t^{high} are firstly concatenated to l_c^{high}, which collects all information from the source and target images. Then l_s^{high} is refined to $\hat{l}_s^{high}$ through a two-step non-linear trajectory:

$$\mathcal{T}(l_c^{high}, l_s^{high}) = \mathcal{T}_2(l_c^{high}, \mathcal{T}_1(l_c^{high}, l_s^{high})) \tag{1}$$

in which

$$\begin{aligned} \mathcal{T}_1(a, b) &= sigmoid(K_1(a)) \times b \\ \mathcal{T}_2(a, b) &= Tanh(K_2(a)) + b \end{aligned} \tag{2}$$

where $K_1(\cdot)$ and $K_2(\cdot)$ denote two linear layers. The trajectory is crafted based on the following illustrations. In the first step, the multiplication coefficients are scaled in range $(0, 1)$ after sigmoid activation, where l_s^{high} is designed to discard irrelevant semantics except for the identity information. In the second step, l_s^{high} accepts a small amount of target semantic attributes by shifting in the latent space. Similarly, l_t^{high} is processed in parallel but for discarding target identity while holding other semantics. Finally, the transferred latent code $l_{s2t} \in L_{s2t}$ can be predicted as

$$l_{s2t} = \sigma(\omega)\hat{l}_t^{high} + (1 - \sigma(\omega))\hat{l}_s^{high} \tag{3}$$

where $\omega \in \mathbb{R}^{1 \times 512}$ is a trainable weight vector, and σ stands for the sigmoid activation. The transferred latent codes L_{s2t} is composed by gathering all predicted l_{s2t}.

3.3. High-Fidelity Face Generation

Finally in the third stage, C_s and L_s^{low} are discarded since they contain negligible identity information from x_s. The swapped face image y_{s2t} can be generated by feeding StyleGAN2 generator with C_t, L_t^{low} and L_{s2t}.

By taking StyleGAN2 as the decoder, face swapping through latent space differentiates our method from other face swapping frameworks. Firstly, it provides an extended latent space for complete face representation, which makes detailed face generation feasible. Secondly, it makes our method operating globally in $\mathcal{W}^{++}$ instead of locally on feature maps, which is desirable as it can conduct non-linear transformations through latent code manipulations without local distortions. Thirdly, it does not require explicit attributes disentanglement, which makes the training process straightforward without tricky loss functions and hyper-parameter settings.

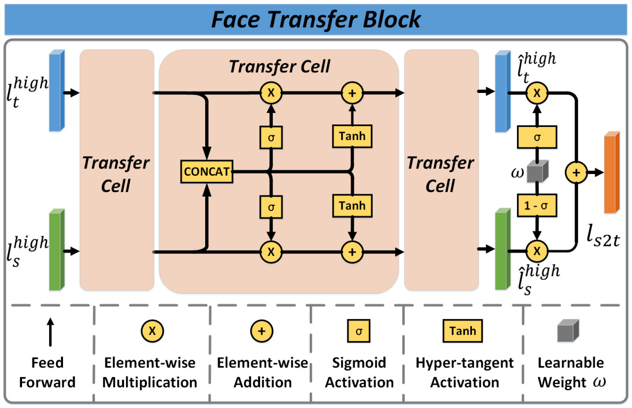

Figure 4. Inside FTM, each Face Transfer Block contains three identical transfer cell. After being processed by three cells, two refined vectors are weighted by a learnable weight ω and summed as the final output.

3.4. Objective Functions

For each part of MegaFS, HieRFE and FTM are trained sequentially, while StyleGAN2 generator remains intact.

Objective function of HieRFE: Following the previous work [44], we make use of three objectives for supervising a pair of input image x and its reconstruction image $\hat{x}$, including pixel-wise reconstruction loss $\mathcal{L}_{rec}$, Learned Perceptual Image Path Similarity (LPISP) loss $\mathcal{L}_{LPIPS}$ [58], and identity loss $\mathcal{L}_{id}$ as follows:

$$\mathcal{L}_{rec} = \|x - \hat{x}\|_2 \tag{4}$$

$$\mathcal{L}_{LPIPS} = \|F(x) - F(\hat{x})\|_2 \tag{5}$$

$$\mathcal{L}_{id} = 1 - \cos(R(x), R(\hat{x})) \tag{6}$$

where $\|\cdot\|_2$ denotes ℓ_2 distance, $F(\cdot)$ denotes the perceptual feature extractor, $R(\cdot)$ denotes the ArcFace [12] recognition model, $\cos(\cdot, \cdot)$ denotes the cosine similarity of two face embeddings.

In addition, as face swapping needs pose and expression controllability, we introduce landmarks loss $\mathcal{L}_{ldm}$ to measure ℓ_2 difference between the predicted landmarks of the input faces and the corresponding ones of reconstructed faces as following:

$$\mathcal{L}_{ldm} = \|P(x) - P(\hat{x})\|_2 \tag{7}$$

where $P(\cdot)$ denotes the facial landmark predictor [54]. The overall loss function for training HieRFE is

$$\mathcal{L}_{inv} = \lambda_1\mathcal{L}_{rec} + \lambda_2\mathcal{L}_{LPIPS} + \lambda_3\mathcal{L}_{id} + \lambda_4\mathcal{L}_{ldm} \tag{8}$$

where $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are loss weights. Besides, x and $\hat{x}$ need resizing as the input of each model before calculating the loss function.

Objective function of FTM: For training FTM, four losses are proposed, including:

$$\mathcal{L}'_{rec} = \|x_s - \hat{x}_s\|_2 + \|x_t - \hat{x}_t\|_2 \tag{9}$$

$$\mathcal{L}'_{LPIPS} = \|F(x_t) - F(y_{s2t})\|_2 \tag{10}$$

$$\mathcal{L}'_{id} = 1 - \cos(R(x_s), R(y_{s2t})) \tag{11}$$

$$\mathcal{L}'_{ldm} = \|P(x_t) - P(y_{s2t})\|_2 \tag{12}$$

Besides, $\mathcal{L}_{norm}$ is leveraged to stabilizes the training process.

$$\mathcal{L}_{norm} = \left\| L_s^{high} - L_{s2t} \right\|_2 \tag{13}$$

Similarly, the overall loss function for training FTM is

$$\mathcal{L}_{swap} = \varphi_1 \mathcal{L}'_{rec} + \varphi_2 \mathcal{L}'_{LPIPS} + \varphi_3 \mathcal{L}'_{id} + \varphi_4 \mathcal{L}'_{ldm} + \varphi_5 \mathcal{L}_{norm} \tag{14}$$

where $\varphi_1, \varphi_2, \varphi_3, \varphi_4$ and φ_5 are loss weights. Finally, when FTM converges, the proposed method is ready for face swapping on megapixels.

4. Experiments

In this section, we will first show the effectiveness of the proposed method by comparing it with other state-of-the-art methods provided in FaceForensics++ [46]. Then the superiority of our method is demonstrated by conducting face swapping on CelabA-HQ [21]. Finally, an ablation study is presented to reveal the necessity of each component of our method.

4.1. Datasets and Implementation Details

CelebA [32]: This dataset is built for face detection, facial landmark localization, attribute recognition and control, and face synthesis. It contains 202,599 celebrity images with 40 labeled attributes and 5 landmark location annotations.

CelebA-HQ [21]: It is a high-quality version of CelebA dataset. All 202,599 images in CelebA are processed by two pre-trained neural nets for denoising and super-resolution, resulting in 30,000 high-quality images.

FFHQ [22]: The dataset contains 70,000 megapixel face images collected from Flickr. FFHQ has considerable variations of age, ethnicity, gender, and background.

FaceForensics++ [46]: It is a forensics dataset consisting of 1,000 original video sequences from YouTube that have been manipulated with five automated face manipulation methods: Deepfakes, Face2Face, FaceSwap, NeuralTextures, and FaceShifter, in which Deepfakes, FaceSwap, and FaceShifter are face swapping methods, while Face2Face and NeuralTextures are reenactment algorithms.

Implementation Details: In all experiments, learning rate of the Adam optimizer [25] is set to 0.01. We set $\lambda_1, \lambda_2, \lambda_3$ and λ_4 to 1, 0.8, 1, and 1000. We set $\varphi_1, \varphi_2, \varphi_3, \varphi_4$ and φ_5 to 1, 32, 32, 24, and 100000. In addition, 200,000 faces are randomly sampled as auxiliary data by running StyleGAN2.

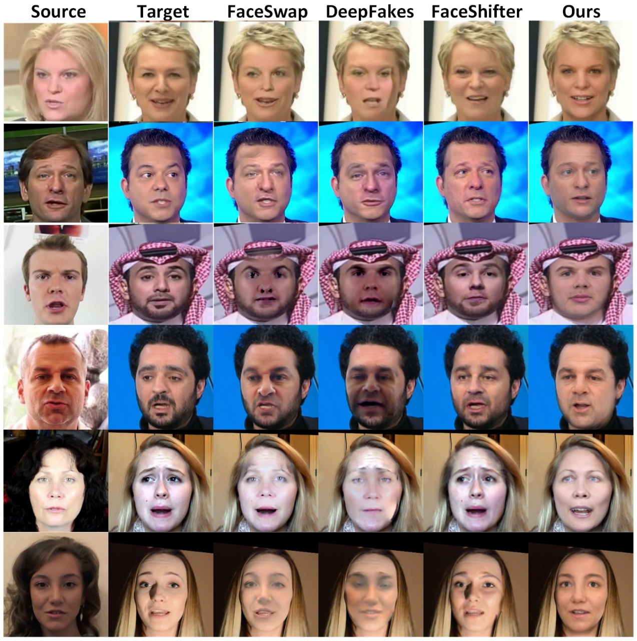

Figure 5. Qualitative comparison results of FaceSwap, DeepFakes, FaceShifter, and ours. FaceShifter and our method generate obviously better results than other methods. For FaceShifter, it generates wrong expressions in row 1 (fierce) and row 2 (fear), of which expressions are from source faces. Besides, FaceShifter keeps the beard of target faces in rows 3 and 4, which makes the swapped faces close to their target faces. In the last three rows, FaceShifter fails to swap faces. However, our method successfully preserves identity information from source images.

For experiments on FaceForensics++, HieRFE and FTM are sequentially trained ten epochs in total on CelebA, CelebA-HQ, FFHQ, and the auxiliary data. As for experiments on CelebA-HQ, HieRFE and FTM are sequentially trained seventeen epochs in total on FFHQ and the auxiliary data. As for training time, it takes about five days on three Tesla V100 GPUs.

4.2. Experiments on FaceForensics++

Qualitative Comparison: As FaceForensics++ contains images generated by three face swapping methods: FaceSwap, DeepFakes, and FaceShifter, we extract frames of the same index from this dataset and compare them with the proposed MegaFS.

As shown in Fig.5, FaceShifter and our method generate more visually pleasant results than other methods. For example, FaceSwap and DeepFakes suffer from blending inconsistency, distortions, and artifacts. For FaceShifter, the disentanglement of identity information from other attributes is sub-optimal because of its fixed identity encoder. In Fig.5, FaceShifter generates unnatural expressions in the first and second rows, which seems to keep unnecessary expressions from the source images. FaceShifter also fails to transfer identity information from source faces to target faces by incorrectly maintaining the beard from target faces in rows 3 and 4. Additionally as shown in rows 4, 5, and 6, FaceShifter tends to keep excessive attributes from target images, which makes the swapped faces similar to their target faces.

Quantitative Comparison: In order to make a fair comparison with other methods quantitatively, we follow the experiment settings introduced in FaceShifter [28].

Firstly, ten frames per original video are evenly sampled and processed by MTCNN [57], resulting in 10,000 aligned faces. Then, aligned faces are manually checked in case of incorrect detections. After data cleaning, all corresponding frames in manipulated videos are extracted for testing. However, as FaceForensics++ is not designed for face recognition, some videos display repeated identities. For example, videos numbered 043 and 343 show Vladimir Putin, and videos of 179, 183 and 826 show the same person Barack Obama. Thus, we manually categorize all videos into 885 identities. ID retrieval is measured as the top-1 matching rate of the swapped faces and their corresponding identities from source faces, serving to measure the identity preservation ability of different face swapping methods. As for pose and expression errors, an open-sourced pose estimator [47] and a 3D facial model [13] are used to extract pose and expression vectors. Then ℓ_2 distances between swapped faces and the corresponding target faces are measured and recorded in Tab.1.

Method	ID retrieval ↑	pose ↓	expression ↓
DeepFakes [11]	88.39	4.64	3.33
FaceSwap [15]	72.69	2.58	2.89
Face2Face [53]	-	2.68	**2.09**
Neural Textures [52]	-	**2.21**	**1.64**
FaceShifter [28]	**90.68**	**2.55**	2.82
Ours	**90.83**	2.64	2.96

Table 1. Quantitative comparison results on FaceForensics++. The best two results are shown in red and blue respectively. ↑ means higher is better, and ↓ means lower is better.

As DeepFakes, FaceSwap, and FaceShifter are face swapping methods, while Face2Face and Neural Textures are face reenactment methods, different evaluation criterions should be considered. We report ID retrieval, pose error, and expression error for face swapping methods and neglect ID retrieval for face reenactment methods. As shown in Tab.1, our method achieves the highest ID retrieval thanks to the hierarchical representation for faces. However, our method performs inferior to FaceShifter and reenactment methods in terms of pose and expression errors. Aside from face reenactment methods are mainly designed to control facial movements and expression deformations while neglecting to swap the identity information, two possible reasons hide behind. Firstly, our method is trained on only 500,000 images, which is much less than 2,700,000 images used to train FaceShifter. Besides, the training set for FaceShifter contains VGGFace [41], which contains more pose and expression variations compared with CelebA-HQ and FFHQ. Secondly, StyleGAN2 is trained on FFHQ, which is proved to have data bias [49]. Consequently, StyleGAN2 tends to generate smiling faces.

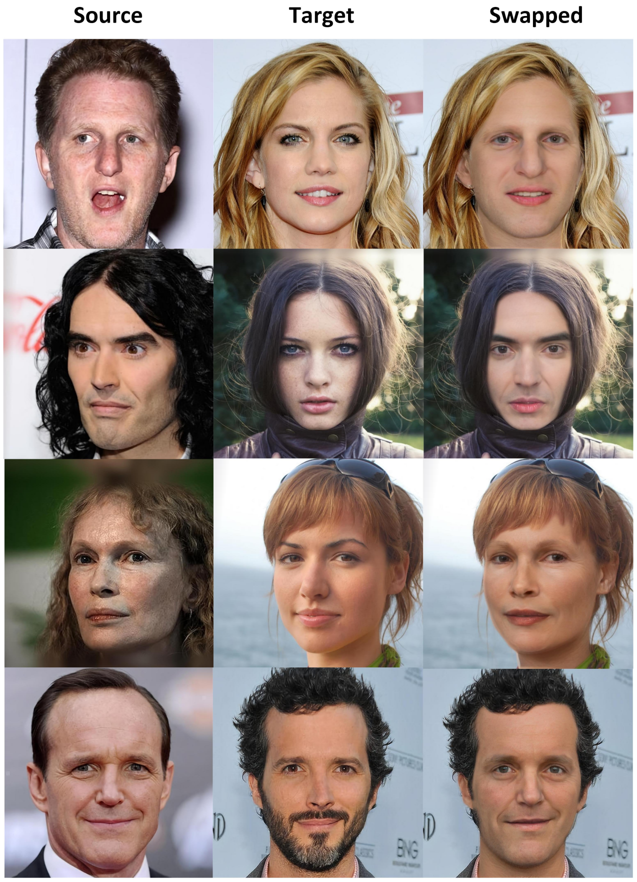

Figure 6. Face swapping results on CelabA-HQ. Images from right to left are source image x_s which provides the identity, target image x_t that offers the attributes, and the swapped face image y_{s2t}. All images are in 1024^2.

4.3. Experiments on CelebA-HQ

Qualitative Result: One superiority of our method is that it can achieve megapixel level face swapping. As shown in Fig.6, faces can be swapped across various expressions and poses. The swapped faces faithfully keep wrinkles, iris colors, eyebrow and nose shapes from source faces. To the best of our knowledge, no other methods can swap faces at the resolution of 1024^2 except for [34]. However, [34] needs to train different decoders for different identities, so it is not compared in this section.

Method	ID similarity ↑	pose ↓	expression ↓	FID ↓
Ours	0.5014	3.58	2.87	10.16

Table 2. Quantitative experimental results on CelebA-HQ. We report ID similarity, pose error, and expression error to demonstrate the megapixel level face swapping performance of the proposed MegaFS. FID is also reported as the similarity between the 300,000 swapped face images and CelebA-HQ dataset.

Quantitative Result: To quantify the capability of the proposed MegaFS on swapping megapixel face images, we randomly swapped 300,000 pairs of face images in CelebA-HQ for testing. For the reason that ID retrieval calculation between 30,000 original faces and 300,000 swapped faces requires *Nine Billion* times of matching, we report cosine similarity of swapped faces and the corresponding source faces using cosface as ID similarity to release the computational burden. Also, both pose error and expression error are measured under the same settings as experimented in subsection 4.2. In addition, Fréchet Inception Distance (FID) is reported to quantify the similarity of the 300,000 swapped face images to CelebA-HQ dataset. The results are summarized in Tab.2 as the baseline for future research.

4.4. Ablation Study

In this section, we conduct ablation experiments on CelebA-HQ to evaluate the effectiveness of the key components in the proposed MegaFS.

4.4.1 The Choise of Latent Space

In this part, we will verify the superiority of the extended latent space $\mathcal{W}^{++}$ over $\mathcal{W}^{+}$. We trained another neural network, which has the same network structure as HieRFE, to project facial images into latent space $\mathcal{W}^{+}$.

Latent Space	LPIPS ↓	MSE ↓	failure rate ↓
$\mathcal{W}^{+}$	0.2486	0.0672	1.28%
$\mathcal{W}^{++}$	**0.2335**	**0.0563**	**0.65%**

Table 3. Quantitative comparison results of latent space $\mathcal{W}^{+}$ and $\mathcal{W}^{++}$ using GAN Inversion metrics. LPIPS ℓ_2 distance and image level MSE are measured to quantify the information preservation capabilities of $\mathcal{W}^{+}$ and $\mathcal{W}^{++}$. The robustness is indicated by the failure rate of facial reconstruction. For reported metrics, HieRFE outperforms its counterpart trained on $\mathcal{W}^{+}$.

For illustrating the information preservation ability, two widely used metrics in GAN Inversion, LPIPS ℓ_2 distance and image level MSE, are reported in Tab.3. Besides, the percentage of unsuccessful reconstructions is defined as the failure rate to quantify the robustness of two inversion models. From the reported results, we can conclude that HieRFE outperforms its counterpart trained on $\mathcal{W}^{+}$ for better information preservation ability as well as the robustness. As to the controllability, we use the same evaluation criterions in subsection 4.3 to evaluate different latent space. The quantitative results are shown in Tab.4, suggesting that $\mathcal{W}^{++}$ is better than $\mathcal{W}^{+}$ in terms of ID similarity, pose and expression preservation ability.

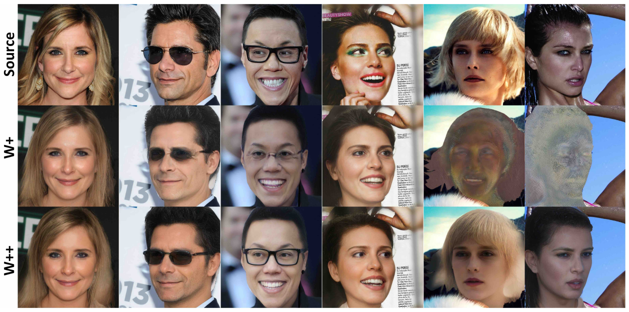

Figure 7. Qualitative comparison results of reconstructed images from latent space $\mathcal{W}^{+}$ and $\mathcal{W}^{++}$. From top to bottom: source images, reconstructed images from $\mathcal{W}^{+}$ and $\mathcal{W}^{++}$. HieRFE and its counterpart perform well in easy cases (the first column), but the latter fails to recast sunglasses, glasses, eye gazes, and faces under complex lighting conditions (from the second to the last columns).

Latent Space	ID similarity ↑	pose ↓	expression ↓
$\mathcal{W}^{+}$	0.5438	4.0640	1.7467
$\mathcal{W}^{++}$	**0.5816**	**3.8179**	**1.6489**

Table 4. Quantitative comparison results of latent space $\mathcal{W}^{+}$ and $\mathcal{W}^{++}$ using face swapping metrics. HieRFE trained on $\mathcal{W}^{++}$ beats its counterpart trained on $\mathcal{W}^{+}$ in terms of ID similarity, pose error, and expression error.

The qualitative results of two inversion models are displayed in Fig.7. HieRFE and its counterpart can reconstruct easy cases well. However, the latter fails to recast sunglasses, eyeglasses, eye gazes, and faces under complex lighting conditions. Thus, the latent space $\mathcal{W}^{++}$ is verified to be better than $\mathcal{W}^{+}$ for both face reconstruction and face swapping tasks in all terms of information preservation ability, robustness, and controllability.

4.4.2 The Design of Latent Code Manipulator

As StyleGAN2 has a layer-wise representation [9, 22, 23], it is heuristically feasible to manipulate latent codes by any network that operates on vectors. However, we argue that the design of the latent code manipulator needs to consider the applicability on vectorized information exchanging. To this end, we make use of $[C_t, L_t^{high}, \boldsymbol{L_s^{high}}]$ instead of $[C_t, L_t^{high}, \boldsymbol{L_{s2t}}]$ for generation, named as Latent Code Replacement (LCR), to envisage the functionality of C, L^{low} and L^{high}. Afterwards, we follow the previous

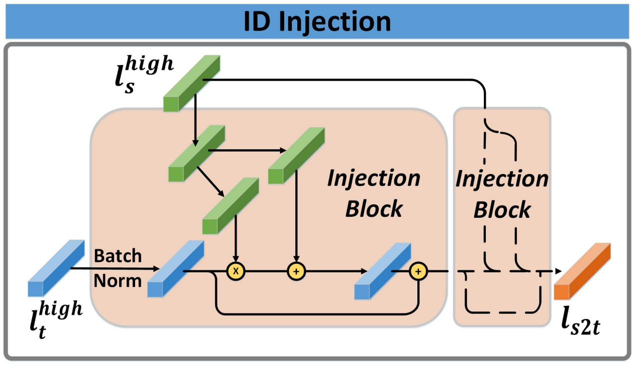

Figure 8. The design of ID Injection follows the SPADE ResBlk [40], with convolutional layers for 2D input subsitituted by linear layers, indicated by black arrows, for vectors.

Latent Control	ID similarity ↑	pose ↓	exp ↓	FID ↓
LCR	0.3997	5.04	3.43	**9.64**
ID Injection	0.4447	3.67	**2.82**	10.32
FTM (Ours)	**0.5014**	**3.58**	2.87	10.16

Table 5. Quantitative comparison results of different latent code manipulation methods ("exp" represents expression error). FTM achieves the best ID similarity and pose preservation results and makes a decent balance among expression error and FID.

method [40] to inject identity information into latent codes. This design is detailed in Fig.8, namely ID Injection. Both designs are compared to the proposed FTM.

For a fair comparison, other two sets of 300,000 swapped face images are generated by adopting LCR and ID Injection respectively. The quantitative results are summarized in Tab.5 and the qualitative comparison is shown in Fig.9. LCR achieves the best FID since it keeps excessive semantic information from source images. However, this is not favorable for face swapping since information from L_t^{high} is lost. As shown in the third column of Fig.9, LCR can swap faces while ignoring target attributes such as skin color and eye state. Thus, we can safely conclude that identity information, to a large extent, is encoded in L^{high}. Thus, C_s and L_s^{low} are discarded in the proposed pipeline.

Based on this observation, ID Injection and FTM are proposed to process only on L^{high}. For ID Injection, it has the lowest expression error as it reserves topology information from target images at the cost of other semantic parts from source images, such as identity information stated in Tab.5, and facial details as shown in the fourth column of Fig.9. Among them, FTM achieves the highest ID similarity level, lowest pose error, and makes a decent balance in terms of expression error, FID, and visual pleasantness. Thus, the proposed FTM shows to be better than the other two latent code manipulators for face swapping.

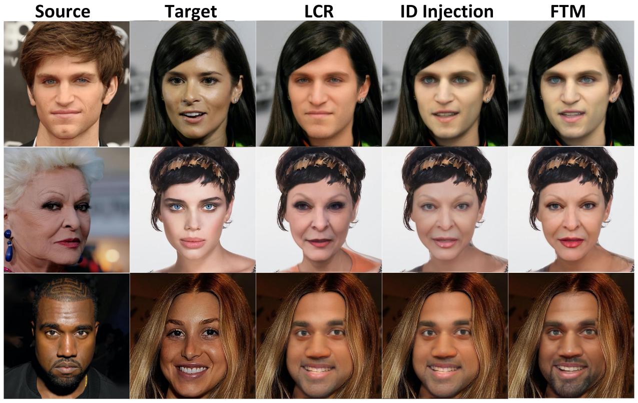

Figure 9. Qualitative comparison results of different latent code manipulation methods: LCR, ID Injection, and FTM. LCR keeps skin color and eye state from the source image x_s as shown in the first two rows. For ID Injection, attributes are dominated by the target image x_t. For example, the red lip in row 2 and beard in the last row from source images are neglected. FTM achieves the best balance among the three latent code manipulation methods.

5. Conclusion

In this paper, we have analyzed three unsettled key issues in previous works for high resolution face swapping and proposed a general face swapping pipeline named MegaFS to resolve these difficulties in a three-stage procedure. HieRFE in the first stage projects faces into hierarchical represSentaions in an extended latent space $\mathcal{W}^{++}$ for complete facial information deposit. FTM in the second stage transfers the identity from a source image to the target by a nonlinear trajectory without explicit feature disentanglement. Finally, StyleGAN2 is used to synthesize the swapped face and avoid unstable adversarial training. The modular design of MegaFS requires little GPU memory with a negligible performance cost and it performs comparatively when compared to other state-of-the-art face swapping methods at the resolution of 256^2. Besides, to the best of our knowledge, MegaFS is the first method that can conduct one shot face swapping on megapixels. Finally, based on MegaFS, the first megapixel level face swapping database is built and released to the public for future research of forgery detection and face swapping.

6. Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2020AAA0140002, in part by the Science and Technology Development Fund of Macau SAR under grant 0015/2019/AKP, in part by the Natural Science Foundation of China under Grant 62076240, Grant 61721004, Grant U1836217, and in part by the Macao Youth Scholars Program under Grant AM201913.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4432–4441, 2019. 2

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8296–8305, 2020. 2

[3] Rameen Abdal, Peihao Zhu, Niloy Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *arXiv preprint arXiv:2008.02401*, 2020. 2, 3

[4] Oleg Alexander, Mike Rogers, William Lambeth, Matt Chiang, and Paul Debevec. Creating a photoreal digital actor: The digital emily project. In *Conference for Visual Media Production*, pages 176–187, 2009. 1

[5] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6713–6722, 2018. 2

[6] Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. Exchanging faces in images. In *Computer Graphics Forum*, volume 23, pages 669–676, 2004. 1

[7] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, pages 187–194, 1999. 2, 3

[8] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1511–1520, 2017. 2

[9] Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. Editing in style: Uncovering the local semantics of gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5771–5780, 2020. 7

[10] Antonia Creswell and Anil Anthony Bharath. Inverting the generator of a generative adversarial network. *IEEE Transactions on Neural Networks and Learning Systems*, 30(7):1967–1974, 2018. 2

[11] DeepFakes. https://github.com/ondyari/FaceForensics/tree/master/dataset/DeepFakes. Accessed: 2020-10-08. 1, 2, 6

[12] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 4

[13] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 6

[14] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019. 2

[15] FaceSwap. https://github.com/ondyari/FaceForensics/tree/master/dataset/FaceSwapKowalski. Accessed: 2020-10-08. 6

[16] Shanyan Guan, Ying Tai, Bingbing Ni, Feida Zhu, Feiyue Huang, and Xiaokang Yang. Collaborative learning for faster stylegan embedding. *arXiv preprint arXiv:2007.01758*, 2020. 2

[17] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546*, 2020. 2, 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[19] Ali Jahanian, Lucy Chai, and Phillip Isola. On the "steerability" of generative adversarial networks. In *International Conference on Learning Representations*, 2020. 2

[20] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2886–2895, 2020. 2

[21] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 5

[22] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 5, 7

[23] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 2, 7

[24] Ira Kemelmacher-Shlizerman. Transfiguring portraits. *ACM Transactions on Graphics*, 35(4):1–8, 2016. 1

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015. 5

[26] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018. 2

[27] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3677–3685, 2017. 1, 2

[28] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020. 2, 6

[29] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020. 2

[30] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 3

[31] Zachary C Lipton and Subarna Tripathi. Precise recovery of latent vectors from generative adversarial networks. *International Conference on Learning Representations Workshops*, 2017. 2

[32] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015. 5

[33] Saleh Mosaddegh, Loic Simon, and Frédéric Jurie. Photorealistic face de-identification by aggregating donors' face components. In *Asian Conference on Computer Vision*, pages 159–174, 2014. 1

[34] J Naruniec, L Helminger, C Schroers, and RM Weber. High-resolution neural face swapping for visual effects. In *Computer Graphics Forum*, volume 39, pages 173–184, 2020. 1, 2, 6

[35] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. Fsnet: An identity-aware generative model for image-based face swapping. In *Asian Conference on Computer Vision*, pages 117–132, 2018. 2

[36] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7184–7193, 2019. 2

[37] Yotam Nitzan, Amit Bermano, Yangyan Li, and Daniel Cohen-Or. Disentangling in latent space by harnessing a pretrained generator. *arXiv preprint arXiv:2005.07728*, 2020. 2, 3

[38] Yotam Nitzan, Amit Bermano, Yangyan Li, and Daniel Cohen-Or. Face identity disentanglement via latent spacemapping. *arXiv preprint arXiv:2005.07728*, 2020. 2

[39] Kyle Olszewski, Zimo Li, Chao Yang, Yi Zhou, Ronald Yu, Zeng Huang, Sitao Xiang, Shunsuke Saito, Pushmeet Kohli, and Hao Li. Realistic dynamic facial textures from a single image using gans. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5429–5438, 2017. 2

[40] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 8

[41] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference*, pages 41.1–41.12, September 2015. 6

[42] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016. 2

[43] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *International Conference on Learning Representations*, 2016. 2, 3

[44] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. *arXiv preprint arXiv:2008.00951*, 2020. 4

[45] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018. 2

[46] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2, 5

[47] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition Workshops*, pages 2074–2083, 2018. 6

[48] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020. 2, 3

[49] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *arXiv preprint arXiv:2005.09635*, 2020. 6

[50] Ayush Tewari, Mohamed Elgharib, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, Christian Theobalt, et al. Pie: Portrait image embedding for semantic control. *arXiv preprint arXiv:2009.09485*, 2020. 2, 3

[51] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020. 2, 3

[52] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics*, 38(4):1–12, 2019. 6

[53] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016. 6

[54] Jingdong Wang, Sun ke, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp:1–1, 04 2020. 4

[55] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8261–8265, 2019. 2

[56] Jiangning Zhang, Xianfang Zeng, Yusu Pan, Yong Liu, Yu Ding, and Changjie Fan. Faceswapnet: Landmark guided many-to-many face reenactment. *arXiv preprint arXiv:1905.11805*, 2019. 2

[57] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016. 6

[58] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 4

[59] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. Indomain gan inversion for real image editing. *arXiv preprint arXiv:2004.00049*, 2020. 2