High-resolution Face Swapping via Latent Semantics Disentanglement

Yangyang Xu[1,3]   Bailin Deng[2]   Junle Wang[3]   Yanqing Jing[3]   Jia Pan[4]   Shengfeng He[1*]

[1]South China University of Technology   [2]Cardiff University

[3]Tencent   [4]The University of Hong Kong

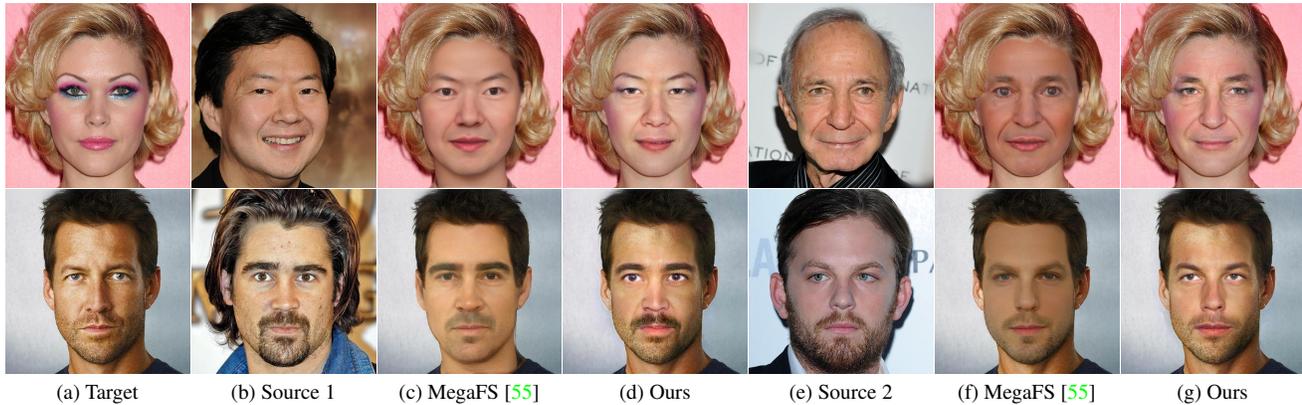| (a) Target | (b) Source 1 | (c) MegaFS [55] | (d) Ours | (e) Source 2 | (f) MegaFS [55] | (g) Ours |

Figure 1. We propose a high-resolution (1024×1024) face swapping method by disentangling the latent semantics of a pre-trained GAN model. Compared with the existing high-resolution method MegaFS [55], our model can preserve target attributes (*e.g.*, skin color and illumination), meanwhile transferring local details and identity from source faces.

Abstract

We present a novel high-resolution face swapping method using the inherent prior knowledge of a pre-trained GAN model. Although previous research can leverage generative priors to produce high-resolution results, their quality can suffer from the entangled semantics of the latent space. We explicitly disentangle the latent semantics by utilizing the progressive nature of the generator, deriving structure attributes from the shallow layers and appearance attributes from the deeper ones. Identity and pose information within the structure attributes are further separated by introducing a landmark-driven structure transfer latent direction. The disentangled latent code produces rich generative features that incorporate feature blending to produce a plausible swapping result. We further extend our method to video face swapping by enforcing two spatio-temporal constraints on the latent space and the image space. Extensive experiments demonstrate that the proposed method outperforms state-of-the-art image/video face swapping methods in terms of hallucination quality and consistency. Code can be found at: https://github.com/cnnlstm/FSLSD_HiRes.

1. Introduction

Face swapping aims at transferring the identity from a source face image to a target face image, while preserving attributes in the target image such as facial expression, head pose, illumination and background. It has received extensive attention in the computer vision and graphics community due to its wide range of potential applications, such as computer games, special effects and privacy protection [13, 22, 41, 42].

The main challenge for face swapping is to identify the highly entangled target facial attributes and source identity information for a natural-looking swapping. Early works such as [8] replace the pixels of face regions and rely on similarities between the source and the target in pose and illumination. 3D-based approaches [14, 28, 37] fit a 3D model to the faces and can handle large pose variation, although the fitting can be largely influenced by the environment and thus unstable. Other works introduce the generative adversarial networks (GANs) for hallucinating target attributes [24, 26, 35] due to its strong generative capability.

Though much progress has been made, many existing GAN-based approaches do not work well on high-resolution faces, due to the compressed representation of end-to-end frameworks [7, 26, 34], the instability of adversarial training [5], and the limitation in GPU memory size. Recently,

*Corresponding author (hesfe@scut.edu.cn).

Zhu *et al*. [55] utilize the inherent prior knowledge of a pre-trained high-resolution GAN model and propose MegaFS for high-resolution face swapping in the latent space of Style-GAN [20,21]. It learns to assemble the inverted latent codes of the source and target images, and directly feed the fused code to a StyleGAN generator to produce the swapped result. However, since the identity and attributes are highly entangled in the latent space, assembling two latent codes without explicit guidance cannot guarantee the transfer of source identity and the preservation of target attributes simultaneously. Moreover, fine details embedded in the latent codes are easily diluted after the assembly, hence, their swapped face tends to have a blurred appearance with some loss of fine details (see Fig. 1c and Fig. 1f for examples). To obtain disentangled semantics in the latent space, we argue that face features should be transferred in a class-specific manner. Intuitively, the structure attributes of a face, such as the facial shape, pose, and expression, should be treated differently than the appearance attributes such as illumination and skin tones. The swapped face should retain the source identity while hallucinating the appearance attributes of the target image. Such separate treatments would require proper disentanglement between the structure and appearance features.

In this paper, we delve into the latent semantics of Style-GAN. StyleGAN is a noise-to-image, coarse-to-fine generation process, and we leverage its progressive nature to disentangle key factors in swapping such as pose, expression, and appearance. Given the inverted source latent code, we decouple the structure (pose and expression) attributes from identity by deriving a structure transfer latent direction. The direction is determined by source and target landmarks, and serves as a latent space operation that transfers the structure. On the other hand, the appearance attributes are controlled in the deeper layers. Therefore, we regroup the target's appearance codes with the structure-transferred source code in the deep layers. In this way, the integrated latent code retains the identity attribute from the source, while having the appearance and structure of the target. This disentangled latent code is fed to the StyleGAN generator to produce generative features. This rich prior knowledge is aggregated with the target features in all scales, effectively eliminating the noticeable blending artifacts.

Furthermore, we extend our model to video face swapping, where we generate a swapped face video from a source face image and a target face video. Since directly applying our model to each video frame can lead to incoherence artifacts, we enforce two spatio-temporal constraints on the disentangled structure and appearance semantics. First, we require the inter-frame structural changes of the swapped faces to be consistent with those in the target faces. This is achieved by enforcing similarity between the swapped and target faces in terms of the latent offsets in the shallow layers (representing structural changes). Secondly, we adopt a linear assumption for the changes of image content between neighboring frames in the output video. These two constraints effectively ensure inter-frame coherence. Extensive experiments on several benchmarks for face swapping methods demonstrate superior performance against state-of-the-arts. As far as we are aware, this is the first feasible solution for high-resolution video face swapping.

In summary, our contributions are three-fold:

- We propose a novel framework for high-resolution face swapping. We disentangle the latent semantics of a pre-trained StyleGAN, enabling the transfer of source identity while preserving the appearance and structure of the target.
- We tailor two novel constraints to enforce the coherence of swapped face videos, including a code trajectory constraint that limits the offset between the latent codes of neighboring frames, and a flow trajectory constraint that works in the RGB space to guarantee the video smoothness.
- Experiments on several datasets demonstrate state-of-the-art results from our method, which can potentially serve as new high-resolution test cases for face forgery detection.

2. Related Work

Face Swapping. Face swapping has been an active research topic. Early works such as [8] can only handle subjects of the same pose. 3D-based approaches fit a 3D template to the source and target faces, to better handle large pose variations [14, 28, 37]. Face2Face [45] fits a bilinear face model to the source and target faces to transfer the expression, with additional steps to synthesize realistic mouth appearance. However, 3D-based methods cannot handle appearance attributes like illumination and style. Recently, many works introduce GANs for face swapping. RS-GAN [35] handles face and hair regions in the latent spaces, and conducts swapping by replacing latent representations. Similarly, FSNet [34] encodes the face region and the non-face region to different latent codes. FSGAN [36] performs face reenactment and swapping simultaneously. Recently, Faceshifter [26] encodes the source identity and target attributes separately. However, none of them can swap face at the high resolution due to the unstable adversarial training, or the compressed representation of end-to-end frameworks. Recently, Naruniec *et al*. [33] swap faces at high resolutions but their model is subject-specific. MegaFS [55] utilizes the prior knowledge of pre-trained StyleGAN [20, 21]. It inverts both source and target faces to the latent space, then designs a face transfer block to assemble the latent codes. However, the identity and attributes are entangled in the latent space, and assembling two latent codes without explicit guidance is insufficient. On the contrary, we propose to transfer attributes in a class-specific manner.

Generative Prior. Generative models show great potential in synthesizing high-quality images [10, 16, 19–21]. It has

Figure 2. The pipeline of our disentangled high-resolution face swapping. We transfer different levels of attributes by three modules. We first invert both the source and target faces to the $W+$ latent space of the pre-trained StyleGAN G by the pSp encoder, then learn a structure transfer direction by encoding the source and target landmarks. The appearance attributes are transferred by swapping the latter part of two latent codes. To eliminate the blending boundary, target features are aggregated with the generative features in a multi-scale manner.

been shown that the latent space of a pre-trained GAN encodes rich semantic information. As a result, many works utilize the prior knowledge hidden in the generative model for semantic editing, image translation, super-resolution, and so on [4, 17, 39, 40, 50, 52]. In particular, Gu *et al*. [17] propose a GAN inversion approach for image colorization and semantic editing, while some other works [11, 31, 54] use the pre-trained StyleGAN for image super-resolution. Recently, generative prior has also been introduced for face swapping. Besides MegaFS [55], Nitzan *et al*. [38] transfer attributes from one face to another in the latent space by a fully-connected network. Different from those works, we transfer the target attributes to the source in a more elaborated way. We decompose the attributes into structure attributes and appearance attributes, and take full usage of disentanglement and editability of the generative model.

3. Method

3.1. Overview

Given two high-resolution face images, we aim to construct a face image with the identity of the source face x_s and the attributes of the target face x_t such as the pose, expression, illumination and background. To do so, we first construct a side-output swapped face using a StyleGAN generator, by blending the structure attributes of the source and the target in the latent space while reusing the appearance attributes of the target (Sec. 3.2). To further transfer the

background of the target face, we use an encoder to generate multi-resolution features from the target image, and blend them with the corresponding features from the upsampling blocks of the StyleGAN generator. The blended features are fed into a decoder to synthesis the final swapped face image (Sec. 3.3). An overview of our pipeline is provided in Fig. 2. The networks are trained using a set of loss functions that enforce desirable properties such as the preservation of source identity and target appearance (Sec. 3.4). We also extend our approach to video face swapping, using additional spatio-temporal constraints to enforce coherence (Sec. 3.5).

3.2. Class-Specific Attributes Transfer

A face image typically contains different classes of attributes. For example, the pose and expression are related to the face structure, while the lighting and colors are related to the appearance. Previous face swapping techniques, such as FaceShifter [26] and MegaFS [55], do not explicitly distinguish different classes of attributes when transferring them to the output image. We instead argue that it is beneficial to transfer structure and appearance attributes separately: the structure attributes of the output can be jointly determined from their counterparts in the source and target images, to obtain the same pose and expression as the target face while retaining the identity of the source face. Meanwhile, the appearance attributes of the target face can be directly reused to achieve similar facial appearance in the output.

To this end, we note that StyleGAN [20, 21], a state-of-

the-art generative model, provides a suitable representation for such separate treatments. In particular, to use the Style-GAN generator, it is common practice to encode an image in an extended latent space $W+$ where a latent code consists of multiple high-dimensional vectors, one for each input layer of StyleGAN [3, 40]. As noted in [40, 51], different input layers of StyleGAN correspond to different levels of details. Thus we treat the first K vectors of the latent code, which correspond to the shallow layers, as the encoding for the structure attributes. The remaining vectors, which correspond to the deeper layers, are used for appearance. For 1024×1024 images, the latent code consists of 18 different 512-dimensional vectors, and we follow [40] and choose the first 7 vectors as the structure part. With such disentanglement, the structure part and the appearance part of the latent code can be transferred separately. Specifically, we first use the pre-trained pSp encoder [40] to invert the source face x_s and the target face x_t to obtain their $W+$ latent codes $w_s = (g_s, h_s)$ and $w_t = (g_t, h_t)$ respectively, where g_s, g_t are the structure parts and h_s, h_t are the appearance parts. To construct the structure attributes of a swapped face that has the same identity of the source with the pose and expression of the target, we note that they should be derived from the source structure attributes with modifications that take into account the pose and expression difference between the source and the target. Therefore, we compute the structure part $\widehat{g}_s$ of the latent code for a swapped face by applying a structure transfer latent direction $\overrightarrow{n}$ that is derived from the source and target structures:

$$\widehat{g}_s = g_s + \overrightarrow{n}. \tag{1}$$

To derive $\overrightarrow{n}$, we note that the face structure can be indicated by the facial landmarks. Thus we train a landmark encoder $E_{le}(\cdot, \cdot)$ that generates $\overrightarrow{n}$ from the heat-map encodings of the source landmarks l_s and the target landmarks l_t:

$$\overrightarrow{n} = E_{le}(l_s, l_t).$$

To transfer the appearance attributes from the target face to the swapped face, we directly reuse the appearance part h_t of the target latent code. It is then re-integrated with $\widehat{g}_s$ to form a latent code for the swapped face:

$$\widehat{w}_s = \text{Cat}(\widehat{g}_s, h_t), \tag{2}$$

where $\text{Cat}(\cdot, \cdot)$ denotes the concatenation operator. The code is fed into a pre-trained StyleGAN generator to obtain a side-output swapped face y_s.

3.3. Background Transfer

For face swapping applications, the background of the target face also needs to be retained in the output image. This often cannot be guaranteed for the swapped face y_s

computed in Sec. 3.2. A common solution is to apply Poisson Blending as the post-process that blends the swapped inner-face with the target image. However, this can lead to unnatural appearance around the boundary of the inner-face.

To address this issue, we discard the side-output face y_s, but retain the features $F_s = \{f_s^0, f_s^1, ...f_s^N\}$ produced by each upsampling block of the StyleGAN generator from the latent code $\widehat{w}_s$ in Eq. (2). We note that F_s can be regarded as a representation for different levels of details from the side-output face image. Accordingly, we apply an encoder E_t to the target face x_t such that its layers generate corresponding features $F_t = \{f_t^0, f_t^1, ...f_s^N\}$, where f_t^i is of the same dimension as f_s^i and represent the details of the target face image at the same resolution. We then aggregate each pair of corresponding features (f_s^i, f_t^i) by replacing the components of f_t^i for the inner-face region with their counterparts in f_s^i. All aggregated features are fed into a decoder to produce the final face image y_f, which can be written as:

$$y_f = \text{Dec}(F_t, F_s, m_t),$$

where $\text{Dec}(\cdot, \cdot, \cdot)$ denotes the decoder, and m_t is the inner-face mask for the target face image. In this way, the decoder transfers the background in multi-level features, which not only eliminates the need for explicit background blending but also enables the code $\widehat{w}_s$ to focus on the facial region and facilitates attributes transfer.

3.4. Loss Functions

We tailor several losses that are applied on the side-output swapped face y_s or the final face y_f for efficient attributes transfer, as explained in the following. We also introduce a style-preservation loss on the final output to narrow the style gaps between the swapped one and the target.

Adversarial loss. We utilize an adversarial loss for the distribution alignment between the final swapped faces and the real faces. In particular, we align the final face y_f with the target face with the following loss:

$$L_{adv} = \mathbb{E}_{y_f \sim P_{Y_f}} [-\log(D_f(y_f))],$$

where P_{Y_f} denotes the distribution of the final faces. The discriminator D_f that distinguishes the final face from the real target face x_t is trained with a loss

$$L_{D_f} = \mathbb{E}_{y_f \sim P_{Y_f}} [-\log(1-D_f(y_f)] + \mathbb{E}_{x_t \sim P_{X_t}} [-\log(D_f(x_t)],$$

where P_{X_t} denote the distribution of real faces.

Identity-preservation loss. To preserve the identity of the source face, we introduce an identity-preservation loss for the final face y_f with the source face x_s:

$$L_{id} = 1 - \cos(\Phi_{id}(y_f), \Phi_{id}(x_s)),$$

where $\Phi_{id}(\cdot)$ is the pre-trained ArcFace network for face recognition [15], and $\cos(\cdot, \cdot)$ denotes the cosine similarity.

Landmark-alignment Loss. Since we use facial landmarks to represent structure attributes, we introduce the following loss to align the landmarks of the side-output swapped face y_s, the final face y_f, and the target face x_t:

$$L_{lmk} = \|E_{lmk}(y_s) - E_{lmk}(x_t)\|_2 + \|E_{lmk}(y_f) - E_{lmk}(x_t)\|_2,$$

where $E_{lmk}(\cdot)$ is a pre-trained landmark estimator [47] and $\|\cdot\|_2$ denotes the ℓ_2-norm.

Reconstruction Loss. Intuitively, if the source face x_s and the target face x_t are the same image, the network should reconstruct this image for both the side-output swapped face y_s and the final face y_f. Hence, we follow pSp [40] and use pixel-wise similarity and perceptual similarity to define a reconstruction loss that penalize the deviation between y_s, y_f and x_t when $x_s = x_t$:

$$L_{rec} = \begin{cases} \|y_f - x_t\|_2 + \alpha\|F(y_f) - F(x_t)\|_2 \\ \quad + \|y_s - x_t\|_2 + \alpha\|F(y_s) - F(x_t)\|_2, & \text{if } x_s = x_t, \\ 0, & \text{otherwise,} \end{cases}$$

where $F(\cdot)$ is the perceptual feature extractor, and α is a weight that balances the pixel-wise similarity and perceptual similarity terms. We set $\alpha = 0.8$ in our experiment.

Style-transfer Loss. As mentioned in Sec. 3.2, we transfer the appearance attributes in the latent space. However, if the difference between the styles of the source and target faces is too large, simple latent code replacement may not reduce the style difference effectively. Inspired by BeautyGAN [27], we create a guidance face image $\text{HM}(y_f, x_t)$ through histogram mapping, and align the final face y_f with the guidance face via the following loss:

$$L_{st} = \|y_f - \text{HM}(y_f, x_t)\|_2.$$

Compared to simple latent code replacement, this provides a stronger guidance for the appearance attributes transfer.

Final Objective. The final loss function for training our model is a weighted combination of above losses:

$$L_{total} = \lambda_1 L_{adv} + \lambda_2 L_{id} + \lambda_3 L_{lmk} + \lambda_4 L_{rec} + \lambda_5 L_{st}, \quad (3)$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ are weights for the loss terms.

3.5. Video Face Swapping

Our method can be extended to video face swapping. Given a source face x_s and a sequence of target faces with M consecutive frames $S_t = \{x_t^0, x_t^1, ..., x_t^{M-1}\}$, we would like to obtain a swapped face sequence $Y = \{y^0, y^1, ..., y^{M-1}\}$. Most of the existing works apply the image-based face swapping methods to each video frame separately, which can lead to incoherent results between neighboring frames and cause

artifacts such as flickering. Such artifacts are particularly noticeable in high resolutions. To address this issue, we need to enforce consistency between neighboring frames in terms of both the structure and the appearance, such that these attributes vary smoothly across the frames. Existing temporal consistency works [9, 25] only take the appearance consistency into account and cannot be used directly for face swapping. We propose two spatio-temporal constraints for the latent space and the image space respectively, to achieve consistency in both the structure and the appearance.

Code Trajectory Constraint. Since the structure attributes of the target frames vary smoothly, we can enforce structure consistency for the swapped face frames by requiring them to have similar changes in structure attributes as the target. To this end, we note that the offset between the structure parts of the latent codes in two neighboring frames can be can be regarded as an indication of the change in the structure attributes. Therefore, we use the following loss to enforce similar trajectories of the structure codes between the target video and the output video in the latent space:

$$L_{ct} = \sum_{k=1}^{M} \left\| (\hat{g}_s^k - \hat{g}_s^{k-1}) - (g_t^k - g_t^{k-1}) \right\|_2,$$

where g_t^k denotes the structure part of the latent code for the target frame x_t^k, and $\hat{g}_s^k$ denotes the structure code of the swapped face obtained from x_s and x_t^k using Eq. (1).

Flow Trajectory Constraint. For appearance consistency, we follow the local linear model [30, 32] that assumes uniform changes between neighboring frames [1]. Specifically, we denote the optical flow from the frame y_f^i to a nearby frame y_f^j in the swapped face sequence as:

$$f_{i \Rightarrow j} = \Phi(y_f^i, y_f^j),$$

where $\Phi(\cdot, \cdot)$ is the pre-trained PWC-Net [44] for the flow prediction. From the local linear assumption, the dense correspondence between two neighboring frames in the swapped face sequence can be approximated by interpolating the forward flow and backward flow across the two frames, which leads to the following loss:

$$L_{ft} = \sum_k \|(f_{k \Rightarrow k+2} + f_{k+2 \Rightarrow k})/2 - f_{k \Rightarrow k+1}\|_2.$$

We note that L_{ft} is effectively the ℓ_1/ℓ_2 norm for the temporal Laplacians of the optical flows. This promotes sequences that have small Laplacians at most frames while allowing for large Laplacians at some frames [6], which allows for motion sequences that are piecewise smooth temporally.

---

[1] Here we follow the local linear assumption for simplicity, we can also follow the acceleration-aware assumption [29, 49] for better approximating real-world scene motion.

Figure 3. Qualitative comparison of face swapping on the resolution of 1024×1024. We can see that MegaFS always present the blurry faces with a distinct boundary with the background. Besides, they cannot preserve the source identity effectively during the swapping process. Meanwhile, when the appearance attributes between source and target are huge, they cannot transfer the target to the output. In opposite, our method can transfer the structure and appearance attributes as desired with identity-preserving. Zoomed in for the best view.

4. Experiments

4.1. Implementation Details

We utilize the StyleGAN2 generator [21] pre-trained on the FFHQ dataset [20] with the resolution of 1024×1024. The pSp encoder is also pre-trained on this dataset. We use a modification of pSp as our landmark (more details of the structure can be seen in the supplementary materials). We use the Adam optimizer [23] to train the model with a learning rate of 1×10^{-4}, and the exponential decay rates for the 1_{st} and 2_{nd} moment estimate being $\beta_1 = 0.9$ and $\beta_2 = 0.999$ respectively, and $\epsilon = 1 \times 10^{-8}$. The batch size is set to be 8 and the model is trained with 500,000 iterations. We empirically set the weights in Eq. (3) as $\lambda_1 = 1$, $\lambda_2 = 2$, $\lambda_3 = 0.1$, $\lambda_4 = 2$ and $\lambda_5 = 0.2$. We implement our method in Pytorch with four Tesla V100 GPUs. It takes about two days to train the whole model.

4.2. Datasets

Datasets. We evaluate our model on three datasets:
- **CelebA-HQ** contains 30,000 celebrities faces with a resolution of 1024×1024 [19]. Due to its high quality, it has been widely used in many face editing works.
- **FaceForensics++** consists of 1,000 original talking videos downloaded from YouTube and manipulated with 5 face swapping methods [42]. This dataset serves as a benchmark in many face swapping works.

Evaluation Metrics. We use several metrics in our quantitative experiments. The *ID retrieval rate*, measured by the top-1 identity matching rate based on the cosine similarity, indicates the identity preservation ability. For some experiments, we follow MegaFS [55] and compute the *ID similarity* instead, which is the cosine similarity between swapped faces and their corresponding sources using CosFace [46], to reduce the computational cost. The *pose error* and the *expression error* are the ℓ_2 distance between the pose and expression feature vectors respectively using pre-trained estimators [12,43] on the swapped and target faces, which indicates the ability to transfer structure attributes. The *Fréchet Inception Distance* (FID) [18] computes the Wasserstein-2 distance between the distribution of real faces and swapped images, which measures the image quality of swapped faces.

4.3. Comparison on CelebA-HQ Dataset

Qualitative Comparison. We first conduct experiments on the high-resolution CelebA-HQ dataset [19]. We compare our method with MegaFS [55] that swaps faces at 1024×1024 resolution, with qualitative comparison results shown in Fig. 3. We can see the faces produced by MegaFS tend to have a blurry appearance without vivid details. This is because they perform face swapping solely on the low-dimensional latent codes without the explicit disentanglement, which easily dilutes the latent encoding of details.

Besides, there is a notable boundary between facial re-

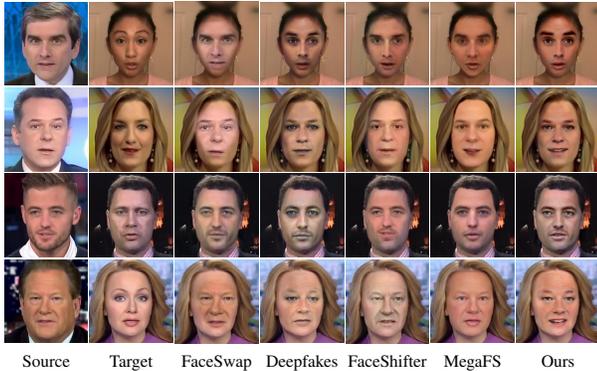| Source | Target | FaceSwap | Deepfakes | FaceShifter | MegaFS | Ours |

Figure 4. Qualitative comparison with MegaFS and other low-resolution non-GAN priors face swapping methods.

gions and the background in their results. In comparison, our approach utilizes multi-resolution spatial features from the StyleGAN generator and aggregates them with the background features from the target decoder, which helps to retain high-quality facial details. Moreover, MegaFS does not transfer the target attributes to the output effectively, especially when there is a large semantic gap between the source and the target. For example, in the second row of Fig. 3, MegaFS cannot transfer the illumination or styles from targets, due to the large difference between the source and the target in these attributes. In comparison, thanks to our disentangled attributes transfer, our results maintain the target attributes more effectively. For the results on the left of the fourth row in Fig. 3, we can also see that MegaFS does not preserve the source identity, where it wrongly enlarges the eyes compared to the source. This is because the identity and the attributes are highly entangled in the latent space, and the lack of explicitly disentanglement in MegaFS can lead to unsatisfactory results. On the contrary, we transfer different levels of attributes separately, which helps to maintain the source identity and the target attributes.

Quantitative Comparison. We also perform quantitative comparison on the CelebA-HQ dataset. We follow MegaFS and conduct a comparison on 300,000 swapped faces for a fair comparison. Tab. 1 shows the average values of evaluation metrics for each method. Our method has a stronger ability both on identity preservation and attributes transfer. Furthermore, our result has a lower FID, which indicates that our swapped faces are more realistic than those of MegaFS.

4.4. Comparison on FaceForensics++ Dataset

To allow comparing with other face swapping methods that can be only applied to low-resolution images, we further evaluate our method on the Face-Forensics++ dataset. Fig. 4 presents qualitative comparison results against MegaFS as well as three non-GAN prior based and low-resolution methods: FaceSwap [2], Deepfakes [1], and FaceShifter [26]. We can see that FaceSwap and MegaFS can lead to notice-

Table 1. Quantitative evaluation of face swapping on CelebA-HQ dataset with four metrics. ↓ denotes the lower the better and vice versa, the best results are marked in **bold**.

Methods	ID Simi.↑	Pose Err.↓	Exp. Err.↓	FID↓
MegaFS [55]	0.5214	3.498	2.95	11.645
Ours	**0.5688**	**2.997**	**2.74**	**9.987**

Table 2. Quantitative evaluation of face swapping on FaceForensics++ dataset with four metrics. ↓ denotes the lower the better and the better and vice versa, the best results are marked in **bold**.

Methods	ID Retri.(%)↑	Pose Err.↓	Exp. Err.↓
FaceSwap [2]	72.69	2.58	2.89
Deepfakes [1]	88.39	4.64	3.33
FaceShifter [26]	90.68	2.55	2.82
MegaFS [55]	**90.83**	2.64	2.92
Ours	90.05	**2.46**	**2.79**

able blending boundaries or artifacts, while our method can eliminate them due to our background transfer. However, the quality of our swapped results is not that high as high-resolution face swapping in the Fig. 3. This is potentially due to domain gap: the styleGAN generator and the pSp encoder used in our method are both pre-trained on high-resolution data, while the majority of the data from FaceForensics++ are of lower resolutions. The large domain gap between low- and high-resolution datasets can potentially decrease the performance of our method.

We also perform quantitative evaluation on this dataset. In particular, we follow MegaFS [55] that samples 10 frames from each video evenly then processed by MTCNN [53]. After filtering the repeated identities, we obtain 885 videos with 88500 frames in total. Tab. 2 shows the average values of evaluation metrics for different methods. We can see that our method preserves the target poses and expressions more effectively, thanks to our structure attributes transfer that takes the landmarks as input and provides a strong guidance signal to the final swapped faces. Meanwhile, our model is outperformed by FaceShifter and MegaFS on ID retrieval rate. We conjecture that this is due to the domain gap between the low- and high-resolution images as mentioned above, which makes our inversion model less effective in preserving the identity information from low-resolution images.

4.5. Ablation Study

In this section, we perform ablation study and use the CelebA-HQ dataset to evaluate the effectiveness of our disentangled approach in transferring the attributes. We use MegaFS as a baseline, since they swap faces solely on two latent codes without disentanglement between attributes. We further include three variants of our approach with modifications on the modules and the loss functions. For the first variant (Var.1), we keep the appearance part of source code
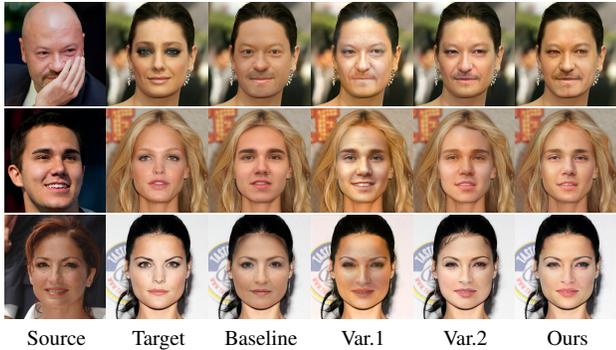
Figure 5. Qualitative comparison of different variants. Each variant cannot transfer the target attributes effectively due to the entangled semantic transfer.
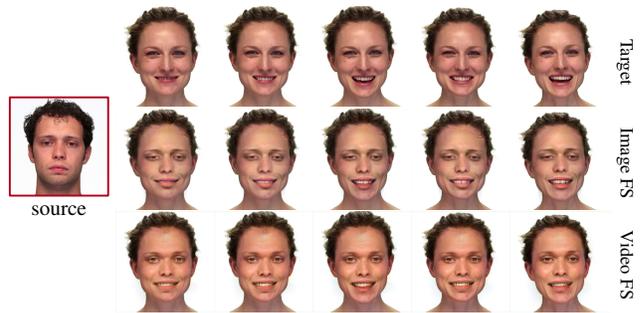

Figure 6. Examples of video face swapping. The figure in the red box is the source face, and the top row shows the target video frames. Applying face swapping on each frame individually leads to the incoherent structure and appearance (middle row), while our trajectory constraints inhibit such incoherence (bottom row).

w_s unchanged for producing the side-output face and discard the style-transfer loss and the latent code swapping operation for appearances transfer, *i.e.*, there is no explicit transfer guidance that contributes to the appearance attributes. For the second variant (Var.2), we discard our background transfer module and directly blend the side-output face with the target face image as the final output. The qualitative comparison between different variants and the baseline is shown in Fig. 5. We can see that the MegaFS baseline leads to blurry faces and sharp boundaries. As mentioned previously, this is due to the highly entangled identity and attributes in the latent code, which can lead to information loss during the transfer. The first variant presents the unnatural styles, which verifies the necessity of our appearance attributes transfer that provides explicit appearance guidance in the swapping process. The second variant presents a notable blending boundary between the swapped face region and the background (see the 2_{nd} and 3_{rd} samples), due to the lack of the background transfer module which fuses multi-resolution features from the source and the target for a natural blending. Meanwhile, our target encoder-decoder structure also can tolerance the structure difference between source and targets, and produce the final plausible results. Thanks to our disentangled attributes transfer, our final swapped faces present the successful semantic transfer from the targets, with the source identity well preserved.

4.6. Face Swapping on High-resolution Videos

Fig. 6 shows qualitative results of our method on high-resolution videos. Applying our method individually to each frame leads to incoherence between neighboring frames, while our code and flow trajectory constraints improve the coherence and visual quality considerably.

5. Conclusion and Discussions

We present a novel high-resolution face swapping based on the inherent prior knowledge of pre-trained StyleGAN. We categorize the attributes into structure and appearance ones, and transfer them separately in the disentangled latent space. We propose a landmark encoder that predicts a latent direction for the structure attributes transfer. The StyleGAN generative features resulting from the transferred latent code are aggregated with the multi-resolution features from a target image encoder to transfer the background information and generate a high-quality result. We further extend the method to video face swapping by enforcing two spatio-temporal constraints. Extensive experiments demonstrate the superiority of our disentangled attributes transfer in terms of hallucination quality and consistency.

Limitations. Since we transfer the attributes in the latent space of StyleGAN, the quality of the result relies heavily on the GAN inversion method. In particular, if the inversion does not produce faithful latent codes for the source and target faces, the result is not guaranteed to preserve the identity of the source. Our method transfers all the target appearance attributes to the result image and does not support selective transfer of appearance from both the source and the target. Such fine-grained control can be beneficial for some applications such as content creation, but would require further disentanglement between different categories of appearance attributes. This can be an avenue for further research.

Potential Negative Impact. Although not the purpose of this work, realistic face swapping can potentially be misused for deepfakes-related applications. The risk can be mitigated by gated release of the model and by forgery detection methods that are able to spot CNN-generated images [48]. In addition, our method can be potentially used to generate new high-resolution test cases for benchmarking and further developing forgery detection techniques [42].

References

[1] Deepfakes. https://github.com/ondyari/FaceForensics/tree/master/dataset/DeepFakes. 7

[2] Faceswap. https://github.com/ondyari/FaceForensics/tree/master/dataset/FaceSwapKowalski. 7

[3] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, pages 4432–4441, 2019. 4

[4] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM TOG*, 40(3):1–21, 2021. 3

[5] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *ICLR*, 2017. 1

[6] Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et al. Optimization with sparsity-inducing penalties. *Found. Trends Mach. Learn.*, 4(1):1–106, 2012. 5

[7] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *CVPR*, pages 6713–6722, 2018. 1

[8] Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree K Nayar. Face swapping: automatically replacing faces in photographs. In *ACM TOG*, pages 1–8. 2008. 1, 2

[9] Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. Blind video temporal consistency. *ACM TOG*, 34(6):1–9, 2015. 5

[10] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2018. 3

[11] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *CVPR*, pages 14245–14254, 2021. 3

[12] Bindita Chaudhuri, Noranart Vesdapunt, and Baoyuan Wang. Joint face detection and facial motion retargeting for multiple faces. In *CVPR*, pages 9719–9728, 2019. 6

[13] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *ACM Multimedia*, pages 2003–2011, 2020. 1

[14] Yi-Ting Cheng, Virginia Tzeng, Yu Liang, Chuan-Chang Wang, Bing-Yu Chen, Yung-Yu Chuang, and Ming Ouhyoung. 3d-model-based face replacement in video. In *SIGGRAPH*, pages 1–1. 2009. 1, 2

[15] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 5

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 27, 2014. 3

[17] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *CVPR*, pages 3012–3021, 2020. 3

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, pages 6626–6637, 2017. 6

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 3, 6

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 2, 3, 4, 6

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 2, 3, 4, 6

[22] Ira Kemelmacher-Shlizerman. Transfiguring portraits. *ACM TOG*, 35(4):1–8, 2016. 1

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[24] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *ICCV*, pages 3677–3685, 2017. 1

[25] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, pages 170–185, 2018. 5

[26] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *CVPR*, pages 5074–5083, 2020. 1, 2, 3, 7

[27] Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *ACM Multimedia*, pages 645–653, 2018. 5

[28] Yuan Lin, Shengjin Wang, Qian Lin, and Feng Tang. Face swapping under large pose variations: A 3D model based approach. In *ICME*, pages 333–338, 2012. 1, 2

[29] Yihao Liu, Liangbin Xie, Li Siyao, Wenxiu Sun, Yu Qiao, and Chao Dong. Enhanced quadratic video interpolation. In *ECCV*, pages 41–56, 2020. 5

[30] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *ICCV*, pages 4463–4471, 2017. 5

[31] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, pages 2437–2445, 2020. 3

[32] Simone Meyer, Oliver Wang, Henning Zimmer, Max Grosse, and Alexander Sorkine-Hornung. Phase-based frame interpolation for video. In *CVPR*, pages 1410–1418, 2015. 5

[33] Jacek Naruniec, Leonhard Helminger, Christopher Schroers, and Romann M Weber. High-resolution neural face swapping for visual effects. In *Computer Graphics Forum*, volume 39, pages 173–184, 2020. 2

[34] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. FSNEt: An identity-aware generative model for image-based face swapping. In *ACCV*, pages 117–132, 2018. 1, 2

[35] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. Rsgan: face swapping and editing using face and hair representation in latent spaces. In *SIGGRAPH*, pages 1–2. 2018. 1, 2

[36] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *ICCV*, pages 7184–7193, 2019. 2

[37] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and face perception. In *FG*, pages 98–105, 2018. 1, 2

[38] Yotam Nitzan, Amit Bermano, Yangyan Li, and Daniel Cohen-Or. Face identity disentanglement via latent space mapping. *ACM TOG*, 39(6):1–14, 2020. 3

[39] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. In *ECCV*, pages 262–277, 2020. 3

[40] 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. In *CVPR*, pages 2287–2296, 2021. 3, 4, 5

[41] Arun Ross and Asem Othman. Visual cryptography for biometric privacy. *IEEE TIFS*, 6(1):70–81, 2010. 1

[42] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *CVPR*, pages 1–11, 2019. 1, 6, 8

[43] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *CVPR Workshop*, pages 2074–2083, 2018. 6

[44] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, pages 8934–8943, 2018. 5

[45] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, pages 2387–2395, 2016. 2

[46] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, pages 5265–5274, 2018. 6

[47] Jingdong Wang, Ke Sun, 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 TPAMI*, 43(10):3349–3364, 2021. 5

[48] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. CNN-generated images are surprisingly easy to spot... for now. In *CVPR*, pages 8695–8704, 2020. 8

[49] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. *NeurIPS*, 32, 2019. 5

[50] Yangyang Xu, Yong Du, Wenpeng Xiao, Xuemiao Xu, and Shengfeng He. From continuity to editability: Inverting gans with consecutive images. In *ICCV*, pages 13910–13918, 2021. 3

[51] Ceyuan Yang, Yujun Shen, and Bolei Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis. *IJCV*, pages 1–16, 2021. 4

[52] Huiting Yang, Liangyu Chai, Qiang Wen, Shuang Zhao, Zixun Sun, and Shengfeng He. Discovering interpretable latent space directions of gans beyond binary attributes. In *CVPR*, pages 12177–12185, 2021. 3

[53] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *Signal Process Letter*, 23(10):1499–1503, 2016. 7

[54] Yang Zhou, Yangyang Xu, Yong Du, Qiang Wen, and Shengfeng He. Pro-pulse: Learning progressive encoders of latent semantics in gans for photo upsampling. *IEEE TIP*, 31:1230–1242, 2022. 3

[55] Yuhao Zhu, Qi Li, Jian Wang, Cheng-Zhong Xu, and Zhenan Sun. One shot face swapping on megapixels. In *CVPR*, pages 4834–4844, 2021. 1, 2, 3, 6, 7