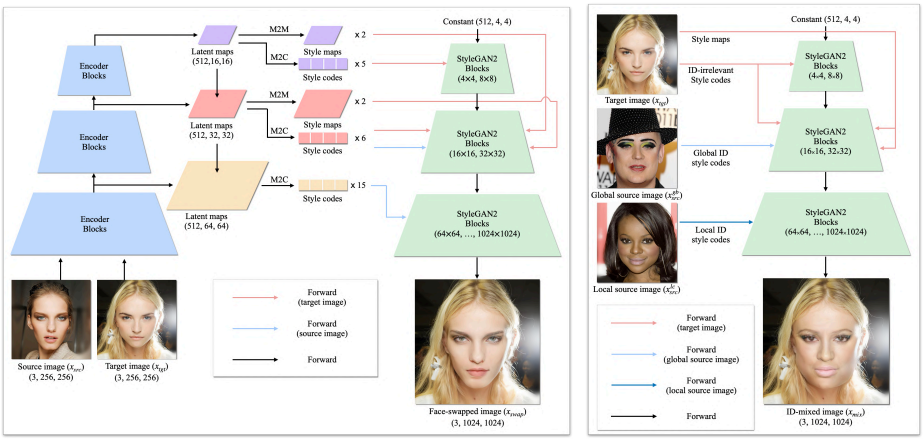
Supplementary Material for
MFIM: Megapixel Facial Identity Manipulation



(a) Face swapping (b) ID mixing

Fig. 1: **The architecture of MFIM.** Figure 1a shows the process of face swapping. The facial attribute encoder extracts style codes and style maps from source and target images. These are given to the pretrained StyleGAN generator as inputs. Figure 1b shows the process of ID mixing. The ID-style codes are extracted from two source images, instead of a single source image.

A Architecture

In this section, we describe the architectures of facial attribute encoder, generator and discriminator.

A.1 Facial Attribute Encoder.

Our facial attribute encoder, which is based on the psp [10] encoder, uses the same encoder backbone (blue structures denoted as 'Encoder Blocks' in Figure 1a) as the psp encoder. As shown in Figure 1a, the encoder backbone extracts the hierarchical latent maps from the given image. The M2C and M2M blocks of our facial attribute encoder extract the style codes and style maps from the hierarchical latent maps extracted from the backbone, respectively. The details of encoding process are as follows.

Table 1: **Architecture of M2M block.** M2M block has shared convolutional layers at the top, but separated convolutional layers at the bottom. All convolutional layers have kernel size of 3×3, stride of 1, and padding size of 1. C_{in} and C_{out} for the convolutional layer denotes the input and output channel dimensions, respectively. a for the LeakyReLU layer denotes the negative slope. To encourage the style maps to be similar to the noise inputs which is used in StyleGAN pretraining, M2M block has an instance normalization [12] layer at the last which makes the style maps to be normally distributed.

(**Input**): latent maps (c, h, w)	
Conv ($C_{in} = c$, $C_{out} = c$) LeakyReLU ($a = 0.01$)	
Conv ($C_{in} = c$, $C_{out} = c'$) LeakyReLU ($a = 0.01$)	
Conv ($C_{in} = c'$, $C_{out} = c'$) InstanceNorm	Conv ($C_{in} = c'$, $C_{out} = c'$) InstanceNorm
(**Output 0**): style maps (c', h, w)	(**Output 1**): style maps (c', h, w)

Style codes. The architecture of the M2C block is the same as that of the Map2Style block of the pSp encoder. However, the pSp encoder produces eighteen style codes because it maps the image to $\mathcal{W}^+$ space [1], whereas our facial attribute encoder maps the image to $\mathcal{S}$ space [14], so twenty-six style codes, $\{c_i\}_{i=0}^{25}$. Then, the style codes go through the following additional steps:

$$s_i = \alpha_i c_i + \mu_i, \tag{1}$$

where $\{\alpha_i\}_{i=0}^{25}$ is a set of learnable parameters and $\{\mu_i\}_{i=0}^{25}$ is a set of style codes that maps an average latent code of $\mathcal{W}$ space [5] to $\mathcal{S}$ space. α_i, c_i, and μ_i have the same dimensions.

We extract the style codes from the source image, x_{src}, and the target image, x_{tgt}, respectively, and combine them to construct the final style codes. Let us denote the style codes extracted from x_{src} and x_{tgt}, $\{s_i^{src}\}_{i=0}^{25}$ and $\{s_i^{tgt}\}_{i=0}^{25}$, respectively. We construct the ID-irrelevant style codes, $\{s_i^{tgt}\}_{i=0}^{b-1}$, by taking a subset of $\{s_i^{tgt}\}_{i=0}^{25}$, and the ID style codes $\{s_i^{src}\}_{i=b}^{25}$ from $\{s_i^{src}\}_{i=0}^{25}$, where b is a hyperparameter for the border index between the ID and ID-irrelevant style codes. We set $b = 8$. Then, the final style codes, $\{s_i\}_{i=0}^{25}$, are constructed by combining $\{s_i^{tgt}\}_{i=0}^{b-1}$ and $\{s_i^{src}\}_{i=b}^{25}$. Finally, $\{s_i\}_{i=0}^{25}$ is used in weight demodulation operation [6].

Style maps. Our facial attribute encoder introduces an M2M block with the architecture depicted in Table 1 to extract the style maps from the target image. As shown in Table 1, the M2M block takes the latent maps as input and produces

two groups of style maps, which are denoted as **Output 0** and **Output 1** in Table 1 respectively, of the same spatial size as the input latent maps.

Our encoder produces a total of *four* groups of style maps: two groups with a spatial size of 16×16, $\left\{m_i^{16\times16}\right\}_{i=0}^{1}$, and the remaining two groups have a spatial size of 32×32, $\left\{m_i^{32\times32}\right\}_{i=0}^{1}$. All of these style maps have the channel dimensions of 512. Finally, these style maps are given to the pretrained StyleGAN generator as noise inputs.

A.2 Generator

We use the pretrained generator of StyleGAN [6], so we use the same architecture with StyleGAN without modification except for the mapping network that maps a random vector $z \in \mathcal{Z}$ to an intermediate latent space $\mathcal{W}$. We replace the mapping network with the facial attribute encoder which produces the ID-irrelevant style codes, ID style codes and style maps. These are forwarded appropriately to each layer of the pretrained StyleGAN generator, as shown in Tables 2 and 3. Table 2 describes the process of face swapping, which uses a single source image, x_{src}, but Table 3 describes the process of id mixing, which uses the global and local source images, x_{src}^{gb} and x_{src}^{lc}.

A.3 Discriminator

We use the pretrained discriminator of StyleGAN [6], so we use the same architecture with StyleGAN without modification.

B Hyperparameters

Table 4 shows weights for each loss to train our model. Following StyleGAN [6], we use R1 regularization [9] every sixteen training steps. Table 5 shows additional hyperparameters for optimization. For the optimizer, we use the Ranger optimizer, which is a combination of RAdam [8] and Lookahead [15], following pSp [10]. We use a learning rate of $1e-4$ and decrease it by $2e-5$ every 40,000 steps after 500,000 steps. We use a batch size of four, which means that we use four pairs of source and target images for training. However, for one of the four pairs, we make the source image and the target image the same, so that the generator performs self-reconstruction on that pair.

C Preprocess and Postprocess

C.1 Data preprocess

We use FFHQ [5], which consists of 70,000 human faces at 1024×1024 resolution, for the training dataset. It is noteworthy that the most of the previous face-swapping models [7, 13, 16, 3] extend the training dataset by combining multiple

Table 2: **Inputs that each layer of the pretrained StyleGAN generator takes for face swapping.** The ID-irrelevant style codes, $\{s_i^{tgt}\}_{i=0}^{7}$, and style maps, $\{m_i^{16\times16}\}_{i=0}^{1}$ and $\{m_i^{32\times32}\}_{i=0}^{1}$ are extracted from x_{tgt}, while the ID style codes, $\{s_i^{src}\}_{i=8}^{25}$ are extracted from x_{src}.

$\mathcal{S}$ layer index	Resolution	Layer name	Style code	Style code type	Style maps
0	4×4	Conv	s_0^{tgt}	ID-irrelevant	-
1	4×4	ToRGB	s_1^{tgt}	ID-irrelevant	-
2	8×8	ConvUp	s_2^{tgt}	ID-irrelevant	-
3	8×8	Conv	s_3^{tgt}	ID-irrelevant	-
4	8×8	ToRGB	s_4^{tgt}	ID-irrelevant	-
5	16×16	ConvUp	s_5^{tgt}	ID-irrelevant	$m_0^{16\times16}$
6	16×16	Conv	s_6^{tgt}	ID-irrelevant	$m_1^{16\times16}$
7	16×16	ToRGB	s_7^{tgt}	ID-irrelevant	-
8	32×32	ConvUP	s_8^{src}	ID	$m_0^{32\times32}$
9	32×32	Conv	s_9^{src}	ID	$m_1^{32\times32}$
10	32×32	ToRGB	s_{10}^{src}	ID	-
11	64×64	ConvUP	s_{11}^{src}	ID	-
12	64×64	Conv	s_{12}^{src}	ID	-
13	64×64	ToRGB	s_{13}^{src}	ID	-
14	128×128	ConvUP	s_{14}^{src}	ID	-
15	128×128	Conv	s_{15}^{src}	ID	-
16	128×128	ToRGB	s_{16}^{src}	ID	-
17	256×256	ConvUP	s_{17}^{src}	ID	-
18	256×256	Conv	s_{18}^{src}	ID	-
19	256×256	ToRGB	s_{19}^{src}	ID	-
20	512×512	ConvUP	s_{20}^{src}	ID	-
21	512×512	Conv	s_{21}^{src}	ID	-
22	512×512	ToRGB	s_{22}^{src}	ID	-
23	1024×1024	ConvUP	s_{23}^{src}	ID	-
24	1024×1024	Conv	s_{24}^{src}	ID	-
25	1024×1024	ToRGB	s_{25}^{src}	ID	-

Table 3: **Inputs that each layer of the pretrained StyleGAN generator takes for ID mixing.** The ID-irrelevant style codes, $\{s_i^{tgt}\}_{i=0}^{7}$, and style maps, $\{m_i^{16\times16}\}_{i=0}^{1}$ and $\{m_i^{32\times32}\}_{i=0}^{1}$ are extracted from x_{tgt}. However, the global ID style codes, $\{s_i^{src}\}_{i=8}^{9}$, are extracted from x_{src}^{gb}, and the local ID style codes, $\{s_i^{src}\}_{i=10}^{25}$, are extracted from x_{src}^{lc}.

$\mathcal{S}$ layer index	Resolution	Layer name	Style code	Style code type	Style maps
0	4×4	Conv	s_0^{tgt}	ID-irrelevant	-
1	4×4	ToRGB	s_1^{tgt}	ID-irrelevant	-
2	8×8	ConvUp	s_2^{tgt}	ID-irrelevant	-
3	8×8	Conv	s_3^{tgt}	ID-irrelevant	-
4	8×8	ToRGB	s_4^{tgt}	ID-irrelevant	-
5	16×16	ConvUp	s_5^{tgt}	ID-irrelevant	$m_0^{16\times16}$
6	16×16	Conv	s_6^{tgt}	ID-irrelevant	$m_1^{16\times16}$
7	16×16	ToRGB	s_7^{tgt}	ID-irrelevant	-
8	32×32	ConvUP	s_8^{src}	Global ID	$m_0^{32\times32}$
9	32×32	Conv	s_9^{src}	Global ID	$m_1^{32\times32}$
10	32×32	ToRGB	s_{10}^{src}	Local ID	-
11	64×64	ConvUP	s_{11}^{src}	Local ID	-
12	64×64	Conv	s_{12}^{src}	Local ID	-
13	64×64	ToRGB	s_{13}^{src}	Local ID	-
14	128×128	ConvUP	s_{14}^{src}	Local ID	-
15	128×128	Conv	s_{15}^{src}	Local ID	-
16	128×128	ToRGB	s_{16}^{src}	Local ID	-
17	256×256	ConvUP	s_{17}^{src}	Local ID	-
18	256×256	Conv	s_{18}^{src}	Local ID	-
19	256×256	ToRGB	s_{19}^{src}	Local ID	-
20	512×512	ConvUP	s_{20}^{src}	Local ID	-
21	512×512	Conv	s_{21}^{src}	Local ID	-
22	512×512	ToRGB	s_{22}^{src}	Local ID	-
23	1024×1024	ConvUP	s_{23}^{src}	Local ID	-
24	1024×1024	Conv	s_{24}^{src}	Local ID	-
25	1024×1024	ToRGB	s_{25}^{src}	Local ID	-

Table 4: **Weights for each loss.** Each loss is described in the main manuscript.

λ_{id}	λ_{recon}	λ_{adv}	λ_{R_1}	λ_{shape}	λ_{pose}	λ_{exp}	R_1 step
2.0	1.0	0.1	10.0	5.0	1.0	1.0	16

Table 5: **Hyperparameters for optimization.** The details are described in Section B.

Training steps	Optimizer	Learning rate	Learning rate decay	Batch size	Self-recon size
700,000	Ranger	0.0001	Step	4	1

datasets, but we only use FFHQ. Therefore, our model can be trained more efficiently because our model does not require any additional preprocess steps such as image alignment to combine the multiple datasets.

For training, we basically follow the image preprocess protocol of pSp [10]. However, for $\mathcal{L}_{adv}$ and R_1, we use the images with the size of 1024×1024. Furthermore, for 3DMM supervision, we preprocess the images by following the image preprocess protocol of DECA [2] before forwarding the images to DECA encoder.

C.2 Postprocess: ROI Only Synthesis

Our model can faithfully reconstruct the background or hair style of x_{tgt}, but we can further improve our model to reconstruct the high-frequency details of the background or hair style via ROI only synthesis.

Note that it does not require a segmentation label at all. This process is depicted in Figure 2. Assuming that the image is aligned, we use a mask, which has a size of 1024×1024, with a fixed box at the expected location of the face. Specifically, we set the size of the box to a width of 512 and a height of 608 and top-left coordinates, $(top,\ left)$, to $(384,\ 256)$. The inside of the box has a value of one, and the outside has a value of zero. Then, we blur the boundary by downsampling the mask to the size of 16×16 and upsampling it to the size of 1024×1024 again. With this mask, the final output image is generated as

$$m \odot x_{swap} + (1 - m) \odot x_{tgt}, \tag{2}$$

where m is a mask and $\odot$ is the element-wise product. Note that it is not used at the training phase, only at the inference phase. Also, we use it only in the qualitative results, not in the quantitative results at all.

D Analysis on 3DMM Supervision

We compare our 3DMM supervision method and that of HifiFace [13]. We first describe each method and then compare them with experimental results.

Fig. 2: **ROI only synthesis.** The details are described in Section C.2.

Table 6: **Quantitative comparison between $\mathcal{L}_{lm}$ and $\mathcal{L}_{param}$.** The metrics are the same with those in the main manuscript. Also, the configuration (B) is the same with that in the main manuscript. $\mathcal{L}_{lm}$ and $\mathcal{L}_{param}$ are the 3DMM supervision methods of HifiFace [13] and ours, respectively.

Configuration	Identity ↓	Shape ↓	Expression ↓	Pose ↓	Pose-HN ↓
B.	96.066	0.887	0.424	0.053	3.839
B + $\mathcal{L}_{lm}$.	96.016	0.892	0.418	0.046	3.683
B + $\mathcal{L}_{param}$.	96.153	0.842	0.426	0.060	4.173

D.1 Method

For the 3DMM supervision, our model utilizes the 3DMM parameter reconstruction loss which is formulated as

$$\mathcal{L}_{param} = \lambda_{shape}\mathcal{L}_{shape} + \lambda_{pose}\mathcal{L}_{pose} + \lambda_{exp}\mathcal{L}_{exp}, \tag{3}$$

where $\mathcal{L}_{shape}$, $\mathcal{L}_{pose}$, and $\mathcal{L}_{exp}$ are described in the main manuscript, and λ_{shape}, λ_{pose}, and λ_{exp} are weights for each loss.

On the other hand, HifiFace utilizes the landmark reconstruction loss. Note that 3DMM can reconstruct a 3D face using 3DMM parameters and extract landmark keypoints corresponding to the 3D face. Using this capability, HifiFace encourages the landmark keypoints of the generated image, $\{q_k^{gen}\}_{k=1}^{K}$ to be equal to its ground-truth landmark keypoints, $\{q_k^{gt}\}_{k=1}^{K}$. Here, when using DECA [2], $K = 68$, and the ground-truth landmark keypoints are extracted from the reconstructed 3D face using the shape parameter of the source image and the pose, expression, and cam parameters of the target image. We apply this method to our model to formulate the landmark reconstruction loss as

$$\mathcal{L}_{lm} = \frac{1}{K}\sum_{k=1}^{K} || \{q_k^{gen}\} - \{q_k^{gt}\} ||_1. \tag{4}$$

Table 7: **Full ablation study of MFIM.** This table is the same with the table of the ablation study in the main manuscript, but the configuration (F) is newly added.

Configuration	Identity ↓	Shape ↓	Expression ↓	Pose ↓	Pose-HN ↓
A. Baseline MFIM	70.160	0.383	1.116	0.145	7.899
B. + style maps	91.430	0.823	0.398	0.051	3.795
C. + $\mathcal{L}_{shape}$	86.476	0.635	0.864	0.085	5.091
D. + $\mathcal{L}_{pose}$	86.777	0.634	0.860	0.078	4.797
E. + $\mathcal{L}_{exp}$	91.469	0.782	0.400	0.057	4.095
F. + $\mathcal{L}_{lm}$	92.018	0.778	0.387	0.041	3.876

D.2 Comparison between $\mathcal{L}_{lm}$ and $\mathcal{L}_{param}$

In Table 6, we compare $\mathcal{L}_{lm}$ and $\mathcal{L}_{param}$ on CelebA-HQ [4]. Here, unlike the quantitative experiment on CelebA-HQ in the main manuscript, we use 30,000 face-swapped images instead of 300,000. Specifically, we randomly assign an image to each image in CelebA-HQ and make 30,000 pairs of the source image and target image.

The configuration (B) in Table 6 is the same with that in the main manuscript. Then, we construct the configurations (B+$\mathcal{L}_{lm}$) and (B+$\mathcal{L}_{param}$) by adding $\mathcal{L}_{lm}$ and $\mathcal{L}_{parm}$ to the configuration (B), respectively. The configuration (B+$\mathcal{L}_{param}$) is the same with the configuration (E), our proposed model, in the main manuscript.

As shown in Table 6, adding $\mathcal{L}_{lm}$ to the configuration (B) does not improve the shape score while $\mathcal{L}_{param}$ improves the shape score. However, we can see that $\mathcal{L}_{lm}$ improves the pose score by comparing the configurations (B) and (B+$\mathcal{L}_{lm}$). We think that this may be because the pose, which is the more global attribute than the shape and expression, has a greater effect on the landmark regression than the shape or expression. For this reason, the most effective way to decrease $\mathcal{L}_{lm}$ can be to match the pose of the generated image to that of the target image. As a result, the model focuses on matching poses, and may not be sufficiently motivated to improve the shape score.

In contrast, we use a separate loss for each attribute. In particular, to decrease $\mathcal{L}_{shape}$, the face shape of the generated image should be the same as that of the source image. Due to this difference, $\mathcal{L}_{param}$ can improve the shape score, while $\mathcal{L}_{lm}$ cannot. Although the pose score is somewhat degraded after applying $\mathcal{L}_{param}$, transforming the face shape rather than preserving the pose is one of our important goals. Furthermore, the configuration (B+$\mathcal{L}_{param}$) still shows the visually plausible results in terms of the pose. Therefore, we propose $\mathcal{L}_{param}$ as our 3DMM supervision method.

D.3 Combination of $\mathcal{L}_{lm}$ and $\mathcal{L}_{param}$

Based on the results in Table 6, we further improve our model by combining $\mathcal{L}_{param}$ and $\mathcal{L}_{lm}$ as shown in Table 7. For the results in Table 7, we use 300,000

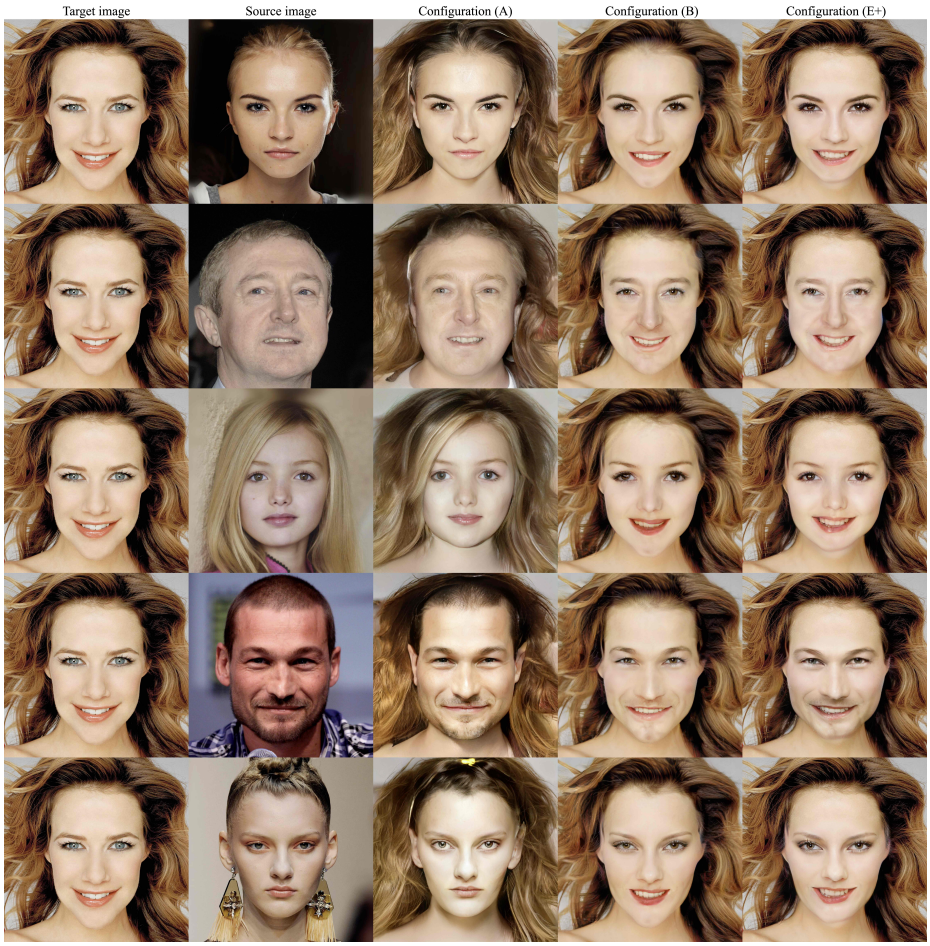


Fig. 3: **Ablation study of MFIM.** The configuration (A) transfers the ID attributes (e.g., eyes and face shape) of the source image while maintaining the overall structure and pose of the target image, but cannot reconstruct the details of the target image. The configuration (B) reconstructs the details of the target image better than the configuration (A), but the face shape of the source image is not sufficiently transferred. Finally, the configuration (E+) sufficiently transfers the face shape of the source image while preserving the ID-irrelevant attributes (e.g., pose and expression) of the target image. Furthermore, ROI only synthesis (Section C.2) allows our model to preserve the high-frequency details on hair or background of the target image.

Fig. 4: **Qualitative comparison on ID mixing.** MegaFS has a trouble in ID mixing because it cannot transfer the face shape of the global source image. In contrast, our model can create a new identity by blending the global (e.g., face shape) and local (e.g., eyes) ID attributes captured from the global and local source images, respectively.

face-swapped images, which is the same setting with that of the quantitative experiment on CelebA-HQ in the main manuscript.

In Table 7, we construct the configuration (F) by adding $\mathcal{L}_{lm}$ with the weight for this loss of 1,000 (i.e., $\lambda_{lm} = 1000$) to the configuration (E). Here, we use only some of the landmark keypoints instead of the full landkark keypoints to encourage our model to further focus on matching the pose. Specifically, we use $\{q_k^{gen}\}_{k\in\{9,31,37,46,49,55\}}$ and $\{q_k^{gt}\}_{k\in\{9,31,37,46,49,55\}}$. As shown in Table 7, the configuration (F) achieves the better pose and pose-HN scores than the configuration (E) without deterioration on the shape and expression scores. As a result, the configuration (F) achieves the better shape, expression, and pose scores than the configuration (B) at the same time. However, $\mathcal{L}_{lm}$ is not our contribution and the configuration (E) also shows the visually plausible results in terms of pose, so we propose the configuration (E) as our final model.

Figure 3 shows the qualitative results for several configurations. We construct the configuration (E+) by adding ROI only synthesis (Section C.2) to the configuration (E). As shown in Figure 3, the configuration (E+) transfers the ID attributes (e.g., eyes and face shape) of the source image actively while preserving the ID-irrelevant attributes (e.g., pose and expression) of the target image. In Figure 3, the differences between the configurations (A) and (B) show the effectiveness of the style maps, and the differences between the configurations (B) and (E+) show the effectiveness of the 3DMM supervision.

E Comparison with MegaFS on ID Mixing

One of the state-of-the-art models, MegaFS [16], has a potential to perform ID mixing because it also exploits the StyleGAN [6] architecture. However, MegaFS is not good at transforming the face shape as demonstrated in the manuscript. As a result, in Fig. 4, MegaFS fails to performing ID mixing because it cannot transfer the round face shape of the global source image to the target image. It only transfers the eyes of the local source image to the target image. For

this reason, the generated image by MegaFS does not seem an ID-mixed image. In contrast, our model, MFIM, can transfer the round face shape of the global source image and the eyes of the local source image at the same time. As a result, the generated image by MFIM seems an ID-mixed image.

F Additional Samples

Figure 5 shows the qualitative results of face swapping on FaceForensics++ [11]. Figures 6, 7, 8, and 9 show the qualitative results of face swapping on CelebA-HQ [4]. Figure 10 shows the qualitative results of ID mixing on CelebA-HQ.

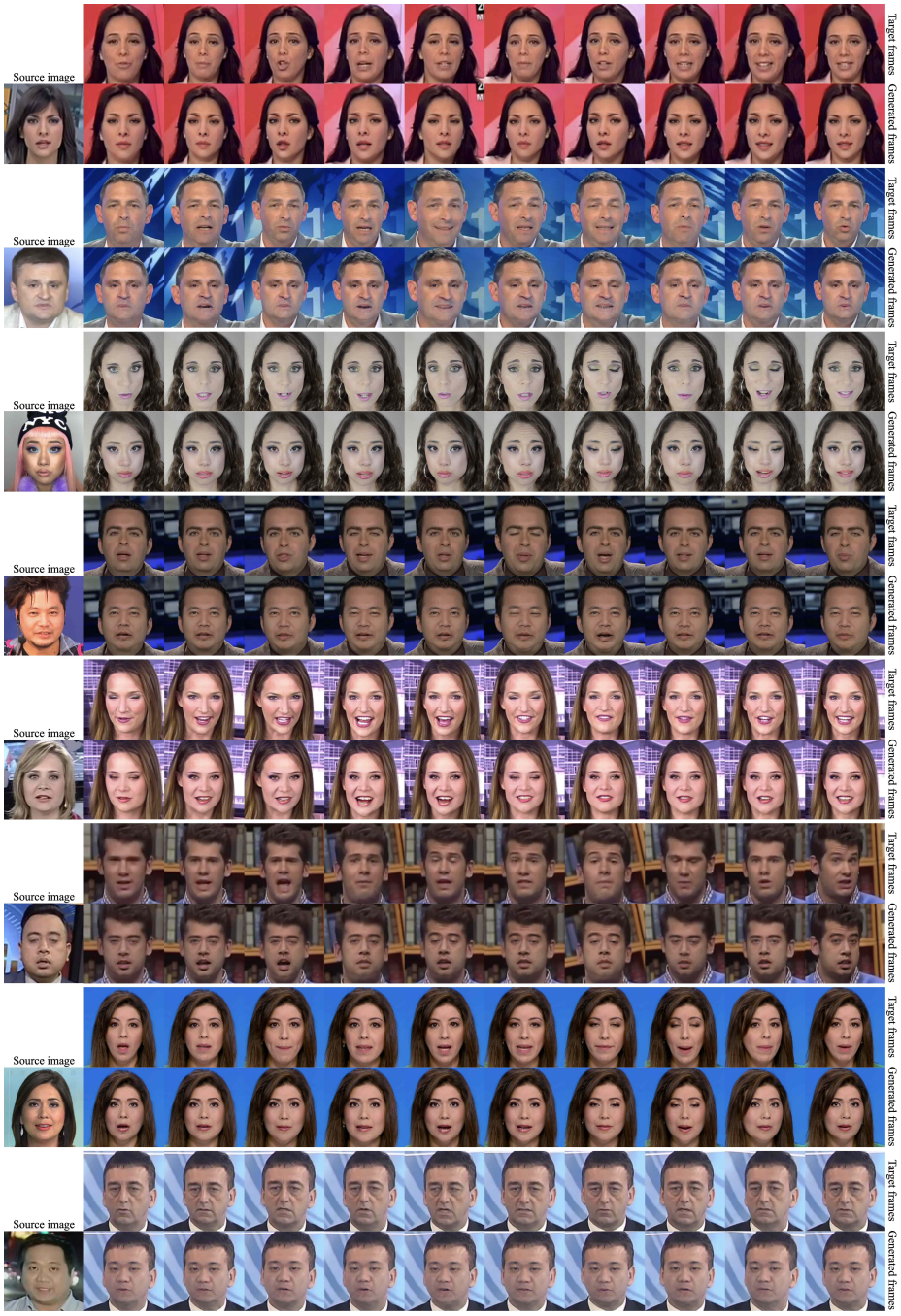


Fig. 5: **Qualitative results of face swapping on FaceForensics++.** The leftmost image is the source image, and the uppermost images are the target frames captured uniformly from the video. The rest of the images are the generated frames.

Fig. 6: **Qualitative results of face swapping on CelebA-HQ.** Our model faithfully captures ID (e.g., eyes and face shape) and ID-irrelevant (e.g., pose and expression) attributes from the source and target images, respectively, and synthesizes a high-quality megapixel image by blending these attributes.

Fig. 7: **Qualitative results of face swapping on CelebA-HQ.** Our model faithfully captures ID (e.g., eyes and face shape) and ID-irrelevant (e.g., pose and expression) attributes from the source and target images, respectively, and synthesizes a high-quality megapixel image by blending these attributes.

Fig. 8: **Qualitative results of face swapping on CelebA-HQ.** Our model faithfully captures ID (e.g., eyes and face shape) and ID-irrelevant (e.g., pose and expression) attributes from the source and target images, respectively, and synthesizes a high-quality megapixel image by blending these attributes.

Fig. 9: **Qualitative results of large-gap face swapping on CelebA-HQ.** Our model faithfully performs face swapping even with a large gap between the source and target images (e.g., gender and age).

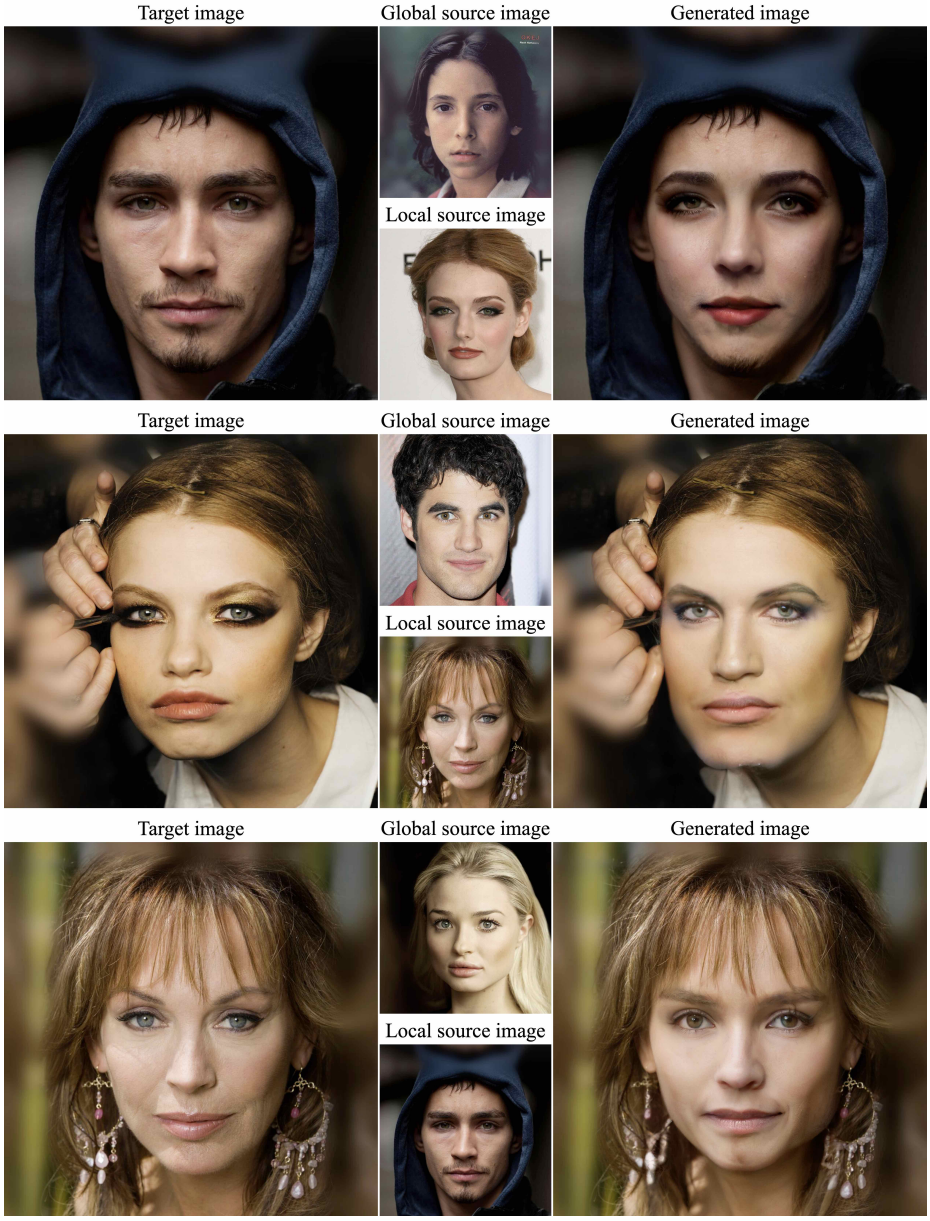


Fig. 10: **Qualitative results of ID mixing on CelebA-HQ.** Our model can create a new identity by blending the global (e.g., face shape) and local (e.g., eyes) ID attributes captured from the global and local source images, respectively.

References

1. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: How to embed images into the stylegan latent space? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4432–4441 (2019)
2. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3D face model from in-the-wild images. vol. 40 (2021), https://doi.org/10.1145/3450626.3459936
3. Gao, G., Huang, H., Fu, C., Li, Z., He, R.: Information bottleneck disentanglement for identity swapping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3404–3413 (2021)
4. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
5. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
6. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020)
7. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Faceshifter: Towards high fidelity and occlusion aware face swapping. arXiv preprint arXiv:1912.13457 (2019)
8. Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J.: On the variance of the adaptive learning rate and beyond. In: Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020) (April 2020)
9. Mescheder, L., Geiger, A., Nowozin, S.: Which training methods for gans do actually converge? In: International conference on machine learning. pp. 3481–3490. PMLR (2018)
10. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2287–2296 (2021)
11. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1–11 (2019)
12. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
13. Wang, Y., Chen, X., Zhu, J., Chu, W., Tai, Y., Wang, C., Li, J., Wu, Y., Huang, F., Ji, R.: Hififace: 3d shape and semantic prior guided high fidelity face swapping. arXiv preprint arXiv:2106.09965 (2021)
14. Wu, Z., Lischinski, D., Shechtman, E.: Stylespace analysis: Disentangled controls for stylegan image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12863–12872 (2021)
15. Zhang, M., Lucas, J., Ba, J., Hinton, G.E.: Lookahead optimizer: k steps forward, 1 step back. Advances in Neural Information Processing Systems **32** (2019)
16. Zhu, Y., Li, Q., Wang, J., Xu, C.Z., Sun, Z.: One shot face swapping on megapixels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4834–4844 (2021)