FlowFace: Semantic Flow-guided Shape-aware Face Swapping

Hao Zeng[1], Wei Zhang[1], Changjie Fan[1], Tangjie Lv[1], Suzhen Wang[1],
Zhimeng Zhang[1], Bowen Ma[1], Lincheng Li[1], Yu Ding[1,3,*], Xin Yu[2]

[1]Virtual Human Group, Netease Fuxi AI Lab
[2]University of Technology Sydney
[3]Zhejiang University
zenghao1110@gmail.com, {zhengwei05, fanchangjie, hzlvtangjie, wangsuzhen,
zhangzhimeng, mabowen01, lilincheng, dingyu01}@corp.netease.com, xin.yu@uts.edu.au

Abstract

In this work, we propose a semantic flow-guided two-stage framework for shape-aware face swapping, namely Flow-Face. Unlike most previous methods that focus on transferring the source inner facial features but neglect facial contours, our FlowFace can transfer both of them to a target face, thus leading to more realistic face swapping. Concretely, our FlowFace consists of a face reshaping network and a face swapping network. The face reshaping network addresses the shape outline differences between the source and target faces. It first estimates a semantic flow (*i.e.*, face shape differences) between the source and the target face, and then explicitly warps the target face shape with the estimated semantic flow. After reshaping, the face swapping network generates inner facial features that exhibit the identity of the source face. We employ a pre-trained face masked autoencoder (MAE) to extract facial features from both the source face and the target face. In contrast to previous methods that use identity embedding to preserve identity information, the features extracted by our encoder can better capture facial appearances and identity information. Then, we develop a cross-attention fusion module to adaptively fuse inner facial features from the source face with the target facial attributes, thus leading to better identity preservation. Extensive quantitative and qualitative experiments on in-the-wild faces demonstrate that our FlowFace outperforms the state-of-the-art significantly.

Introduction

Face swapping refers to transferring the identity information of a source face to a target face while maintaining the attributes (e.g., expression, pose, hair, lighting, and background) of the target. It has attracted many interests due to its wide applications, such as portrait reenactment, film production, and virtual reality.

Recent works (Li et al. 2019; Chen et al. 2020; Xu et al. 2021; Li et al. 2021) have made great efforts to achieve promising face swapping results. However, these methods often focus on inner facial feature transferring but neglect facial contour reshaping. We observe that facial contours also carry the identity information of a person, but few efforts (Wang et al. 2021) have been made on facial contours

transferring. Facial shape transferring is still a challenge for authentic face swapping.

To solve the shape transferring problem, we propose a semantic flow-guided two-stage framework, dubbed FlowFace. Unlike existing methods, FlowFace is a shape-aware face swapping network. In a nutshell, we first present a face reshaping network to warp the target face referring to the source face shape at the first stage. Then, based on the reshaped facial contour, we further employ a face swapping network to transfer the inner facial features to the reshaped target face.

Our face reshaping network addresses the shape outline discrepancy between the source face and the target face. Specifically, we use a 3D face reconstruction model (*i.e.*, 3DMM (Blanz and Vetter 1999)) to obtain shape coefficients of the source and target faces and then project the obtained 3D shapes to 2D facial landmarks. To accurately warp the target face, we need to obtain dense motion between the source and the target faces. Subsequently, we design a semantic guided generator to transform the sparse 2D facial landmarks into the dense flow. The estimated flow, called semantic flow, will be exploited to warp the target face shape explicitly in a pixel-wise manner. In addition, we propose a semantic-guided discriminator to enforce our face reshaping network to produce accurate semantic flow.

After reshaping the target face, we introduce a face swapping network for transferring the inner facial features of the source face to the target ones. Prior works usually use a face recognition model to extract the identity embedding of the source face and then transfer it to the target face. We argue this would lose some personalized appearances during transferring because the identity embedding is often trained under discriminative tasks and thus may ignore intra-class variations (Kim, Lee, and Zhang 2022). Thus, we opt to employ a pre-trained masked autoencoder (MAE) (He et al. 2022) to extract facial features that better capture facial appearances and identity information. Moreover, unlike prior arts that widely employ AdaIN (Liu et al. 2017b) to infuse the source identity embedding to the target face, we develop a cross-attention fusion module to adaptively fuse the source and target features. In doing so, we achieve better face swapping performance.

Extensive quantitative and qualitative experiments vali-

---

date the effectiveness of our FlowFace on in-the-wild faces and our FlowFace outperforms the state-of-the-art. Overall, our contributions are summarized as follows:

- We propose a two-stage framework for shape-aware face swapping, namely FlowFace. It can effectively transfer both the inner facial features and the facial outline to a target face, thus achieving authentic face swapping results.
- We design a semantic flow-guided face reshaping network and validate its effectiveness in transferring the source face shapes to the target ones. The reshaped target faces are more similar to the source faces in terms of face contours.
- We design a pre-trained face masked autoencoder based face swapping network. The encoder captures not only identity information but also facial appearance, thus allowing us to transfer richer information from the source face to the target and achieve identity similarity.
- We design a cross-attention fusion module to adaptively fuse the source and target features. To the best of our knowledge, we are the first to perform face swapping in the latent space of the pre-trained masked autoencoder.

Related Work

The previous face swapping methods can be classified as the target attribute-guided and source identity-guided methods.

Target attribute-guided methods edit the source face first and then blend it to the target background. Early methods (Bitouk et al. 2008; Chen et al. 2019; Lin et al. 2012) directly warp the source face according to the target facial landmarks, thus failing to address large posture differences and expression differences. 3DMM-based methods (Blanz et al. 2004; Thies et al. 2016; MarekKowalski 2021; Nirkin et al. 2018) swap faces by 3D-fitting and re-rendering. However, these methods often cannot handle skin color or lighting differences and suffer from poor fidelity. Later, GAN-based methods improve the fidelity of the generated faces. Deepfakes (DeepFakes 2019) transfers the target attributes to the source face by an encoder-decoder structure while being constrained by two specific identities. FSGAN(Nirkin, Keller, and Hassner 2019) employs the target facial landmarks to animate the source face and proposes a blending network to fuse the generated source face to the target background. However, it fails to tackle drastic skin color differences. Later, AOT (Zhu et al. 2020) focuses on swapping faces with large differences in skin color and lighting by formulating appearance mapping as an optimal transport problem. These methods always need a facial mask to blend the generated face with the target background. However, the mask-guided blending restricts the face shape change.

Source identity-guided methods usually adopt the identity embedding or the latent representation of Style-GAN2 (Karras et al. 2020) to represent the source identity and inject in into the target face. FaceShifter (Li et al. 2019) designs an adaptive attentional denormalization generator to integrate the source identity embedding and the target features. SimSwap (Chen et al. 2020) introduces a weak feature matching loss to help preserve the target attributes.

MegaFS (Zhu et al. 2021), RAFSwap (Xu et al. 2022a) and HighRes (Xu et al. 2022b) utilize the pre-trained StyleGAN2 to swap faces and can achieve high-resolution face swapping. FaceController (Xu et al. 2021) exploits the identity embedding with 3D priors to represent the source identity and design a unified framework for identity swapping and attribute editing. InfoSwap (Gao et al. 2021) leverages the information bottleneck principle to disentangle the identity and identity-irrelevant information. FaceInpainter (Li et al. 2021) also utilizes the identity embedding with 3D priors to implement controllable face in-painting under heterogeneous domains. Smooth-Swap (Kim, Lee, and Zhang 2022) builds smooth identity embedding that makes the training of face swapping fast and stable.

However, most of these methods neglect the facial outlines during face swapping. Recently, HifiFace (Wang et al. 2021) can control the face shape using a 3D shape-aware identity. However, it injects the shape representation into the latent feature space, making it hard for the model to correctly decode the face shape. Moreover, these methods always need a pre-trained face recognition model during the inference time, which is not friendly to deployment.

Proposed Method

The face swapping task aims to generate a face with the identity of the source face and the attributes of the target face. This paper proposes a semantic flow-guided two-stage framework for shape-aware face swapping, namely Flow-Face. As shown in Figure 1, FlowFace consists of a face reshaping network F^{res} and a face swapping network F^{swa}. Let I_s be the source face and I_t be the target face. F^{res} first transfers the shape of I_s to the target face I_t. The reshaped image is denoted as I_t^{res}. Then F^{swa} generates the inner face of I_t^{res} and outputs the result image I_o.

Face Reshaping Network

We design the face reshaping network, F^{res}, to address the shape discrepancy between the source and target faces. It warps the target face shape explicitly pixel-wise with an estimated semantic flow. To achieve this goal, F^{res} requires a face shape representation that models the shape differences between the source and target faces. Then it estimates a semantic flow according to the above shape differences. Finally, the semantic flow is used to warp the target face shape.

Face Shape Representation. Since our face reshaping network needs to warp the face shape pixel-wisely, we choose the explicit facial landmarks as the shape representation. We use a 3D face reconstruction model to obtain facial landmarks. As shown in Figure 1, the 3D face reconstruction model E_{3D} extracts 3D coefficients of the source and target:

$$(\beta_*, \theta_*, \psi_*, c_*) = E_{3D}(I_*), \qquad (1)$$

where $\beta_*, \theta_*, \psi_*, c_*$ are the FLAME coefficients (Li et al. 2017) representing the face shape, pose, expression, and camera, respectively. $*$ is s or t, representing the source or the target, respectively. With these coefficients, the target face can be modeled as:

$$M_t(\beta_t, \theta_t, \psi_t) = W\left(T_P(\beta_t, \theta_t, \psi_t), \mathbf{J}(\beta_t), \theta_t, \mathcal{W}\right), \quad (2)$$
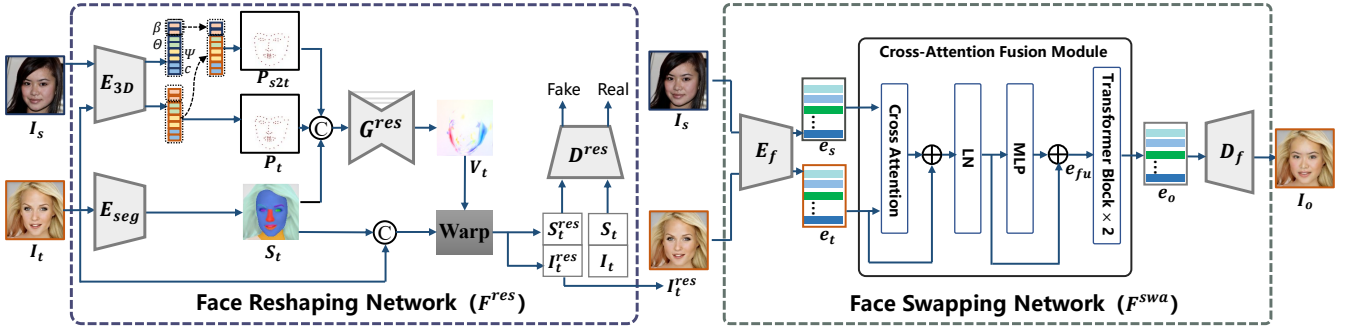
Figure 1: Overview of our two-stage FlowFace. In the first stage, the face reshaping network (F^{res}) transfers the shape of the source face I_s to the target face I_t by warping I_t explicitly with an estimated semantic flow V_t. In the second stage, the face swapping network (F^{swa}) generates the inner facial details by manipulating the latent face representation e_s and e_t using our designed cross-attention fusion module. It should be noted that ⓒ in the figure represents the concatenation operation.

where M_t represents the 3D face mesh of the target face. W is a linear blend skinning (LBS) function that is applied to rotate the vertices of T_P around joint J. $\mathcal{W}$ is the blend weights. T_P denotes the template mesh $\overline{T}$ with shape, pose, and expression offsets (Li et al. 2017).

Then, we reconstruct the source face similarly, except that the source pose and expression coefficients are replaced with the target ones. The obtained 3D face mesh is denoted as M_{s2t}. Finally, we sample 3D facial landmarks from M_t and M_{s2t} and project these 3D points to 2D facial landmarks with the target camera parameter c_t:

$$
\begin{aligned}
P_t &= s\Pi\left(M_t^i\right) + t, \\
P_{s2t} &= s\Pi\left(M_{s2t}^i\right) + t,
\end{aligned}
\quad (3)
$$

where M_*^i is a vertex in M_*, Π is an orthographic 3D-2D projection matrix, and s and t are parameters in c_t, indicating isotropic scale and 2D translation. P_* denotes the 2D facial landmarks. It should be noted that we only use the landmarks at the facial contours as the shape representation since inner facial landmarks contain identity information that may influence the reshaping result.

Semantic Flow Estimation. The relative displacement between P_t and P_{s2t} only describes sparse movement. To accurately warp the target face, we need to obtain dense motion between the source and the target faces. Therefore, we propose the semantic flow, which models the semantic correspondences between two faces, to achieve pixel-wised movement. We design a semantic guided generator G^{res} to estimate the semantic flow. Specifically, G^{res} requires three inputs: P_{s2t}, P_t and S_t, where P_{s2t} and P_t are the 2D facial landmarks obtained above. S_t is the target face segmentation map that complements the semantic information lost in facial landmarks. The output of G^{res} is the estimated semantic flow V_t, the formulation is:

$$
V_t = G^{res}(P_{s2t}, P_t, S_t). \quad (4)
$$

Then, a warping module is introduced to generate the warped faces using V_t. We find that an inaccurate flow is likely to produce unnatural images, and therefore, we design

a semantic guided discriminator D^{res} that ensures G^{res} to produce a more accurate flow. Specifically, the warping operation is conducted on both I_t and S_t:

$$
(I_t^{res}, S_t^{res}) = F(V_t, I_t, S_t), \quad (5)
$$

where F is the warping function in the warping module. We feed the concatenation of the warped face I_t^{res} and the warped segmentation map S_t^{res} to D^{res}. Thus, D^{res} is able to discriminate whether the input is real or fake from the semantic level and the image level. It should be noted that S_t^{res} and D^{res} are only used during training.

Training Loss. We employ three loss functions for F^{res}:

$$
\mathcal{L}^{res} = \mathcal{L}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{ldmk}\mathcal{L}_{ldmk}, \quad (6)
$$

where λ_{ldmk} and λ_{rec} are hyperparameters for each term. In our experiments, we set λ_{ldmk}=800 and λ_{rec}=10.

Adversarial Loss. To make the resultant images more realistic, we adopt the hinge version adversarial loss(Lim and Ye 2017) for training, denoted by L_{adv}:

$$
\mathcal{L}_{adv} = -\mathbb{E}[D^{res}([I_t^{res}, S_t^{res}])], \quad (7)
$$

where D^{res} is the discriminator which is trained with:

$$
\begin{aligned}
\mathcal{L}_D &= \mathbb{E}[\max(0, 1 - D([I_t, S_t]))] \\
&+ \mathbb{E}[\max(0, 1 + D([I_t^{res}, S_t^{res}]))].
\end{aligned}
\quad (8)
$$

Reconstruction Loss. Since there is no ground-truth for face reshaping results, we enforce $I_s = I_t$ with a certain probability when training G^{res}. Then the face reshaping task becomes a reconstruction task, and we introduce a pixel-wise reconstruction loss:

$$
\mathcal{L}_{rec} = \|I_t^{res} - I_t\|_2, \quad (9)
$$

where $\|*\|_2$ denotes the euclidean distance.

Landmark Loss. Since there is not pixel-wised ground truth for I_t^{res}, we exploit the 2D facial landmarks P_{s2t} to constrain the shape of I_t^{res}. Specifically, we first use a pre-trained facial landmark detector (Sun et al. 2019) to predict the the facial landmarks of I_t^{res}, denoted as P_t^{res}. Then the loss is computed as:

$$
\mathcal{L}_{ldmk} = \|P_t^{res} - P_{s2t}\|_2. \quad (10)
$$

At this point, our designed face reshaping network is able to transfer the face shape of the source to the target face. However, the inner facial features are still unchanged.

Face Swapping Network

The face swapping network F^{swa} is used to generate the inner face of I_t^{res} (I_t). As shown in Figure 1, we first utilize a shared face encoder E_f to map both I_s and I_t^{res} into patch embeddings e_s nad e_t. Then a cross-attention fusion module is designed to adaptively fuse the identity information of the source face and the attribute information of the target. Finally, the facial decoder, fed with the manipulated embeddings e_o, outputs the final face swapping result I_o.

Shared Face Encoder. Most previous face swapping methods map the source face into an ID embedding with a pre-trained face recognition model and extract the target face attributes with another face encoder. However, we argue that using two different encoders is unnecessary and even makes deploying more complex. Moreover, the ID embedding is trained on purely discriminative tasks and may lose some personalized appearances during transferring.

Therefore, we employ a shared encoder to project both the source face and the target face into a common latent representation. The encoder is designed following MAE (He et al. 2022) and pre-trained on a large-scale face dataset using the masked training strategy. Compared to the compact latent code of StyleGAN2 (Karras et al. 2020) and the identity embedding, the latent space of MAE can better capture facial appearances and identity information, because masked training requires reconstructing masked image patches from visible neighboring patches, thus ensuring each patch embedding contains rich topology and semantic information.

Based on the pre-trained encoder E_f, we can project a facial image I_* into a latent representation, also known as patch embeddings:

$$e_* = E_f(I_*), \qquad (11)$$

where $e_* \in R^{N*L}$. N and L denote the number of patches and the dimension of each embedding, respectively.

Cross-Attention Fusion Module. The shared face encoder projects the source face and the target face into a representative latent space. The subsequent operation is to fuse the source identity information with the target attribute in this latent space. Intuitively, identity information should be transferred between related patches (*e.g.*, nose to nose, etc.). Therefore, we design a cross-attention fusion module (CAFM) to adaptively aggregate identity information from the source and fuse it into the target.

As shown in Figure 1, our CAFM consists of a cross-attention block and two standard transformer blocks (Dosovitskiy et al. 2020). Given the source patch embeddings e_s and the target patch embeddings e_t, we first compute Q, K, V for each patch embedding in e_s and e_t. Then the cross attention is computed by:

$$CA(Q_t, K_s) = \text{softmax}\left(\frac{Q_t K_s^T}{\sqrt{d_k}}\right), \qquad (12)$$

where CA represents Cross Attention, Q_*, K_*, V_* are predicted by attention heads, and d_k is the dimension of K_*. The cross attention describes the relation between each target patch and the source patches. Next, the source identity

information is aggregated based on the computed CA and fused to the target values via addition:

$$V_{fu} = CA * V_s + V_t. \qquad (13)$$

Then, V_{fu} are normalized by a layer normalization (LN) and processed by multi-layer perceptrons (MLP). The Cross Attention and MLP are along with skip connections. The fused embeddings e_{fu} are further fed into two transformer blocks to obtain the final output e_o.

Finally, we utilize a a convolutional decoder to generate the final swapped face image I_o from e_o. In contrast to the ViT decoder in MAE, we find the convolutional decoder achieves more realistic results.

Training Loss. We employ six loss functions to train our face swapping network F^{swa}:

$$\begin{aligned} \mathcal{L}^{swa} = \mathcal{L}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{id}\mathcal{L}_{id} + \lambda_{exp}\mathcal{L}_{exp} \\ + \lambda_{ldmk}\mathcal{L}_{ldmk} + \lambda_{perc}\mathcal{L}_{perc}, \end{aligned} \qquad (14)$$

where λ_{rec}, λ_{id}, λ_{exp}, λ_{ldmk}, λ_{attr} are hyperparameters for each term. In our experiment, we set λ_{rec}=10, λ_{id}=5, λ_{exp}=10, λ_{ldmk}=5000 and λ_{attr}=2.

As in the face reshaping stage, the adversarial loss is used to make the resultant images more realistic, and the reconstruction loss between I_o and I_t^{res} is used for self-supervision since there is also no ground-truth for face swapping results.

Identity Loss. The identity loss is used to improve the identity similarity between I_s and I_o:

$$\mathcal{L}_{id} = 1 - cos(E_{id}(I_o), E_{id}(I_s)), \qquad (15)$$

where E_{id} denotes a face recognition model (Deng et al. 2019) and cos denotes the cosine similarity.

Posture Loss. We adopt the landmark loss to constrain the face posture during face swapping:

$$\mathcal{L}_{ldmk} = \|P_t^{res} - P_o\|_2, \qquad (16)$$

where P_o represents the landmarks of I_o.

Perceptual Loss. Since high-level feature maps contain semantic information, we employ the feature maps from the last two convolutional layers of pre-trained VGG as the facial attribute representation. The loss is formulated as:

$$\mathcal{L}_{perc} = \|VGG(I_t^{res}) - VGG(I_o)\|_2. \qquad (17)$$

Expression Loss. We utilize a novel fine-grained expression loss (Zhang et al. 2021) that penalizes the $\mathcal{L}_2$ distance of two expression embeddings:

$$\mathcal{L}_{exp} = \|E_{exp}(I_o) - E_{exp}(I_t)\|_2. \qquad (18)$$

Experiments

Our method is validated through qualitative and quantitative comparisons with state-of-the-art ones and a user study. Moreover, several ablation experiments are also reported to validate our design of FlowFace.
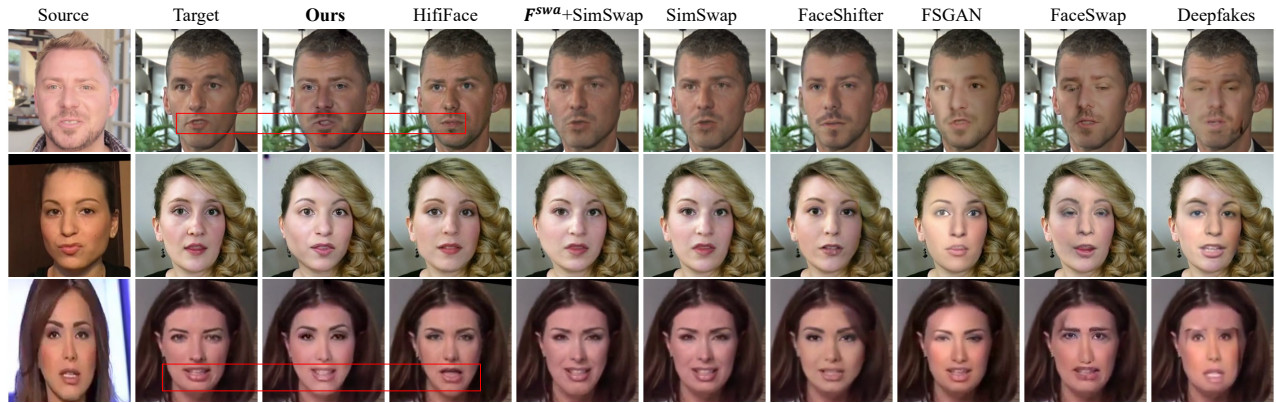
Figure 2: Qualitative comparisons with Deepfakes, FaceSwap, FSGAN, FaceShifter, SimSwap and HifiFace on FF++. Our FlowFace outperforms the other methods significantly, especially in preserving face shapes, identities, and expressions.

Table 1: Quantitative comparisons with state-of-the-art methods on FF++. "†" means the results are cited from their papers.

Methods	ID Acc(%) ↑			Shape↓	Expr.↓	Pose.↓
	CosFace	SphereFace	Avg			
Deepfakes	83.55	86.60	85.08	1.78	0.54	4.05
FaceSwap	70.95	76.77	73.86	1.85	0.40	2.21
FSGAN	48.86	53.85	51.36	2.18	0.27	2.20
FaceShifter	97.38†	80.64	89.01	1.68	0.33	2.28
SimSwap	93.63	96.22	94.43	1.74	0.26	**1.40**
F^{res}+SimSwap	94.31	96.82	95.56	1.67	0.27	2.27
HifiFace	98.48†	90.76	94.62	1.62	0.30	2.29
F^{swa}	99.18	98.23	98.70	1.43	**0.21**	1.99
Ours	**99.26**	**98.40**	**98.83**	**1.17**	0.22	2.66

Implementation Details

Dataset. The training dataset is collected from three commonly-used face datasets: CelebA-HQ (Karras et al. 2017), FFHQ (Karras, Laine, and Aila 2019), and VG-GFace2 (Cao et al. 2018). Faces are aligned and cropped to 256×256. Particularly, low-quality faces are removed to ensure high-quality training. The final dataset contains 350K face images, and 10K images are randomly sampled as the validation dataset. For the comparison experiments, we construct the test set by sampling FaceForensics++(FF++) (Rössler et al. 2019), following (Li et al. 2019). Specifically, FF++ consists of 1000 video clips, and the test set is collected by sampling ten frames from each clip of FF++, in a total of 10000 images.

Training. Our FlowFace is trained in a two-stage manner. Specifically, F^{res} is first trained for 32K steps with a batch size of eight. As for F^{swa}, we first pre-trained the face encoder following the training strategy of MAE on our face dataset. Then we fix the encoder and train other components of F^{swa} for 640K steps with a batch size of eight. We adopt Adam (Kingma and Ba 2014) optimizer with β_1=0 and β_2=0.99 and the learning rate is set to 0.0001. More details are in the supplementary materials and our codes will be made publicly available upon publication of the paper.

Metrics. The quantitative evaluations are performed in terms of four metrics: identity retrieval accuracy (ID Acc), shape error, expression error (Expr Error), and pose error. We follow the same test protocol in (Li et al. 2019; Wang

et al. 2021). However, since some pre-trained models used in their testing are not available, we leverage different ones. For ID Acc, we employ two face recognition models, including CosFace (Wang et al. 2018) and SphereFace (Liu et al. 2017a), to perform identity retrieval for a more comprehensive comparison. For expression error, we adopt a different expression embedding model (Vemulapalli and Agarwala 2019) to compute the euclidean distance of expression embeddings between the target and swapped faces.

Comparisons with the State-of-the-art

Quantitative Comparisons. Our method is compared with six methods including Deepfakes (DeepFakes 2019), FaceSwap (MarekKowalski 2021), FSGAN (Nirkin, Keller, and Hassner 2019), FaceShifter (Li et al. 2019), Sim-Swap (Chen et al. 2020), and HifiFace (Wang et al. 2021). For Deepfakes, FaceSwap, FaceShifter, and HifiFace, we use their released face swapping results of the sampled 10,000 images. For FSGAN and SimSwap, the face swapping results are generated with their released codes.

Table 1 shows that our method achieves the best scores under most evaluation metrics, including ID Acc, shape error, and Expr Error. These results validate the superiority of our FlowFace. We obtain a slightly worse result than other methods for the pose error, which can be attributed to our FlowFace changing the face shape while the employed head pose estimator is sensitive to face shapes.

Qualitative Comparisons. The qualitative comparisons are conducted on the same FF++ test set collected in the quantitative comparisons. As shown in Figure 2, our Flow-Face maintains the best face shape consistency. Note that most methods do nothing to transfer the face shape, so their resulting face shapes are similar to the target ones.

Although HifiFace is specifically designed to change the face shape, our method still obtains better results. As observed in Figure 2, our generated face shapes are more similar to the source ones than HifiFace. Since HifiFace injects the shape representation into the latent feature space, it is harder to accurately decode the face shape from the latent feature than our explicit semantic flow. Meanwhile, our
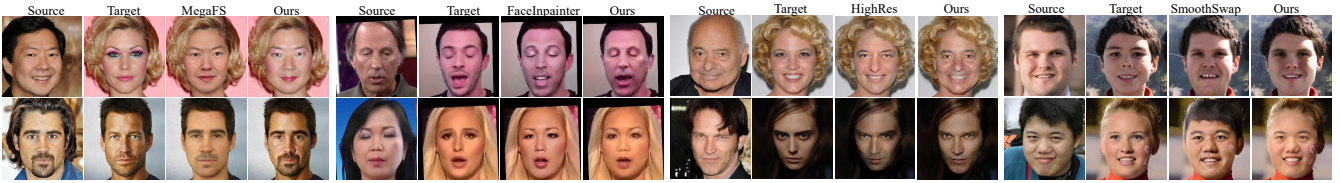
Figure 3: Qualitative comparisons with more methods including MegaFS (Zhu et al. 2021), FaceInpainter (Li et al. 2021), HighRes (Xu et al. 2022b) and SmoothSwap (Kim, Lee, and Zhang 2022). The shown images of the compared methods are cropped from their original papers or their released results.
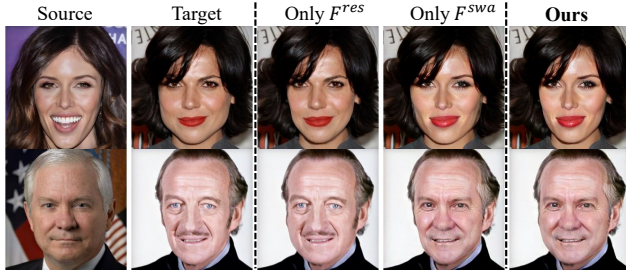

Figure 4: Qualitative ablation results of FlowFace.

Table 2: Subjective comparisons with SimSwap and Hifi-Face on FF++.

Method	Shape. (%)↑	ID. (%)↑	Exp. (%)↑	Realism (%)↑
SimSwap	20.67	27.78	34.44	14.67
HifiFace	35.11	34.67	30.45	41.78
Ours	**44.22**	**37.55**	**35.11**	**43.55**

method can better preserve the fine-grained target expressions (marked with red boxes in rows 1,3).

We further compare our methods with four more SOTA face swapping methods: (Zhu et al. 2021), FaceInpainter (Li et al. 2021), HighRes (Xu et al. 2022b) and Smooth-Swap (Kim, Lee, and Zhang 2022). Among them, MegaFS and HighRes are based-on the latent space of StyleGAN2. As shown in Figure 3, our method can better transfer the shape of the source to the target than all other methods. Although SmoothSwap can change the face shape, it destroys the target attributes (*e.g.*, hairstyle and hair color). Besides, our results are also more similar to the source face in terms of inner facial features (*e.g.*, beard), validating that our face encoder can better capture facial appearances than the identity embedding or the latent code of StyleGAN2. Moreover, our method also preserves the target attributes (*e.g.*, skin color, lighting, and expression) better than other methods.

User Study. To further validate our FlowFace, we conduct a subjective comparison with SimSwap and HifiFace, two SOTA methods that release their codes or results. Fifteen participants are instructed to choose the best result in terms of shape consistency, identity consistency, expression consistency, or image realism, involving comparisons of 30 swapped faces by three methods. Table 2 shows that our method outperforms the two baselines in terms of all four metrics, validating the superiority of our method.

Analysis of FlowFace

Three ablation studies are conducted to validate our two-stage FlowFace framework and several components used in F^{res} and F^{swa}, respectively.

Ablation Study on FlowFace. We conduct ablation experiments to validate the design of our two-stage framework. Figure 4 shows the swapped images by only F^{res}, only F^{swa} and the full model (FlowFace). It can be seen that F^{res} transforms the face shape naturally according to the source, while F^{swa} is good at capturing the identity of the source inner face and other facial attributes of the target. Benefiting from the strengths of both F^{res} and F^{swa}, our FlowFace is able to create results with accurate identity and consistent facial attributes. Table 1 records the quantitative results, which further illustrates the effectiveness of our two-stage framework. The above observation validates the effectiveness of F^{res} and confirms that face shapes are essential for identifying a person.

To further validate our F^{res}, we plug it into the open-sourced SimSwap. As shown in Figure 2 and Table 1, after reshaping by F^{res}, the face swapping result of SimSwap is more similar to the source face in terms of face contours. The ID Acc. also rises from 93.63% to 94.31%. The results demonstrate the effectiveness of our F^{res} and also reveal that the face shape carries the identity information, thus improving identity similarity.

Ablation study on F^{res}. We first conduct an ablation experiment to validate our proposed semantic guided generator G^{res}. Specifically, we remove the semantic input S_t of G^{res} (G^{res} w/o Seg). It can be seen from Figure 5 that some inaccurate flow occurs in the generated face, which implies that only facial landmarks cannot guide G^{res} to produce accurate dense flow due to the lack of semantic information. The results also demonstrate that the semantic information is beneficial for accurate flow estimation and validates the effectiveness of G^{res}.

Then, we conduct two ablation experiments to validate D^{res}: (1) removing the semantic inputs (S_t and S_t^{res}) of D^{res} (D^{res} w/o Seg). Compared with F^{res}, the generated faces suffer from unnaturalness, like the eyes are stretched, as observed in Figure 5. It implies that structured information in the semantic inputs can provide more fine-grained discriminative signals, thus enforcing G^{res} to produce a more accurate flow. (2) removing D^{res} (w/o D^{res}). As observed in Figure 5, compared with F^{res}, there are many artifacts in the generated images, and the estimated flow also
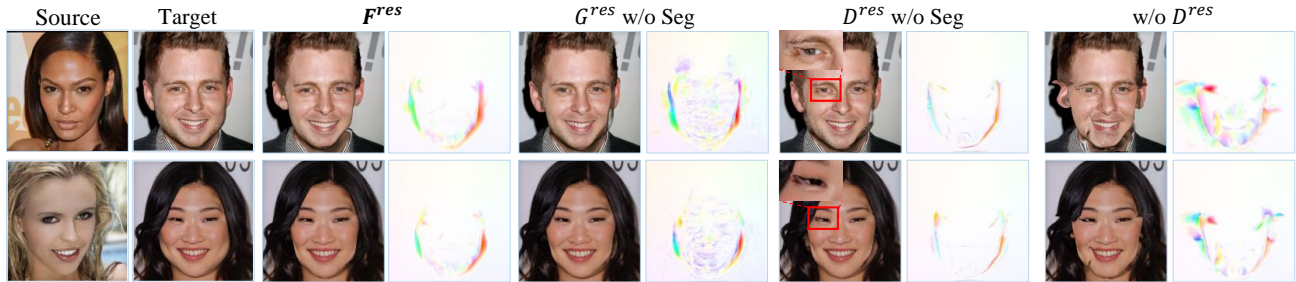
Figure 5: Qualitative ablation results of each component in F^{res}.

Figure 6: Qualitative ablation study of F^{swa}.

Table 3: Quantitative ablation study of F^{swa} on FF++.

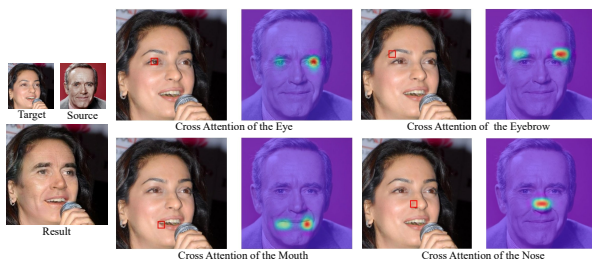| Methods	ID Acc			Expr	Pose
	CosFace	SphereFace	Avg		
Addition	**99.38**	**99.44**	**99.41**	0.43	4.90
AdaIN	97.31	97.15	97.23	0.33	3.27
Id Embed.	97.10	96.90	97.00	<u>0.22</u>	<u>2.10</u>
Vit Decoder	98.44	97.73	98.09	0.23	2.80
F^{swa}	<u>99.18</u>	<u>98.23</u>	<u>98.71</u>	**0.21**	**1.99**

Figure 7: Visualize the cross-attention of different facial parts. For each part in the target, our CAFM can accurately focus on the corresponding parts in the source.

contains many noises. The above observation validates the effectiveness of our proposed D^{res}.

Ablation study on F^{swa}. Three ablation experiments are conducted to evaluate the design of F^{swa}:

(1) Choices on CAFM, *Addition* and *AdaIN*. To verify the effectiveness of CAFM, we compare with two other methods: *Addition* that directly adds the source values to the target values; *AdaIN* that first averages source patch embeddings and then injects it into the target feature map using AdaIN residual blocks. As shown in Figure 6 and Table 3, *Addition* simply brings all information of the source face to the target face, thus leading to severe pose and expression mismatch. *AdaIN* impacts the non-face parts (*e.g.*, hair) due to its global modulation. In contrast, F^{res} with CAFM obtains a high ID Acc and preserves the target attribute well, which proves that CAFM can accurately extract identity information from the source face and adaptively infuse it into

the target counterpart.

To further validate the effectiveness of our CAFM, we visualize the cross attention computed by CAFM. As shown in Figure 7, given a specific part (marked by red boxes) of the target face, CAFM accurately focuses on the corresponding parts of the source face, validating our CAFM can adaptively transfer the identity information from the source patches to corresponding target patches.

(2) Latent Representation vs. ID Embedding (*ID Embed.*). To verify the superiority of using the latent representation of MAE, we train a new model which adopts the identity embedding as the identity representation and employs AdaIN as the injection method. As can be seen from Figure 6, *ID Embed.* misses some fine-grained face appearances, such as eyes color, beard. In contrast, F^{swa} contains richer identity information and achieves higher ID Acc, as shown in Tab 3.

(3) Convolutional Decoder vs. ViT Decoder (*ViT Decoder*). We try two different decoders to find out the better one. As shown in Figure 6, the results of *ViT Decoder* contains a lot of artifacts. In contrast, *Convolutional Decoder* achieves realistic results with high fidelity.

Conclusion

This work proposes a semantic flow-guided two-stage framework, FlowFace, for shape-aware face swapping. In the first stage, the face reshaping network transfers the shape of the source face to the target face by warping the face pixel-wisely using semantic flow. In the second stage, we employ a pre-trained masked autoencoder to extract facial features that better capture facial appearances and identity information. Then, we design a cross-attention fusion module to better fuse the source and the target features, thus leading to better identity preservation. Extensive quantitative and qualitative experiments are conducted on in-the-wild faces, demonstrating that our FlowFace outperforms the state-of-the-art significantly.

Acknowledgments

References

Bitouk, D.; Kumar, N.; Dhillon, S.; Belhumeur, P.; and Nayar, S. K. 2008. Face swapping: automatically replacing faces in photographs. In *ACM SIGGRAPH 2008 papers*, 1–8.

Blanz, V.; Scherbaum, K.; Vetter, T.; and Seidel, H.-P. 2004. Exchanging faces in images. In *Computer Graphics Forum*, volume 23, 669–676. Wiley Online Library.

Blanz, V.; and Vetter, T. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 187–194.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th international conference on automatic face & gesture recognition (FG 2018)*, 67–74. IEEE.

Chen, D.; Chen, Q.; Wu, J.; Yu, X.; and Jia, T. 2019. Face swapping: realistic image synthesis based on facial landmarks alignment. *Mathematical Problems in Engineering*, 2019.

Chen, R.; Chen, X.; Ni, B.; and Ge, Y. 2020. SimSwap: An Efficient Framework For High Fidelity Face Swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2003–2011.

DeepFakes. 2019. DeepFakes. https://github.com/deepfakes/faceswap. Online; Accessed March 1, 2021.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gao, G.; Huang, H.; Fu, C.; Li, Z.; and He, R. 2021. Information bottleneck disentanglement for identity swapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3404–3413.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.

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

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4401–4410.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.

Kim, J.; Lee, J.; and Zhang, B.-T. 2022. Smooth-Swap: A Simple Enhancement for Face-Swapping with Smoothness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10779–10788.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, J.; Li, Z.; Cao, J.; Song, X.; and He, R. 2021. FaceInpainter: High Fidelity Face Adaptation to Heterogeneous Domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5089–5098.

Li, L.; Bao, J.; Yang, H.; Chen, D.; and Wen, F. 2019. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*.

Li, T.; Bolkart, T.; Black, M. J.; Li, H.; and Romero, J. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6): 194:1–194:17.

Lim, J. H.; and Ye, J. C. 2017. Geometric gan. *arXiv preprint arXiv:1705.02894*.

Lin, Y.; Lin, Q.; Tang, F.; and Wang, S. 2012. Face replacement with large-pose differences. In *Proceedings of the 20th ACM international conference on Multimedia*, 1249–1250.

Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; and Song, L. 2017a. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 212–220.

Liu, X.; Vijaya Kumar, B.; You, J.; and Jia, P. 2017b. Adaptive deep metric learning for identity-aware facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 20–29.

MarekKowalski, M. 2021. FaceSwap. [EB/OL]. https://github.com/MarekKowalski/FaceSwap Accessed March 1, 2021.

Nirkin, Y.; Keller, Y.; and Hassner, T. 2019. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7184–7193.

Nirkin, Y.; Masi, I.; Tuan, A. T.; Hassner, T.; and Medioni, G. 2018. On face segmentation, face swapping, and face perception. In *IEEE International Conference on Automatic Face & Gesture Recognition*, 98–105. IEEE.

Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *International Conference on Computer Vision (ICCV)*.

Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; and Wang, J. 2019. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*.

Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; and Nießner, M. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2387–2395.

Vemulapalli, R.; and Agarwala, A. 2019. A compact embedding for facial expression similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5683–5692.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5265–5274.

Wang, Y.; Chen, X.; Zhu, J.; Chu, W.; Tai, Y.; Wang, C.; Li, J.; Wu, Y.; Huang, F.; and Ji, R. 2021. HifiFace: 3D Shape and Semantic Prior Guided High Fidelity Face Swapping. *arXiv preprint arXiv:2106.09965*.

Xu, C.; Zhang, J.; Hua, M.; He, Q.; Yi, Z.; and Liu, Y. 2022a. Region-Aware Face Swapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7632–7641.

Xu, Y.; Deng, B.; Wang, J.; Jing, Y.; Pan, J.; and He, S. 2022b. High-resolution face swapping via latent semantics disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7642–7651.

Xu, Z.; Yu, X.; Hong, Z.; Zhu, Z.; Han, J.; Liu, J.; Ding, E.; and Bai, X. 2021. FaceController: Controllable Attribute Editing for Face in the Wild. *arXiv preprint arXiv:2102.11464*.

Zhang, W.; Ji, X.; Chen, K.; Ding, Y.; and Fan, C. 2021. Learning a Facial Expression Embedding Disentangled From Identity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6759–6768.

Zhu, H.; Fu, C.; Wu, Q.; Wu, W.; Qian, C.; and He, R. 2020. AOT: Appearance Optimal Transport Based Identity Swapping for Forgery Detection. In *Neural Information Processing Systems (NeurIPS)*.

Zhu, Y.; Li, Q.; Wang, J.; Xu, C.-Z.; and Sun, Z. 2021. One Shot Face Swapping on Megapixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4834–4844.