FuseAnyPart: Diffusion-Driven Facial Parts Swapping via Multiple Reference Images

Zheng Yu[*] [†]
Shanghai Jiao Tong University & Alibaba Group
cs-yuzheng@sjtu.edu.cn

Yaohua Wang [*] [‡]
Alibaba Group
xiachen.wyh@alibaba-inc.com

Siying Cui [†]
Peking University & Alibaba Group
cuisiying.csy@alibaba-inc.com

Aixi Zhang
Alibaba Group
aixi.zhax@alibaba-inc.com

Wei-Long Zheng
Shanghai Jiao Tong University
weilong@sjtu.edu.cn

Senzhang Wang
Central South University
szwang@csu.edu.cn

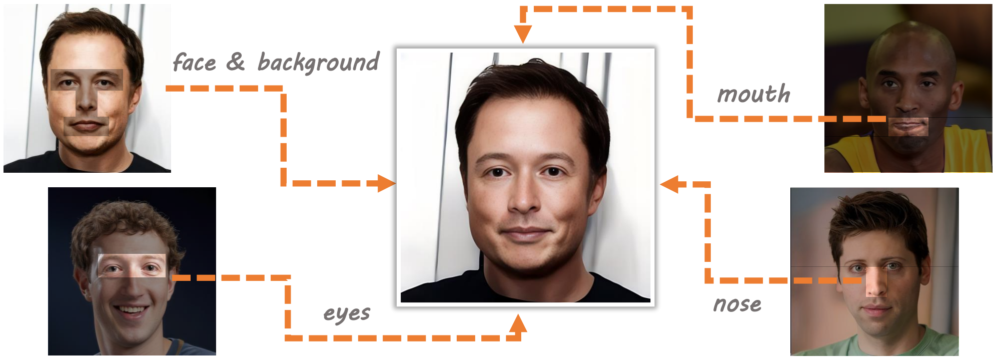

Figure 1: Results of facial parts swapping using the proposed FuseAnyPart at 512×512 resolution. The swapped face (central image) is generated by fusing the original face (top-left image) with three facial part reference images (bottom-left, top-right, bottom-right). Notably, FuseAnyPart can seamlessly blend facial parts from multiple reference images with significant differences in appearance, producing high-fidelity and natural-looking swapped faces.

Abstract

Facial parts swapping aims to selectively transfer regions of interest from the source image onto the target image while maintaining the rest of the target image unchanged. Most studies on face swapping designed specifically for full-face swapping, are either unable or significantly limited when it comes to swapping individual facial parts, which hinders fine-grained and customized character designs. However, designing such an approach specifically for facial parts swapping is challenged by a reasonable multiple reference feature fusion, which needs to be both efficient and effective. To overcome this challenge, FuseAnyPart is proposed to facilitate the seamless "fuse-any-part" customization of the face. In FuseAnyPart,

[*] Equal contribution.
[†] Work done during the internship at Alibaba Group.
[‡] Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

facial parts from different people are assembled into a complete face in latent space within the Mask-based Fusion Module. Subsequently, the consolidated feature is dispatched to the Addition-based Injection Module for fusion within the UNet of the diffusion model to create novel characters. Extensive experiments qualitatively and quantitatively validate the superiority and robustness of FuseAnyPart. Source codes are available at `https://github.com/Thomas-wyh/FuseAnyPart`.

1 Introduction

Imagine a person who possesses the face of Elon Musk, the eyes of Mark Zuckerberg, the nose of Sam Altman, and the mouth of Kobe Bryant. What would the visual composite of such a person look like? Currently, this dream can be realized through the facial parts swapping technology as shown in Fig. 1. Different from traditional face swapping, which is coarse and typically replaces the entire face at once, facial parts swapping aims to transfer the individual facial components, e.g., nose, mouth or eyes from varied sources onto the target image while maintaining the rest of the target image unchanged. A growing interest in facial parts swapping technology has emerged due to its broad applications such as innovative character creation, popular entertainment, privacy protection and beyond [29, 20].

Most of the studies so far have primarily focused on full-face swapping, and can be roughly divided into GAN-based and diffusion-based approaches. The GAN-based methods [7, 34, 27, 15, 12] usually perform face swapping by extracting the identity feature from the source images and then injecting them into generative adversarial networks [6]. Nevertheless, the GAN-based techniques may not succeed in completely transferring the identity features, especially when there is a significant difference in shape between the source and the target. In addition, the GAN-based methods often involve an array of losses about image fidelity, identity, and facial attributes to guide the training, which increases the complexity of the training process. On the other hand, diffusion-based models [23, 24, 36, 13, 31, 26, 33] have demonstrated a powerful capability in generating images with high resolution and complex scenes. Some efforts try to swap face [38, 18] through diffusion models, achieving pleasant results.

However, the aforementioned methods, designed for full-face swapping, are either unable or significantly limited when it comes to swapping individual facial parts. If one needs to replace a facial part, it is only possible to swap the entire face, rather than independently swapping one or several facial parts individually, let alone the facial parts from different individuals. This high degree of coupling poses an inconvenience for users who seek more fine-grained and customized designs. Therefore, the focus on face-swapping shifts from an identity-centric to an attribute-level perspective.

The primary challenge in facial parts swapping lies in the fusion mechanism. Popular face-swapping techniques [15, 38, 36, 18] perform the fusion of source and target images in the latent space for harmonious generated images. Therefore, the feature fusion mechanism becomes critical in affecting the quality of the generated images. In the facial parts swapping task, the number of source images increases from one to multiple, further complicating the fusion process and making this issue more prominent. Previous methods [36, 13] utilize adapters implemented by a cross-attention mechanism to fuse reference information into the UNet of diffusion models. However, as the cross-attention is initially designed for multi-modal tasks [24], like text and image, it may be sub-optimal for facial parts swapping due to the difficulty in aligning fine-grained facial region features. What is more, the inclusion of multiple references increases computational needs, thus efficient fusion is essential.

To tackle these challenges, an innovative diffusion-driven approach dubbed FuseAnyPart is proposed to facilitate the seamless "fuse-any-part" customization of faces. In FuseAnyPart, a facial image is initially detected by an open-set detector to derive its various facial part masks. Then an image encoder extracts the facial part features based on the facial image and the aforementioned masks. Subsequently, these facial part features are assembled according to the masks within the Mask-based Fusion Module included in FuseAnyPart to generate a complete face in latent space. After this step, the cohesive feature is forwarded to the Addition-based Injection Module proposed by FuseAnyPart for fusion within the UNet of the diffusion model. The Addition-based Injection Module adds a minimal amount of parameters yet is highly effective in preserving the positional information and fine details of the image features, which demonstrates obvious superiority compared to the conventional cross-attention mechanism. During the training stage, FuseAnyPart is trained by **reconstructing** a facial image conditioned on different facial parts, inspired by [15, 38]. In the inference stage, facial parts from images of various people can be fed into FuseAnyPart to create a novel character.

Overall, the contributions can be summarized as follows: (1) To the best of our knowledge, FuseAnyPart is the first diffusion-based work specifically designed for facial parts swapping, which is capable of simultaneous, multi-source and fine-grained facial parts swapping. (2) The proposed Masked-based Fusion Module in FuseAnyPart allows dynamically aggregating specific parts from different faces in latent space. Then, the Addition-based Injection Module of FuseAnyPart injects this conditional information into UNet, which is more effective and efficient than the conventional cross-attention-based adapter methods. (3) Extensive experiments validate the superiority of FuseAnyPart. Ablation studies confirm the soundness of our design choices and the robustness of our proposed approach.

2 Related Work

Image Generation with Multiple References. InstantBooth [26], InstanID [31], and IDAdapter [3] use the average feature of all reference images, which contributes to improving generation quality. And photoMaker [13] generates an ID embedding by stacking embeddings from multiple ID reference images, which results in improved ID representation. Moreover, it can create a mixed ID embedding by controlling the proportion of identity images within the input reference image collection. Although the aforementioned methods introduce a multiple reference image mechanism, the role of these reference images is similar to that of providing a single reference image, affecting only the generation of a specific subject. FastComposer [33], on the other hand, achieves the generation of multiple subjects by injecting different reference image features into distinct word embeddings. Currently, image generation using multiple reference images remains an area ripe for exploration, such as generating human faces with multiple inputs.

Facial Parts Swapping. In recent years, region-controllable face swapping has emerged as a fascinating subfield within the broader domain of facial manipulation and generation. This technology enables precise control over specific regions of a face in an image, allowing the exchange or modification of features such as the eyes, nose, or mouth, while maintaining the integrity of the original image's context. The E4S model [15] achieves precise editing results by manipulating masks of specific regions, such as the eyes or lips, using a reference image as a guide. Meanwhile, Diffswap [38] is a technique that selectively determines the regions to swap by constructing masks that cover varying facial areas. Although these methods are capable of transferring specific facial regions from a source image to a target image, the results often exhibit unnatural boundaries. Furthermore, to achieve the replacement of facial features from multiple reference images onto a single face, multiple iterations are typically required, complicating the process. This iterative approach can be time-consuming and may not consistently produce seamless, natural-looking results, indicating that there is still room for improvement in the field of region-controllable face swapping technology.

3 Method

3.1 Preliminary

StableDiffusion. Our model is based on the StableDiffusion [24] model, which progresses the diffusion process in low-dimensional latent space with a pre-trained autoencoder. Using a latent representation, StableDiffusion can maintain the essential features and structure of the data while requiring fewer steps and less time to generate high-quality samples. First, the variational autoencoder compresses the input image x to a latent representation z_0, which is gradually added Gaussian noise with a fixed Markov chain of T steps. Let $z_t = \alpha_t z_0 + \sigma_t \epsilon$ be the noised data at the t-th timestep, where α_t, σ_t are predefined functions of t and $\epsilon \in \mathcal{N}(0, I)$, UNet ϵ_θ is responsible for the denoising process by predicting the noise ϵ. The denoising process can be conditioned by the additional condition C. The training objective is to minimize the ELBO of the denoising process, which is defined as:

$$\mathcal{L} = \mathbb{E}_{z_t,t,C,\epsilon}[||\epsilon - \epsilon_\theta(z_t, t, C)||_2^2]. \tag{1}$$

During inference time, UNet gradually predicts the noise $\epsilon_\theta(x_t, t)$ and recovers the initial latent representation z_0 from random noise $z_T \in \mathcal{N}(0, I)$. Finally, the image is generated by mapping z_0 back to pixel space with the variational autoencoder.

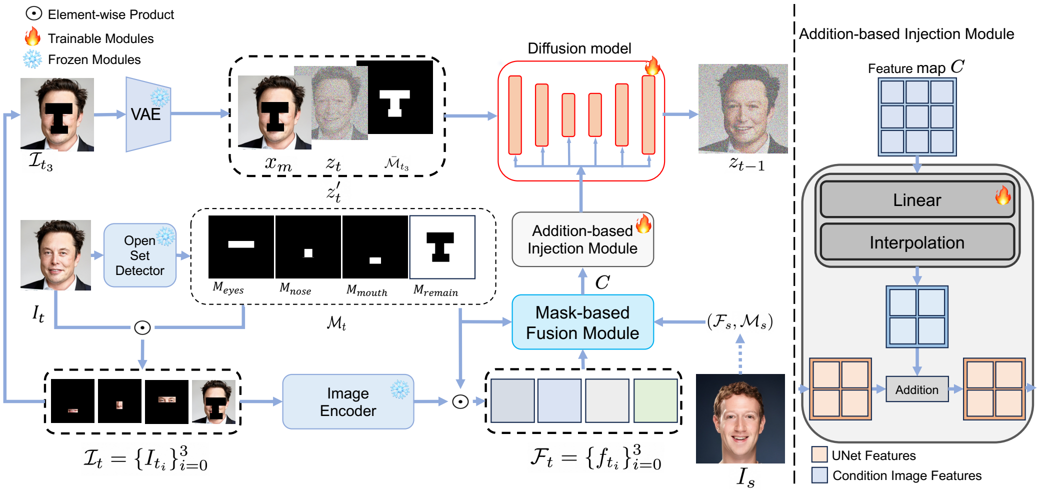

Figure 2: Illustration of FuseAnyPart. The process begins with an open-set detector identifying a facial image to obtain various facial part masks. Following this, an image encoder uses these masks and the facial image to derive the corresponding facial part feature. These facial part features and masks are then fed into the Mask-based Fusion Module to piece together a complete face in latent space. Subsequently, the consolidated feature is dispatched to the Addition-based Injection Module for fusion within the UNet of the diffusion model.

Image Prompt Adapter. Image prompt adapter [36] is an innovative approach to incorporate image features into the generation process without model fine-tuning for each new concept. This approach addresses the challenge faced by previous methods, which struggled to effectively extract and utilize detailed image features from image prompts. Similarly to text prompts, image prompts can also condition the generative process through the cross-attention mechanism. Specifically, a decoupled cross-attention strategy is employed in which an additional cross-attention layer is added to every original cross-attention layer to inject image features. The output of this new cross-attention layer can be articulated as

$$Z^{new} = \textbf{Attention}(Q, K, V) + \lambda \cdot \textbf{Attention}(Q, K^i, V^i), \tag{2}$$

where λ is weight factor, $\textbf{Attention}(Q, K^i, V^i)$ is cross-attention of the new added cross-attention layer, $K^i = c_i W_k^i$ and $V^i = c_i W_v^i$ are key and values matrices of the corresponding operation. c_i are the image features, and W_k^i and W_v^i are the relevant weight matrices.

3.2 Overview

The goal of facial parts swapping is to selectively transfer regions of interest such as the eyebrows, eyes, nose, or mouth from the source image onto the target image while maintaining the rest of the image unchanged. Some methods [15, 8] swap region feature obtained by mask pooling to facilitate facial parts swapping and use masks as structure guidance to maintain detail and coherence in the generated results. However, the applicability of this approach is limited in advanced StableDiffusion models utilizing cross-attention mechanisms, due to the difficulty in aligning fine-grained facial region features with the latent features in UNet. Consequently, it is necessary to fine-tune the entire SD model with a significant volume of data. Additionally, there is a notable paucity of methodologies that permit the incorporation of multiple source images to selectively transfer features from sources onto a target image in one step.

To address these problems, we map a face image into multiple non-overlapping region image features and perform mask-based fusion at the feature map level between the source and target image features. This fusion results in new facial image features, which are then integrated into the generation process through an Addition-based Injection Module.

3.3 Facial Feature Decomposition and Aggregation

For simplicity, we consider three regions for swapping including the eyes (including the eyebrows), nose, and mouth. We use an open-set detection model to get region masks M_{eyes}, M_{nose}, M_{mouth}, and the remaining region mask is $M_{remain} = \bar{M}_{eyes} \odot \bar{M}_{nose} \odot \bar{M}_{mouth}$. Let $\mathcal{M} = \{M_{eyes}, M_{nose}, M_{mouth}, M_{remain}\}$ be the region masks, a given face image I can be represented as the union of multiple regional images: $\mathcal{I} = \{I_i\}_{i=0}^{3}$, where $I_i = I \odot \mathcal{M}_i$.

Following most of the previous methods [2, 3, 37, 32], we utilize a pre-trained CLIP [22] image encoder ϕ to extract image representations from regional images. Contrary to the preceding work that harnesses the more abstract, global, and high-level features from the last layer, we use the uncompressed feature map from the penultimate layer, which retains greater spatial information and finer details. For a face image I, its feature representation can be decomposed into multiple components $\mathcal{F} = \{f_i\}_{i=0}^{3}$, where $f_i = \phi(I_i)$. Replacement is conducted at the feature map level, where a target image's features $\mathcal{F}_t = \{f_{t_i}\}_{i=0}^{3}$, with $i = 0, 1, 2$, are replaceable, corresponding respectively to eyes, nose, and mouth. Mathematically, the feature replacement between target image features $\mathcal{F}_t = \{f_{t_i}\}_{i=0}^{3}$ and source image features $\mathcal{F}_s = \{f_{s_i}\}_{i=0}^{3}$ is realized in the **Mask-based Fusion Module**, described as follows:

$$f'_{t_i} = \begin{cases} f_{t_i}, & R_i = True \\ f_{t_i} \odot \bar{\mathcal{M}}_{t_i} + \mathbf{G}(f_{s_i} \odot \mathcal{M}_{s_i}, \mathcal{M}_{t_i}, \mathcal{M}_{s_i}), & R_i \neq True \end{cases} \tag{3}$$

where $i \in \{0, 1, 2, 3\}$, R_i indicates whether the region feature f_{g_i} is replaced by f_{s_i}. $\mathbf{G}(I, m_1, m_2)$ is an interpolation function that resizes the region covered by mask m_1 in image I to fit the region of mask m_2. The resulting features $\mathcal{F}'_t = \{f'_{t_i}\}_{i=0}^{3}$ are aggregated and then fed into Multi-Layer Perceptron (**MLP**) to generate the final condition feature map, which is input into the UNet as $C = \textbf{MLP}(\sum_{i=0}^{3} \mathcal{F}'_{t_i} \odot \mathcal{M}_{t_i})$ providing the conditional information to guide the generative network.

3.4 Addition-based Injection

It has been discussed for a long time how to inject image features into the UNet using a cross-attention mechanism, and there are two principal methodologies. One is a direct method that feeds the concatenation of image features and text features into the layers of cross-attention. However, this method can be ineffective when image features are misaligned with textual features in the concatenation process. The other one, which has been widely adopted in numerous works, employs adapter modules with decoupled cross-attention [36]. Nevertheless, the cross mechanism still faces the challenge of inaccurate feature fusion because attention maps may fail to focus appropriately on the correct regions.

Thus, we propose the **Addition-based Injection Module** for integrating image features into the UNet which directly adds fine-grained image features to latent features within the UNet. Specifically, the output of the injected layer is described as follows:

$$Z' = Z + \lambda \cdot \textbf{Inter}(\textbf{Linear}(C)), \tag{4}$$

where Z is the latent feature within the UNet, C is the swapped face image feature map that servers as the condition information, $\textbf{Linear}(\cdot)$ is a linear layer, $\textbf{Inter}(\cdot)$ is a function which is capable of resizing C to match the dimensions of Z and λ is weight factor. It is feasible because the latent space features within the UNet also comprise a feature map that contains positional information, which has a corresponding spatial relationship with the condition feature map C. By adding fine-grained image features at their respective locations, we ensure alignment of the newly introduced features with the original latent space features in terms of position. Furthermore, the injection of such image features is not confined exclusively to the cross-attention layers; it can be integrated at any level within the UNet architecture. Compared to the cross-attention mechanism, this method reduces the number of added parameters and computational load while increasing the flexibility and controllability of feature injection. It enables the model to be fine-tuned with less training data, enhancing efficiency without compromising on the richness of the generated details.

3.5 Training and Inference

To preserve the regions of the face image that are not subject to replacement, FuseAnyPart follows the practice of [1]. Specifically, We concatenate the latent vector x_m, derived from the masked image $\mathcal{I}_{t_3}$, the associated mask $\bar{\mathcal{M}}_{t_3}$, and the noised latent vector z_t to form a new latent vector $z'_t = \text{Concat}(x_m, \bar{\mathcal{M}}_{t_3}, z_t)$, which is fed into a convolution layer for dimension adjustment. The feature vector $\hat{z}_t = \text{Conv}(z'_t)$ is subsequently introduced into the UNet, serving as the query. Our training objective is similar to the original StableDiffusion model, formulated as:

$$\mathcal{L} = \mathbb{E}_{z_t, t, x_m, \bar{\mathcal{M}}_{t_3}, C, \epsilon}[||\epsilon - \epsilon_\theta(z_t, t, x_m, \bar{\mathcal{M}}_{t_3}, C)||_2^2]. \tag{5}$$

During the inference phase, our model possesses the capability to transfer facial regions from multiple source images onto a target image. By deconstructing and reassembling facial image features, we can construct mixed facial features, facilitating flexible and controllable facial parts swapping.

4 Experiment

Dataset. We train our model on the CelebA-HQ [11] dataset. The CelebA-HQ dataset contains 30,000 high-resolution face images of celebrities widely used for face generation and face swapping tasks. This dataset has been pre-processed and aligned, and is available in three different resolutions. In our experiments, we use the 1024×1024 resolution. Our evaluation set is sampled from the FaceForensics++ [25] dataset, which contains 1,000 videos. We randomly sample 10 frames from each video and obtain 10,000 images. Then we use GPEN [35] for portrait enhancement and crop and align these images by landmarks to the resolution of 512×512. Additionally, we collected some high-quality face images from the internet intended for qualitative visual results.

Implementation Details. Our implementation is based on HuggingFace diffusers [30] library and we use StableDiffusion v1-5 [24] and OpenAI's clip-vit-large-path14 vison model [22]. We train our model on 16 NVIDIA A100 GPUs (80GB) with a batch size of 16 per GPU using the AdamW optimizer [16] with a constant learning rate of 1e-4 and weight decay of 0.01. During training, facial part reference images are randomly sampled from images with the same ID, and the target image is consistent with the face reference image. During the inference stage, we use the DDIM [28] sampler with 50 steps and set $\lambda = 1.0$. Since we do not use a text prompt, we set the text prompt to empty.

4.1 Qualitative Comparisons

We collected a series of high-quality celebrity images from the Internet to conduct qualitative experiments. To demonstrate the effectiveness of our approach, we have structured the qualitative experiments into three sets: fuse any part, multiple parts replacement, and multiple parts replacement with reference images in different styles.

4.1.1 Fuse Any Part

With grounding-dino [14], an open-set object detection model, our method is capable of extracting region-specific features of faces based on text, such as "eyes", "nose", and "mouth". Limited by the performance of grounding-dino, the "eyes" include the "eyebrows" in the subsequent chapters. In this paper, "reference image" is used to describe both the source and target images. The face reference image is the target, while the facial part reference image is the source.

We select eyes, nose, and mouth for attribute-level facial parts swapping. For single part replacement, we compare our method with StableDiffusion (SD)[24], IP-Adapter [36], FacePartsSwap [5], E4S [15] and Diffswap [38], and the results for eyes, nose and mouth are presented respectively in Fig. 3, Fig. 8 and Fig. 9. Since SD and IP-Adapter aren't designed for facial parts replacement, we cut out the desired attributes from the source image and overlaid them onto the target image, resulting in a pixel blend image, the inputs are source-target image pairs. By utilizing SD's image-to-image generation function with the denoising strength set to 0.5, we can reconstruct the spliced image. For the IP-Adapter, the spliced image serves as an image prompt, acting as additional information to condition the generation process. FacePartsSwap specifically focuses on exchanging facial parts and E4S and Diffswap are face swapping methods that can utilize different masks during the inference process to achieve partial facial region replacement.

4.1.2 Multiple Parts Replacement

Multi-attribute replacement differs from swapping the entire facial information onto a target image, as it only involves replacing certain attributes within their corresponding regions, and the number of

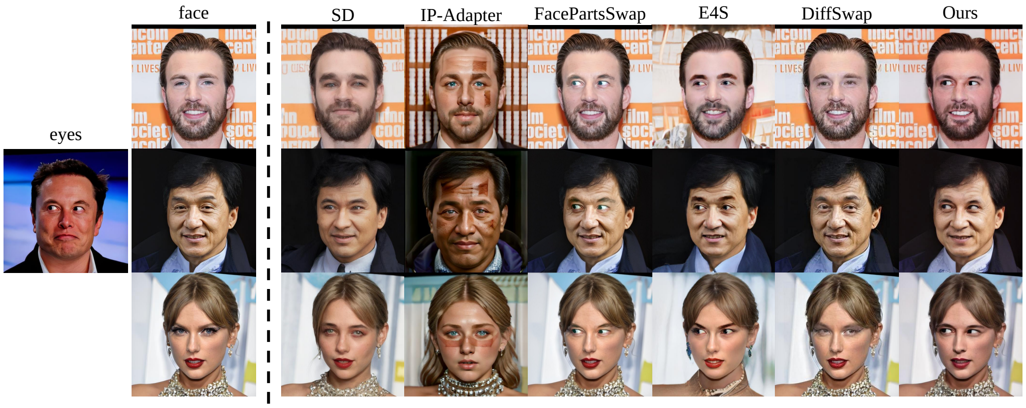

Figure 3: **Qualitative comparison of eyes swapping.** Our method produces high-fidelity results that maintain the consistency of facial features while ensuring a natural appearance.

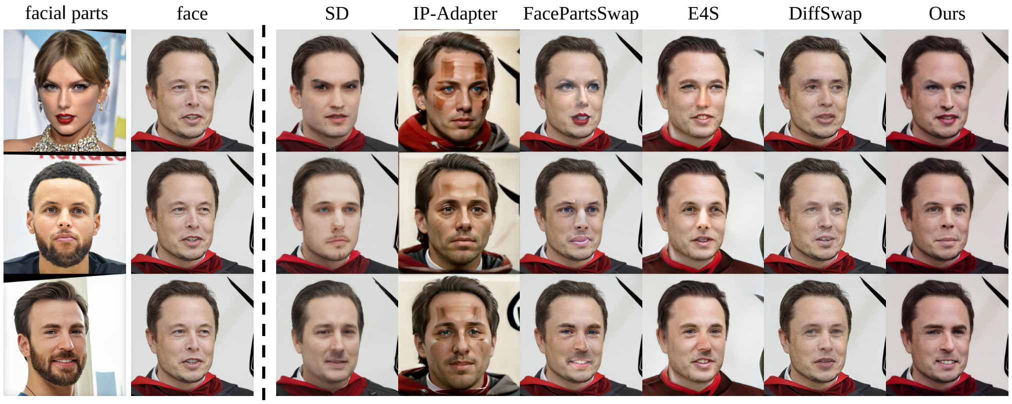

Figure 4: **Qualitative comparison of multiple facial parts swapping with a single reference face.** Our method can naturally replace multiple facial parts of one face with those of another and better preserve both the characteristics and the facial part shape. More results are presented in Fig. 10.

replaced attributes can be arbitrary. We demonstrate the simultaneous replacement of eyes, nose and mouth. Fig. 4 showcases the results of multi-attribute replacement using source-target pairs.

Both the SD and IP-Adapter struggle to maintain the non-replacement areas unchanged, and the similarity of the replaced attributes is not high, highlighting the limitations of pixel space manipulations. While FacePartsSwap and E4S can retain a higher degree of attribute similarity, the replacement results often appear visually inconstant and forced, particularly when there is a significant difference between the source image and the target image, such as in skin tone or facial angle. In contrast, the replacement effect of Diffswap is not pronounced.

Our method outperforms these approaches by offering better consistency across both the replaced attributes and the unaffected areas, leading to a more seamless and natural integration of the replaced features regardless of discrepancies in the reference images. Moreover, Fig. 5 demonstrates the results when each replaced attribute originates from different source images. Our method still significantly surpasses other approaches, as it can combine distinct attribute features and generate natural-looking facial photos. This superior performance indicates that our method effectively extracts and integrates the characteristics of individual attributes, even when dealing with varied sources. It reinforces our method's flexibility and robustness in handling complex face manipulation scenarios where each facial feature may require a different treatment based on its unique reference image.

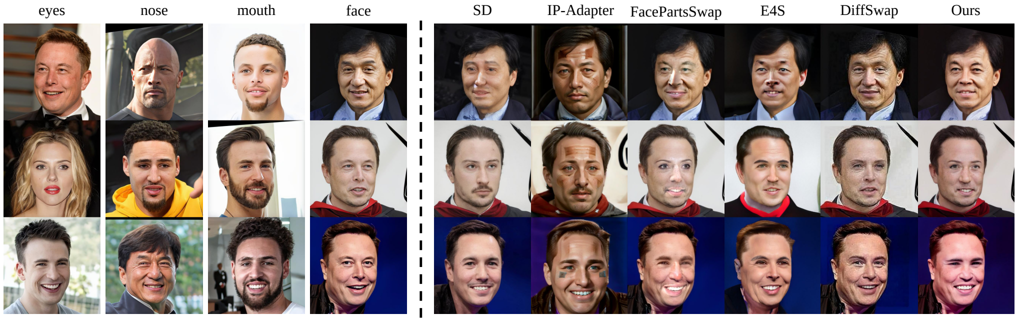

Figure 5: **Qualitative comparison of multi swapping with multiple reference faces.** Our method remains robust to different appearances of various reference facial parts.

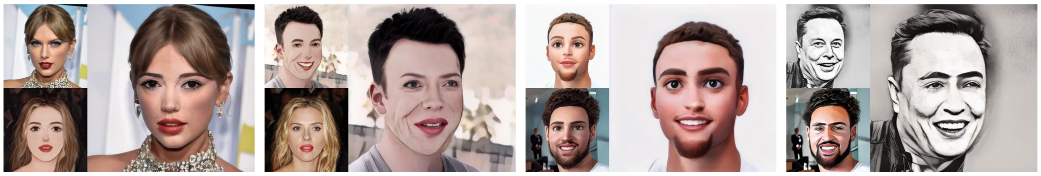

Figure 6: **Facial parts swapping on images with different styles.** As an extended application, FuseAnyPart can use the facial parts of reference images with different styles to generate harmonious faces without changing the style of the target face. We show that the features of the facial parts and the image style can be well decoupled.

4.1.3 Fusion Across Different Styles

Fig. 6 shows the results of our method performing multi-attribute replacements on images spanning various styles. This is a limitation often encountered in many face swapping methods, as they typically rely on facial segmentation models which are bound by the constraints of their training data and tend to have poor generalization on unseen data. Benefiting from the open-set detection model, our method can extract any regional feature from the reference images, which helps our method to generalize well on the data with different distributions.

To get images in different styles, we apply a style modification model [19] to real photographs to generate a series of images in different styles, including cartoon, 3D, sketch, and more. We then use these stylized images as references to perform multi-attribute replacements following the experimental setup described earlier. Despite the style discrepancies between reference images, our method is still able to accurately extract and fuse the targeted features. The generated results maintain the characteristics of the reference images while preserving the style of the original face reference. This demonstrates our method's robust capability to handle diverse styles and perform complex attribute fusion tasks effectively.

4.2 Quantitative Comparisons

Evaluation Metric. Following common practice, we adopt Fréchet Inception Distance (FID) [10, 21] to evaluate the quality of the generated images. Our method is capable of generating faces with multiple reference images, including a face reference image and three facial part reference images (an eyes image, a nose image and a mouth image). To evaluate the effect of facial part reference images, we propose a metric named **FPSim** (Facial Part Similarity) to measure the similarity between the corresponding facial parts of the generated face and those of reference images, and FPSim-E, FPSim-N and FPSim-M are distinct metrics that respectively measure the similarity of the eyes, nose, and mouth. FPSim is defined as $\frac{f_a \cdot f_b}{\|f_a\| \|f_b\|}$, where f_a and f_b are the attribute-level features of the generated face and the corresponding reference images. To extract attribute-level features, we train three facial attribute-level feature extractor models with ResNet50 [9] and ArcFace loss [4] on the CelebA-HQ dataset for eyes, noses and mouths respectively. To measure the ability to reconstruct

Table 1: **Quantitative Comparisons on FF++.** We report Fréchet inception distance, eye similarity, nose similarity, mouth similarity and Mean Square Error and show that our method achieves SoTA or competitive results compared with existing methods. FacePartsSwap is essentially a cut & paste method, rather than a generative one, and thus has a higher FPSim. Therefore, we only present its results here and do not include it in the quantitative comparisons.

Methods	FID ↓	FPSim-E↑	FPSim-N↑	FPSim-M ↑	MSE↓
StableDiffusion [24]	18.57	0.3080	0.2215	0.2127	1.66
IP-Adapter [36]	69.35	0.2865	0.2066	0.1886	13.72
FacePartsSwap [5]	44.23	0.3269	0.2190	0.2220	24.40
E4S [15]	30.61	0.2764	**0.4047**	0.1903	3.03
Diffswap [38]	12.07	0.2461	0.1967	0.1731	**0.15**
Ours	**10.54**	**0.3186**	0.2234	**0.2196**	0.77

Table 2: **Quantitative comparison of feature fusion under different ablative configurations.** Both the generation quality and facial part similarity are measured.

Method	FID↓	FPSim-E↑	FPSim-N↑	FPSim-M ↑	MSE↓
Cross-attention	15.81	0.2542	0.1763	0.1771	1.02
Multiple Cross-attention	15.32	0.2407	0.1834	0.2063	1.94
Cross-attention + Addition-in-Conv	16.66	0.2706	0.1897	0.1797	**0.66**
Cross-attention + Addition-in-CA	**10.51**	0.3108	0.2128	0.2158	0.71
Cross-attention + Addition-in-CA + Hierarchy	28.96	0.2808	0.2034	0.2077	1.33
Addition-in-CA (Ours)	10.54	**0.3186**	**0.2234**	**0.2196**	0.77

Cross-attention: Using cross-attention to inject conditional features.
Add-in-Conv: Addition within convolutional layers.
Add-in-CA: Addition within cross-attention layers.
Hierarchy: Use of hierarchical features.

faces with reference images from the same ID, we compute the Mean Square Error (MSE) between generated images and reference images.

Quantitative Comparison. As indicated in Tab. 1, we compare our method with previous methods on the FaceForensics++ [25] dataset. The results show that our method outperforms previous methods in FID significantly, indicating that we can generate high-fidelity swapped faces and can better preserve naturalness and harmony. Meanwhile, we also achieve comparable results on attribute-level metrics, demonstrating that our method can also keep the characteristics of the swapped facial parts. Notably, we observed a limitation in DiffSwap [38], with its tendency to yield results more resemble the source face rather than the intended target face, as illustrated in Fig. 4. Therefore, when both the source and target faces come from the same ID, DiffSwap achieves a lower MSE (0.15) compared to ours (0.77). Nonetheless, it is more common for the IDs of the source and target faces to differ; in these cases, our method consistently shows superior performance.

4.3 Ablation Study

Qualitative comparisons are in Fig. 7 and the quantitative comparison is shown in Tab. 2, where both the generation quality and facial part similarity are measured.

Cross-attention *vs.* Addition. To inject conditional features of reference facial parts, we directly add the swapped face image feature to the UNet latent feature instead of using cross-attention. As shown in the first and last rows of Tab. 2, direct addition significantly enhances the swapping performance. We also try to use different cross-attention modules for different facial parts and add their results together to form the latent features of the UNet (denoted by "Multiple Cross-attention"). This method provides only a limited improvement to the model's performance (the 2nd row of Tab. 2 and the 4th column of Fig. 7). Moreover, we try to combine the two fusion methods, adding the swapped face image feature to the output of cross-attention layers (the 2nd row of Tab. 2). Although there is a slight improvement in image fidelity (FID and MSE), the facial part similarity all decreased (FPSim-E, FPSim-N and FPSim-M).

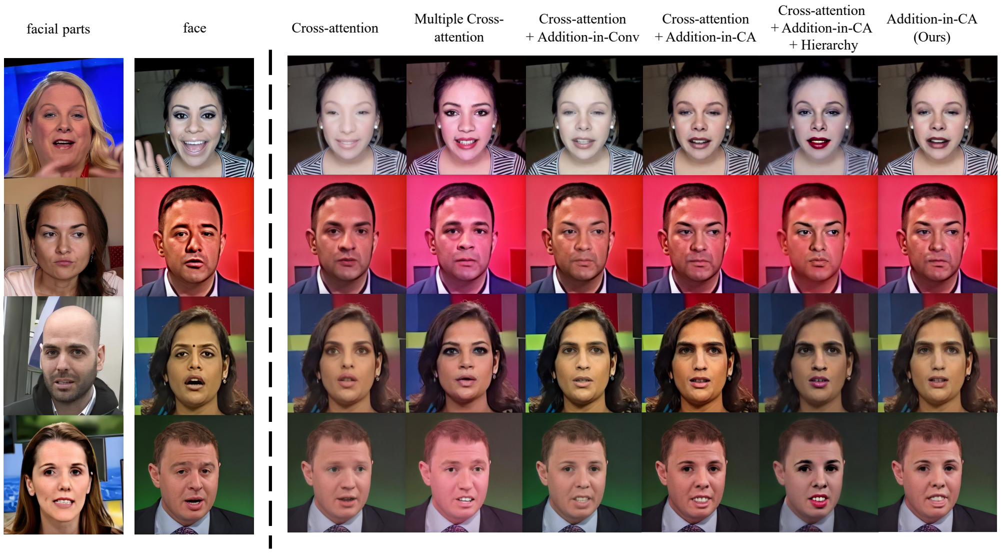
facial parts
face
Cross-attention
Multiple Cross-attention
Cross-attention + Addition-in-Conv
Cross-attention + Addition-in-CA
Cross-attention + Addition-in-CA + Hierarchy
Addition-in-CA (Ours)

Figure 7: Qualitative comparison of different ablative settings.

Feature Injection Across Layers. We inject the swapped face image feature between the two convolutional layers in each ResNet block of the UNet, rather than in the cross-attention layers. The results in Tab. 2 suggest that the Addition-based Injection Module should be positioned in the cross-attention layers (Row 3 and 4). From the 5th and 6th columns of Fig. 7, we can observe that fusion in the cross-attention layer preserves more details and achieves higher similarity than that in the convolution layers.

Hierarchical Feature. Features from the 4th block of the CLIP image encoder and one after every four blocks are extracted and concatenated to form the output of the CLIP image encoder as the hierarchical feature, which contains abundant facial visual information. According to the Row 4 and 5 of Tab. 2, the method is unable to generate high-fidelity and realistic faces with a significant decrease in image quality metrics, which is confirmed by the 7th column in Fig. 7.

5 Societal Impacts, Limitations and Conclusion

Societal Impacts. The proposed FuseAnyPart is fundamentally harmless. Nevertheless, misuse of it, e.g., applications with copyright issues and racial issues, could have negative effects on society. As a result, we call for a conscientious and ethical implementation of FuseAnyPart.

Limitations. While FuseAnyPart demonstrates strong performance in facial parts swapping, it still has some limitations at the current stage. First, while FuseAnyPart performs well on a range of images, there may be challenges with faces that have extreme poses, occlusions, or expressions. Additionally, our method is primarily designed for facial parts swapping and does not directly tackle the challenge of preserving or transforming facial expressions during the process of swapping. FuseAnyPart is based on diffusion models, which typically exhibit high computational complexity due to recursive iterations. Algorithms like Latent Consistency Models (LCM) [17] can accelerate inference by reducing the number of iterations, while techniques such as int8 model quantization can significantly lower computational load. Together, these strategies enhance the speed of FuseAnyPart. Like most generative models, FuseAnyPart relies on high-quality training datasets. The quality of the images can be enhanced using super-resolution methods.

Conclusion. This paper proposes FuseAnyPart, a novel diffusion-driven method for facial parts swapping. FuseAnyPart first extracts multiple decomposed features from face images with masks obtained from an open-set detection model. Then parts from different faces are aggregated in latent space with the Mask-based Fusion Module. An injection module injects this conditional information into UNet for fusing effectively. Extensive experiments validate the superiority of FuseAnyPart.

References

[1] Haoxing Chen, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Changhua Meng, Huijia Zhu, Weiqiang Wang, et al. Diffute: Universal text editing diffusion model. *Advances in Neural Information Processing Systems*, 36, 2024.

[2] Li Chen, Mengyi Zhao, Yiheng Liu, Mingxu Ding, Yangyang Song, Shizun Wang, Xu Wang, Hao Yang, Jing Liu, Kang Du, et al. Photoverse: Tuning-free image customization with text-to-image diffusion models. *arXiv preprint arXiv:2309.05793*, 2023.

[3] Siying Cui, Jiankang Deng, Jia Guo, Xiang An, Yongle Zhao, Xinyu Wei, and Ziyong Feng. Idadapter: Learning mixed features for tuning-free personalization of text-to-image models. *arXiv preprint arXiv:2403.13535*, 2024.

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[5] Claudio Ferrari, Matteo Serpentoni, Stefano Berretti, and Alberto Del Bimbo. What makes you, you? analyzing recognition by swapping face parts. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 945–951. IEEE, 2022.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[7] Jingtao Guo and Yi Liu. Facial parts swapping with generative adversarial networks. *Journal of Visual Communication and Image Representation*, 78:103152, 2021.

[8] Yue Han, Jiangning Zhang, Junwei Zhu, Xiangtai Li, Yanhao Ge, Wei Li, Chengjie Wang, Yong Liu, Xiaoming Liu, and Ying Tai. A generalist facex via learning unified facial representation. *arXiv preprint arXiv:2401.00551*, 2023.

[9] 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.

[10] 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. *Advances in neural information processing systems*, 30, 2017.

[11] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[12] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5074–5083, 2020.

[13] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[14] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[15] Zhian Liu, Maomao Li, Yong Zhang, Cairong Wang, Qi Zhang, Jue Wang, and Yongwei Nie. Fine-grained face swapping via regional gan inversion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8578–8587, 2023.

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[17] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023.

[18] Yiyang Ma, Huan Yang, Wenjing Wang, Jianlong Fu, and Jiaying Liu. Unified multi-modal latent diffusion for joint subject and text conditional image generation. *arXiv preprint arXiv:2303.09319*, 2023.

[19] Yifang Men, Yuan Yao, Miaomiao Cui, Zhouhui Lian, and Xuansong Xie. Dct-net: Domain-calibrated translation for portrait stylization. volume 41, pages 1–9, 2022.

[20] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7184–7193, 2019.

[21] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420, 2022.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[23] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[25] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.

[26] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023.

[27] Sridhar Sola and Darshan Gera. Unmasking your expression: Expression-conditioned gan for masked face inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5907–5915, 2023.

[28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[29] Qianru Sun, Ayush Tewari, Weipeng Xu, Mario Fritz, Christian Theobalt, and Bernt Schiele. A hybrid model for identity obfuscation by face replacement. In *Proceedings of the European conference on computer vision (ECCV)*, pages 553–569, 2018.

[30] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022.

[31] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

[32] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023.

[33] Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv*, 2023.

[34] Zhiliang Xu, Hang Zhou, Zhibin Hong, Ziwei Liu, Jiaming Liu, Zhizhi Guo, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Styleswap: Style-based generator empowers robust face swapping. In *European Conference on Computer Vision*, pages 661–677. Springer, 2022.

[35] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 672–681, 2021.

[36] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023.

[37] Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, and Zhongliang Jing. Ssr-encoder: Encoding selective subject representation for subject-driven generation, 2024.

[38] Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8568–8577, 2023.

A More qualitative results

As discussed in Sec. 4.1.1, the qualitative comparison of the nose and mouth swaps is presented in Fig. 8 and 9.

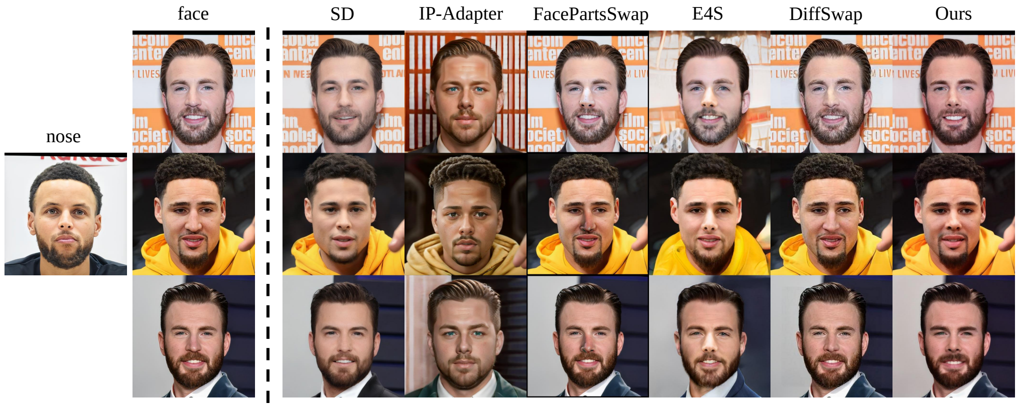

Figure 8: Qualitative comparison of nose swapping.

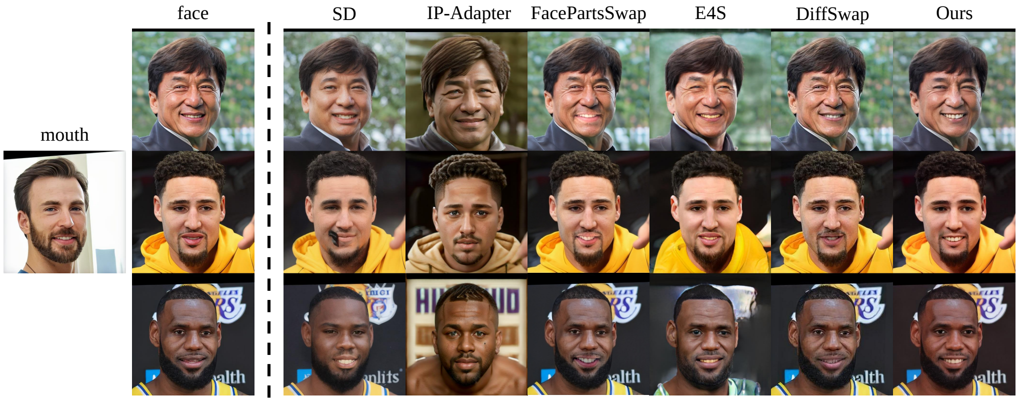

Figure 9: Qualitative comparison of mouth swapping.

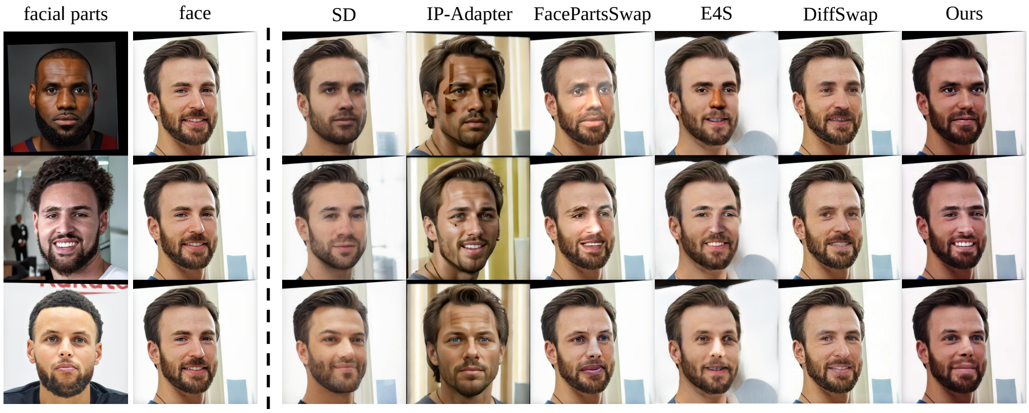

Figure 10: More qualitative comparisons of multiple facial parts swapping. This provides additional examples related to Fig. 4.

As shown in Fig. 11, FuseAnyPart performs well when swapping facial parts from individuals of significantly different racial or age groups.

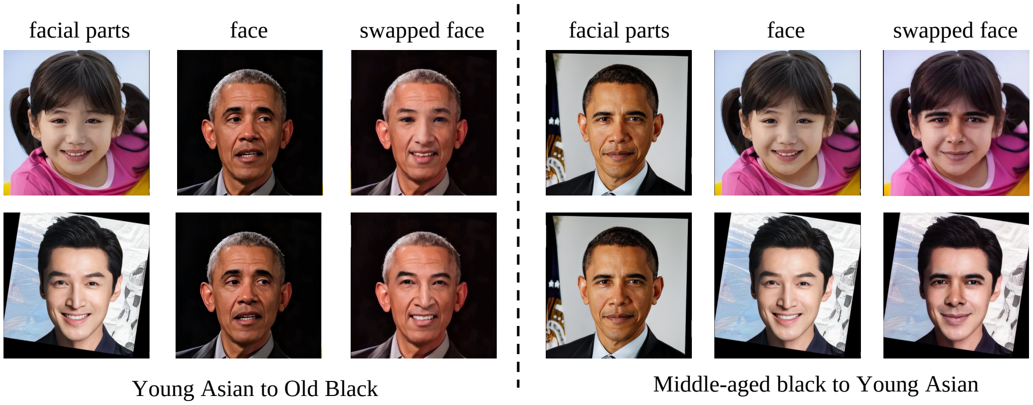

Figure 11: Illustrations of facial parts from significantly different racial and age groups. Facial part swapping between source and target images that significantly differ in age and race.

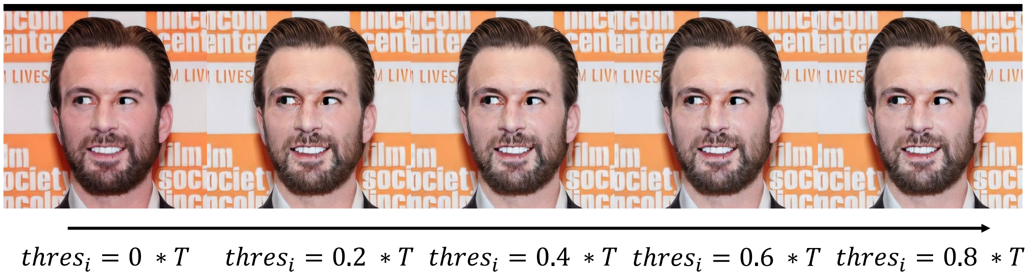

Figure 12: The skin color change issue can be effectively resolved by replacing the generated skin regions with the inverted latent representations of the original skin color using DDIM inversion in the denoising process. The threshold indicates the number of steps performed above the replacement operation in the denoising process.

FuseAnyPart may encounter issues with color changes in generated images, but this problem can be addressed by replacing the generated skin regions with inverted latent representations of the original skin color using DDIM. The results are presented in Fig. 12.

FuseAnyPart was also qualitatively compared with DiffFace, and the results are shown in Fig. 13. More diverse results of multi swapping with multiple reference faces are presented in Fig. 14.

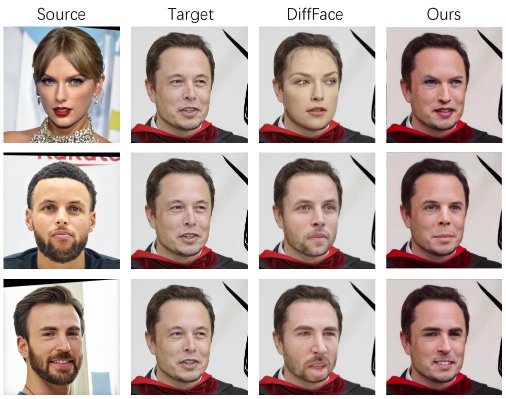

Figure 13: Comparison with DiffFace. DiffFace generates images with local distortions in the eyes and mouth, whereas our method produces cleaner results that are more similar to the source image regarding the facial parts.

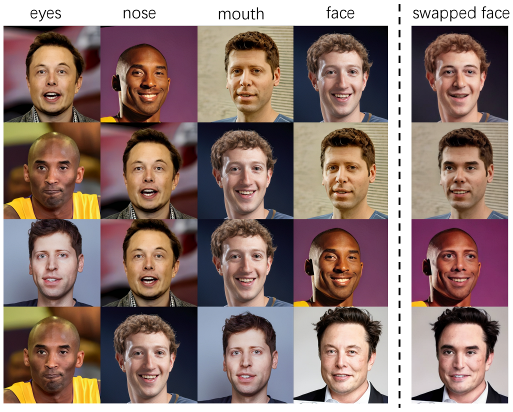

Figure 14: Qualitative results of swapping face parts from different sources to a target face.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in abstract and introduction in Sec. 1 accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations in Sec. 5

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

Justification: The approach presented in this paper utilizes deep learning and adopts a data-driven experimental methodology.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The quantitative findings utilize datasets that are publicly accessible.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code and model files will be made available to the public after the author has finished refining them.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Please see Sec. 4

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Error bars are not reported because it would be too computationally expensive.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see Sec. 4

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper conforms with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please see Sec. 5.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: The authors will personally evaluate the risks and decide on the publication of the model files.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The code and data are mentioned, but we were unable to find the license for the dataset we used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: Currently, the data and code remain private, but there may be consideration for public release in the future.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: No Crowdsourcing used in this work.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We are not sure if we have obtained IRB approvals.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.