Face Anonymization Made Simple

Han-Wei Kung[1] Tuomas Varanka[2] Sanjay Saha[3] Terence Sim[3] Nicu Sebe[1]

[1]University of Trento [2]University of Oulu [3]National University of Singapore

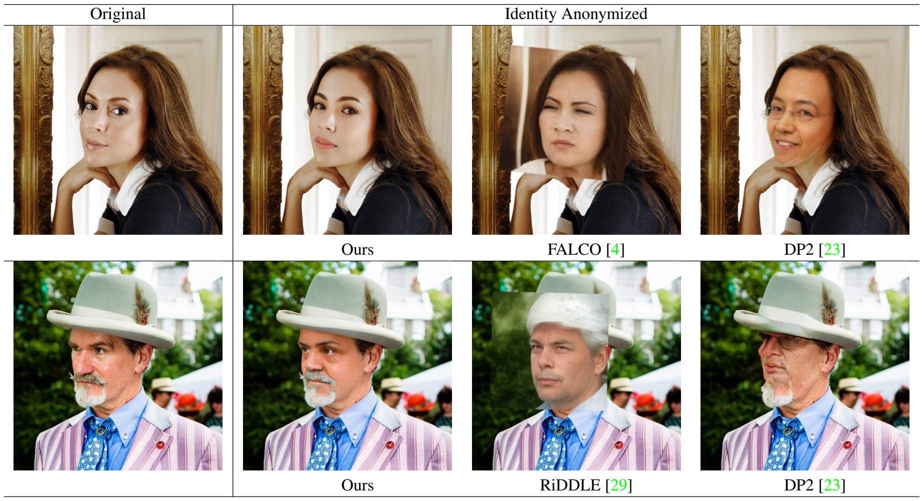

Figure 1. Our face anonymization technique preserves the original facial expressions, head positioning, eye direction, and background elements, effectively masking identity while retaining other crucial details. The anonymized face blends seamlessly into its original photograph, making it ideal for diverse real-world applications.

Abstract

Current face anonymization techniques often depend on identity loss calculated by face recognition models, which can be inaccurate and unreliable. Additionally, many methods require supplementary data such as facial landmarks and masks to guide the synthesis process. In contrast, our approach uses diffusion models with only a reconstruction loss, eliminating the need for facial landmarks or masks while still producing images with intricate, fine-grained details. We validated our results on two public benchmarks through both quantitative and qualitative evaluations. Our model achieves state-of-the-art performance in three key areas: identity anonymization, facial attribute preservation, and image quality. Beyond its primary function of anonymization, our model can also perform face swapping tasks by incorporating an additional facial image as input, demonstrating its versatility and potential for diverse applications. Our code and models are available at `https://github.com/hanweikung/face_anon_simple`.

1. Introduction

In the digital age, our identity and privacy are more vulnerable than ever. People have shared personal information and photos online over recent decades, while advancements in facial recognition technology have made it easier to identify individuals from a single image. This combination allows for the potential linking of our faces to personal information, posing a significant threat to our privacy and identity. In response, various regions have enacted privacy protection laws. These include the European Union's General Data Protection Regulation (GDPR) [1], California's Consumer Privacy Act, and Japan's amended Act on the Protection of Personal Information. Such legislation mandates that organizations implement security measures and maintain transparency in their handling of personal data.

Face anonymization is essential for protecting individuals in photos and videos, thereby reducing the risk of personal data being compromised or misused. Traditional methods like blurring and pixelation are common but have significant drawbacks. These techniques are vulnerable to reconstruction attacks [50], degrade image quality, and apply a uniform transformation across the image without considering which areas are most critical to anonymize.

These limitations make traditional methods impractical for professionals who need to preserve facial expressions and backgrounds. For example, medical practitioners may need to anonymize patient images for case studies or research while retaining crucial facial cues that indicate symptoms. In creative fields, documentary filmmakers might want to protect interviewees' privacy without losing the narrative impact of their facial expressions and reactions. They may also wish to replace an interviewee's face with a specific virtual identity to enhance storytelling clarity. In contrast, recent advances in deep learning have led to more effective anonymization techniques that enhance both privacy protection and usability. Generative Adversarial Networks (GANs) [16], in particular, can anonymize faces by replacing the original with computer-generated alternatives [10, 24, 35, 48]. However, these methods are not without challenges. Some fail to produce natural-looking faces [35], while others [24] struggle to preserve crucial elements like facial expressions, eye direction, head orientation, background details, clothing, and accessories. These limitations greatly restrict the practical application of these techniques.

This paper presents a diffusion-based method for face anonymization. Our goal is to ensure that de-identified facial images remain useful for facial analysis tasks, including pose estimation, eye-gaze tracking, and expression recognition, as well as for broader uses such as interviews and films. Therefore, we approach face anonymization similarly to face swapping, aiming to generate an image where a person's face is replaced by another person's face while maintaining the original facial expression, pose, eye gaze, and background. We designed a framework that initially performs realistic and seamless face swaps given both source and driving images. At its core is a denoising UNet architecture, similar to those used in text-to-image diffusion models, which generates the final output. We enhance this with an image feature extraction mechanism that transfers fine details from input images to the synthesized output throughout the diffusion process. The model is then trained in a dual setting: conditionally with a source image and unconditionally without a source image. This dual method allows the model to replace faces using one single image input. To create a distinct anonymized identity, the system reverses the original face's most distinctive features. This technique produces a believable anonymized face while preserving the original image's quality and essential facial characteristics.

In summary, our contributions are:

- A convenient method that produces realistic anonymized faces while preserving attributes, without needing external data like facial landmarks or masks as required by existing techniques.
- A diffusion-based network that achieves good performance with a single, simple loss function, in contrast to GAN-based models requiring multiple, carefully designed loss functions.
- Simple control of the anonymization level using a single parameter.
- Versatility beyond anonymization, including the ability to perform face swapping tasks with an additional facial image input.

2. Related Work

Face Anonymization. Most deep learning-based image anonymization methods have been developed using GANs and target not only faces [4, 10, 11, 17, 19, 24, 29, 42, 48, 49, 53, 56] but also bodies [9, 23] and other objects [46]. In this study, we focus on face anonymization.

Many GAN-based face anonymization methods use conditional GANs as their foundation. These techniques typically require supplementary data to create anonymized faces. For example, IDeudemon [53] uses face parsing maps or masks to segment image components, while Sun *et al.* [48]'s method employs facial landmarks to guide face inpainting. CIAGAN [35] relies on masks and facial landmarks, and DeepPrivacy [24] utilizes bounding boxes and facial landmarks. These methods depend on additional information, which can be a limitation if the required data are missing or flawed. In contrast, our approach does not rely on such auxiliary data to anonymize faces.

Other techniques like RiDDLE [29] and FALCO [4] use GAN inversion. They map facial images to the latent space of a pre-trained StyleGAN2 [28], leveraging its capabilities to produce high-quality images. However, these techniques may inadvertently alter important identity-irrelevant details such as facial expressions, background, body parts, and accessories. Our method treats face anonymization similarly to face swapping and incorporates image feature extraction networks to capture detailed input features. This allows us to generate anonymized faces that seamlessly integrate with the existing image while preserving the overall integrity of the image.

StyleFace [34] embeds identity vectors from a pretrained face recognition network into the StyleGAN2 [28] model's latent space, sampling random vectors for anonymization. While this approach generates realistic faces, it risks revealing the original identity if the sampled vector is too close to the original identity. In contrast, our model offers an adjustable anonymization degree, allowing users to control the distance between the input and generated images for effective anonymization.

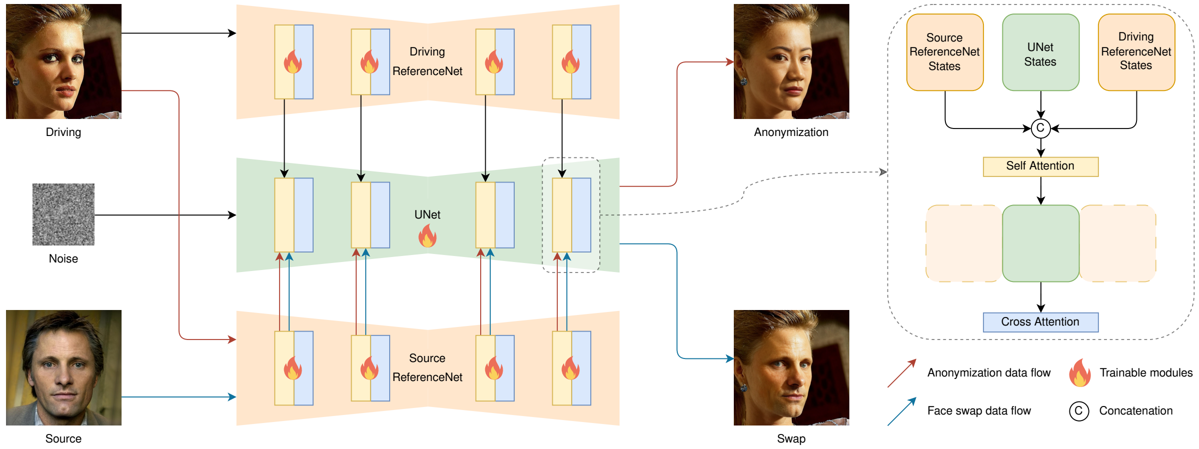

Figure 2. Our network leverages the face swapping mechanism for face anonymization. In both cases, the system encodes source and driving images into latent space and processes them through respective ReferenceNet models. These images are also encoded into intermediate embeddings that guide the UNet via cross-attention. The UNet incorporates states from both ReferenceNet models through concatenation, enabling the transfer of details from source and driving images through self-attention. Using these learned features and intermediate embeddings, the UNet generates the output image. For face anonymization, we use the same image as both source and driving input. However, we modify the intermediate embedding and state from the source ReferenceNet model to achieve the desired anonymization effect.

Face Swapping. Face swapping techniques can be categorized into two main approaches: source-oriented and target-oriented methods.

Source-oriented methods [31,36,55] begin by transforming the source face to match the expression, pose, and lighting of the target face, and then replace the target image with this modified source face. For example, FSGAN [36] employs a two-stage process: it first uses a reenactment network for expression and pose transfer, and then an inpainting network to blend the source face into the target image. Similarly, E4S [31] uses face reenactment to align the source image with the target's pose, followed by swapping faces using masks and texture information. However, these methods are sensitive to the source image; exaggerated expressions or extreme poses in the source can adversely affect the swapping result.

Target-oriented methods [8, 12, 15, 25, 30, 39, 41, 45, 52], on the other hand, modify the features of the target image to incorporate the source identity. Some of these methods [12, 25], based on autoencoder architecture, can swap between specific identities, while others, like GAN-based approaches [8, 15, 30, 39, 41, 45, 52], can generalize to various identities by merging the source identity and target attributes at the feature level. For example, SimSwap [8] offers an efficient framework for high-fidelity face swapping by injecting the source identity into the target features and using a weak feature matching loss to maintain attributes. These methods are more adept at handling variations in the source face compared to source-oriented methods. However, they often struggle to balance competing objectives, such as reconstruction loss and identity loss.

Our diffusion-based approach differs from these methods by relying on a single reconstruction loss for simplicity, while still generating images that both look natural in the target context and preserve the source face's identity.

3. Methodology

Our approach to face anonymization is similar to face swapping, but with a key difference. In face swapping, two images are used: a driving image (containing the face to be replaced) and a source image (providing the new face). Our face anonymization method, however, requires only one input image. Therefore, we developed a framework that initially learns to perform realistic face swaps using both driving and source images. We then expanded this model to work in two scenarios simultaneously: one where a source image is provided, and the other where no source image is available. This dual training allows the model to generate a new face even when given just one image. The result is a system that can synthesize a convincing, anonymous face while preserving the original image's facial expression, head posture, gaze direction, and surrounding elements. This achieves our main goal: replacing a person's face without revealing their identity or compromising the image's overall quality.

3.1. A New Paradigm

We aim to address several common limitations in current face anonymization and face swapping techniques.

First, while facial landmarks and masks provide a structured approach for face anonymization [24, 35, 48, 53] and face swapping [31, 36, 52, 58, 60], they have inherent limitations that can compromise the quality, realism, and flexibility of generated images. These methods identify major features like the eyes, nose, and mouth but miss finer details such as skin texture and nuanced expressions. This oversimplification results in less realistic and detailed facial representations compared to methods that consider pixel-level information. Additionally, the quality of the generated face heavily relies on the accurate detection of landmarks and masks; inaccuracies can lead to distorted or unrealistic faces. Moreover, facial landmarks and masks struggle to effectively capture dynamic expressions and poses, limiting the ability to generate faces with a wide range of emotions and orientations.

Second, using ArcFace [13], a loss function in deep face recognition models, to learn discriminative facial features for face anonymization [4, 29, 42, 56] can have drawbacks. The biases in these encoded features can negatively affect the quality of the anonymized faces. As shown in Fig. 3, ArcFace [13] can sometimes produce misleading identity distances, indicating greater distance between two images of the same person than between two images of different individuals. These errors typically stem from variations in pose, lighting, facial expressions, occlusions, or image quality.

Lastly, training models for face swapping often involves optimizing multiple loss functions, such as reconstruction loss and identity loss, to address different aspects of the output. However, these losses can sometimes conflict, leading to suboptimal results. This issue often arises from insufficient disentanglement between identity and non-identity features. Methods that prioritize preserving the source identity, like those using 3D priors [52], often lose the target's non-identity details. Conversely, approaches like Faceshifter [30] and DiffSwap [58], which focus on preserving the target's low-level attributes, risk allowing the target's facial identity to appear in the final swapped image.

To overcome these limitations, we use networks that capture and utilize pixel-level information, enhancing the quality of the generated faces without relying on additional facial landmarks or masks. Previous research [3, 22, 51, 54] has shown that these networks effectively preserve the fine-grained details of input images. Additionally, we simplify the training process of our networks by employing a single mean squared error loss function, avoiding the complexities associated with multiple loss functions and the dependence on facial features encoded by face recognition models. This approach offers several advantages, including simplicity, stability, and improved quality.

3.2. Architecture

As illustrated in Fig. 2, our architecture uses the Latent Diffusion Model [40], based on a UNet structure, to produce the final output images. Stacked on top of this UNet are two ReferenceNet [22] models that transfer fine-grained details from the input images. The first ReferenceNet model, which we call the source ReferenceNet model, takes the source images as input. These images provide information about the desired identity to be transferred. The second model, named the driving ReferenceNet model, takes the driving images as input. These images set the non-identity related conditions, such as pose, expression, and background details.

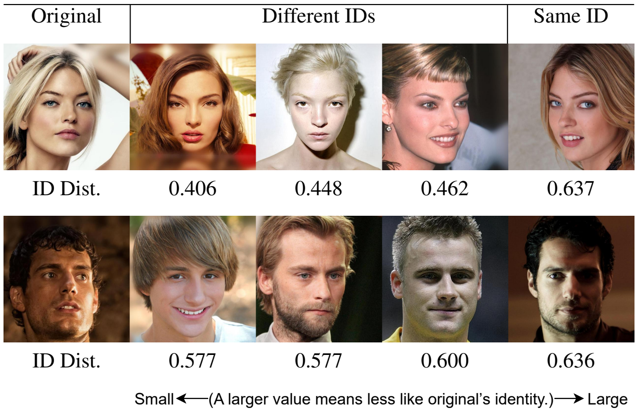

Figure 3. For each row, we show the identity distance of each image from the original image in that row, as calculated by the ArcFace [13] recognition model. The results indicate that the recognition model can generate inaccurate identity distances. It may assign a greater identity distance to two images of the same person than to two images of different people due to variations in head pose, facial expressions, or lighting conditions.

ReferenceNet shares the same structure as UNet. It captures details from input images and modulates the UNet through self-attention at each diffusion step. The process unfolds as follows: First, an input image is encoded into latent space using the CLIP [38] image encoder and then passed to ReferenceNet. Within each attention module of ReferenceNet, self-attention is applied to extract features from the CLIP-encoded image. These extracted features serve as input states for corresponding attention modules in the UNet. Specifically, the self-attention modules within the UNet receive the concatenated states from all three models—the two ReferenceNet models and the UNet itself. The output from these modules is split into three parts, with one part sent to the UNet's cross-attention module for further processing. This step is also depicted in Fig. 2.

Our architectural design offers three main advantages. First, due to their similar network structures, ReferenceNet can retain the extensive knowledge and capabilities that the UNet acquired from a large dataset by adopting its pre-trained weights. This approach prevents the training of ReferenceNet from compromising the UNet's performance and enhances both ReferenceNet's performance and training efficiency. Second, the UNet can utilize knowledge learned by ReferenceNet because of their structural similarities and shared initialization weights. This allows the UNet to extract and incorporate relevant features from ReferenceNet during training, as both networks operate in a shared feature space. Finally, by separating the data flows for source and driving images, the UNet can more effectively identify which features of the driving image to retain and which to replace with those from the source image. This clear distinction is crucial for synthesizing the final output image accurately.

3.3. Anonymization

Our framework's training method enables the UNet to selectively learn identity information from the source ReferenceNet model and non-identity-related information from the driving ReferenceNet model. The UNet then combines these two types of information to synthesize a new facial image. To anonymize a facial image, we use the same image as input for both source and driving ReferenceNet models, while adjusting intermediate inputs to the source ReferenceNet and UNet models. Specifically, we modify two key components:

1. Intermediate image embedding. We adjust the intermediate image embedding from the image encoder using this equation:

$$Z'_{img} = (1 - d) \cdot Z_{img} \tag{1}$$

   Here, Z'_{img} is the adjusted embedding, d controls the degree of anonymization, and Z_{img} is the original embedding. As d increases, more identity information is removed from Z'_{img}. This adjusted embedding influences both the source ReferenceNet (via self-attention) and the UNet (via cross-attention).

2. Source ReferenceNet state. We modify the state of the source ReferenceNet using this equation:

$$S' = (1 - d) \cdot S_{cond} + d \cdot S_{uncond} \tag{2}$$

   S' is the modified state, d is the same factor controlling the degree of anonymization, S_{cond} is the conditional state (with identity information), and S_{uncond} is the unconditional state (without identity information). As d increases, S' shifts further from the conditional state towards the unconditional state. The modified state S' is then incorporated into the UNet's intermediate layers using self-attention.

Simply put, the equations demonstrate that by increasing the parameter d, the original identity is gradually removed from the resulting image while an unknown identity is progressively introduced. This process transforms the original identity into a different one, effectively achieving the desired anonymization.

4. Experiments

This section includes our experimental setup, procedures, findings, and approaches used to analyze our results.

4.1. Implementation Details

We trained our model using three datasets: CelebRef-HQ [32], CelebA-HQ [26], and FFHQ [27]. Face recognition [44] was used to identify images of the same person, and for each identity, two images were randomly selected: one as the source and one as the ground truth. A synthesized driving image was then generated by using a state-of-the-art face-swapping model [18] to replace the face in the ground truth image with another person's face. These three images—the source, synthesized driving, and ground truth—were used to train our model to learn identity changes. For a detailed breakdown of the number of images used in training, please refer to our supplementary material.

The ReferenceNet models and the UNet were initialized from a pre-trained Stable Diffusion [40] v2.1 model. To incorporate classifier-free guidance [21], we applied the unconditional mode to a random 10% of the training data, while the conditional mode was used for the remaining 90%.

During training, we discovered that focusing solely on the attention modules in the ReferenceNet model was as effective as training the entire model. This finding aligns with our understanding that these attention layers play a crucial role in shaping the structure and content of the generated images. As a result, we chose to optimize only the weights of the UNet and the attention modules in the ReferenceNet models. This targeted strategy allowed us to streamline our training process while maintaining effectiveness. We trained the model at a final output resolution of 512 × 512 over 435,000 steps. The training utilized the AdamW [33] optimizer with a batch size of 1 and 8 accumulation steps, maintaining a fixed learning rate of 1e-5. This process was conducted on two A6000 GPUs.

We also observed that using only synthesized images as driving images led to a problem where our model performed well only when the driving image was synthesized. To enhance performance and generalization, we adopted strategies from curriculum learning [5]. Initially, we trained the model with both real and synthesized driving images. When the driving image was real, we used its face-swapped counterpart as the ground truth and an image of the person originally used to swap the face in the driving image as the source. As training progressed, we transitioned to using only synthesized images as driving images and fine-tuned

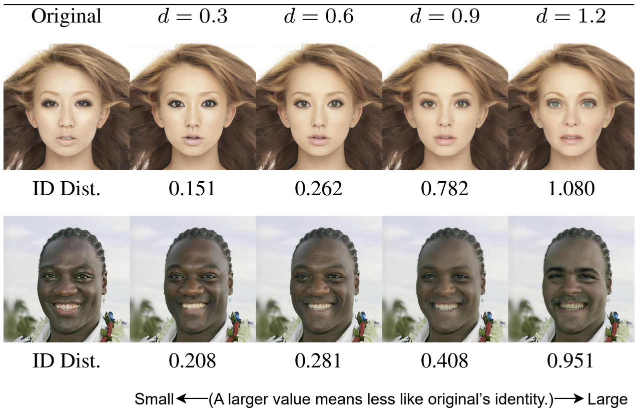

Figure 4. Facial images generated with different degrees of anonymization. Each generated image reflects a different degree of anonymity applied to the original face. Alongside each generated image is a cosine distance score, calculated using the FaceNet [44] recognition model. This score quantifies how different the anonymized face is from the original in terms of identity features.

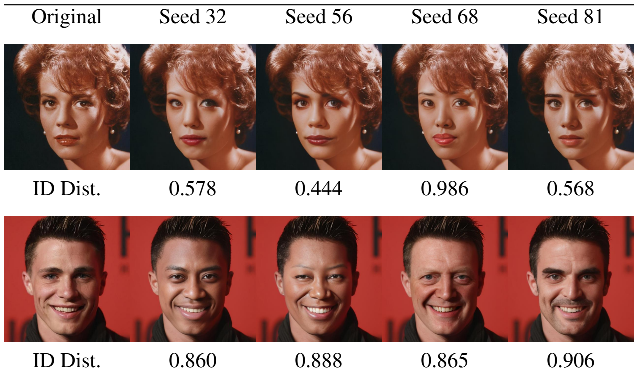

Figure 5. Various anonymized versions created from a single original identity, each using a different integer seed value. For each anonymized version, we present the cosine distance from its original identity, calculated using the FaceNet [44] recognition model.

the model solely on real images as ground truth. This approach allows the model to first learn fundamental representations from a diverse set of data and then improve its capability to generate more realistic images.

Throughout this study, we maintained consistent parameters for image generation. We used the DDPM [20] algorithm with 200 denoising steps and a guidance scale value [21] of 4.0 for all examples presented in this paper.

4.2. Achieving Diverse Anonymization Results

Two methods allow us to vary anonymization results. First, we can modify the floating-point value d, defined in Eqs. (1) and (2), which controls the anonymization intensity. Higher d values produce images that deviate more from the original, as Fig. 4 demonstrates. When d surpasses 1, the process moves in the opposite direction of the original identity's defining characteristics, ensuring the anonymized identity is not overly similar to the original. Second, we can use different integer seed values. This change introduces different initial Gaussian noise, leading to varied outcomes, as shown in Fig. 5.

4.3. Baseline Comparisons

Baselines. We benchmarked our model against three leading face anonymization methods (DP2 [23], FALCO [4], and RiDDLE [29]) and three leading face swapping methods (DiffSwap [58], BlendFace [45], and InSwapper [18]). For evaluation, we used images not included in the training datasets. Specifically, we selected 1,000 images each from CelebA-HQ [26] and FFHQ [27], totaling 2,000 images for testing.

Evaluation Metrics. We evaluate the generated facial images using several metrics: re-identification rate, face shape distance, pose distance, gaze distance, expression distance, and image quality.

To calculate the re-identification rate, we extract identity vectors using the FaceNet [44] recognition model and compute the cosine similarity to measure identity distance. For each generated face in the test set, we find the most similar face within the same test set. If this face matches the original face used for generation, we increment the re-identification count by one; otherwise, the count remains unchanged.

Face shape and expression distances are assessed using a face reconstruction model [14]. This model predicts 3DMM [6] coefficients for both generated and original faces, allowing us to calculate the L2 Euclidean Distance between these coefficients.

For pose distance, we use a head pose estimation model [43] to predict the orientation of both the generated and original faces. We then calculate the quaternion angular distance between these orientations. Gaze distance is computed similarly. We employ a gaze estimation model [2] to predict the gaze direction of both the generated and original faces, then calculate the quaternion angular distance between these predicted directions.

Image quality is measured using an Image Quality Assessment (IQA) network [7] specifically trained on a face IQA dataset [47], which is ideal for evaluating the quality of facial images.

Quantitative Comparison. The quantitative results in Tab. 1 demonstrate our model's performance in face anonymization in comparison to baseline methods. We did not include the quantitative results for FALCO [4] on

	Identity Distance				Attribute Distance						Image Quality	
	Re-ID ↓		Shape ↑		Pose ↓		Gaze ↓		Expression ↓		Face IQA ↑	
	CelebA-HQ	FFHQ	CelebA-HQ	FFHQ	CelebA-HQ	FFHQ	CelebA-HQ	FFHQ	CelebA-HQ	FFHQ	CelebA-HQ	FFHQ
DP2 [23]	0.020	0.046	30.297	29.837	0.140	0.194	0.244	0.252	10.139	9.613	0.459	0.480
FALCO [4]	**0.005**	-	31.816	-	0.088	-	0.258	-	9.290	-	**0.757**	-
RiDDLE [29]	-	**0.007**	-	36.624	-	0.090	-	0.220	-	10.018	-	0.571
Ours ($d = 1.2$)	0.053	0.098	33.046	28.971	**0.048**	**0.047**	**0.161**	**0.166**	**8.256**	**7.769**	0.701	0.698
Ours ($d = 1.4$)	0.008	0.039	**53.244**	**41.695**	0.074	0.061	0.190	0.206	13.125	10.899	0.707	**0.704**

Table 1. Quantitative results on the task of face anonymization for CelebA-HQ [26] and FFHQ [27] test sets, with the best results highlighted in bold and the second-best results underlined.

the FFHQ [27] test set and RiDDLE [29] on the CelebA-HQ [26] test set, as they require additional information that is not readily available. For quantitative results related to face swapping, please see our supplementary material.

Table 1 indicates that our model, with $d = 1.4$, excels in producing faces with highly distinct shapes while maintaining the original pose and gaze across both datasets. Conversely, when we set d to a smaller value of 1.2, our model best preserves all three original facial attributes (pose, gaze, and expressions) across both datasets. However, this smaller d comes with lower re-identification performance and face shapes more similar to the original. Generally, a smaller d value improves attribute preservation, but results in lower re-identification performance and more similar face shapes. This is expected, as the generated image remains closer to the original.

We recognize that our method does not achieve the lowest re-identification rates compared to FALCO [4] and RiDDLE [29] when assessed by the FaceNet [44] recognition model. We examined the cases where the recognition model successfully traced our model's outputs back to their original images. Many involved subjects from underrepresented groups in our training data, particularly infants and ethnic minorities like Asian individuals. This lack of representation led to poorer model performance in these scenarios. This data imbalance also explains why our model performs better on the CelebA-HQ [26] dataset compared to FFHQ [27], as the former contains fewer examples of infants and minority groups. In comparison, RiDDLE [29] achieves the lowest re-identification rate on the FFHQ [27] dataset, as it explicitly uses an identity loss term to distinguishes between real and anonymized faces. However, it also relies on several additional loss terms to preserve non-identity-relevant facial attributes and background. The use of multiple loss terms can lead to conflicts between different objectives, potentially resulting in less-than-ideal outcomes.

Regarding image quality, our model ranks second behind FALCO [4]. This may be due to FALCO's [4] ability to natively generate higher resolution images (1024 × 1024) compared to our model's native resolution of 512 × 512. While the SDXL [37] model allows us to create images exceeding 512 × 512 resolution, training and testing such larger models require significantly more GPU memory, which is currently beyond our available resources.

Qualitative Comparison. Figures 6 and 7 present qualitative comparison results for anonymization tasks on the CelebA-HQ [26] and FFHQ [27] test sets, respectively. For face swapping tasks, Fig. 8 showcases two representative examples. Additional results are available in the supplementary material.

From Figs. 6 and 7, we observe that DP2 [23] sometimes produces artifacts where the anonymized face does not align correctly with the position or orientation of the original face in the image. This issue arises because DP2 [23] approaches anonymization as an image inpainting task. It first detects and crops the face region from the input photo, then applies a predicted mask over the region to be anonymized. An inpainting generator is then used to fill in these masked area with an anonymized face. However, if the mask inaccurately removes parts of the image, it can disrupt the inpainting process, leading to misaligned or distorted results. Our method overcomes these limitations of inpainting-based approaches by generating the entire image from a noise map, avoiding dependency on masks.

We also note that FALCO [4] does not preserve background details because its design does not include background elements in its loss functions. Although FALCO [4] incorporates facial attribute preservation loss, it struggles with maintaining certain facial features, such as eye direction, because it relies on finding similarity within the FaRL [59] feature space, which does not encode eye gaze information. RiDDLE [29] attempts to preserve image quality and similarity at the perceptual feature level by using a perceptual loss [57], but it still fails to accurately replicate specific details like eye direction, clothing, and background elements from the original image. In contrast, our method effectively modifies identity-related facial features while preserving non-identity-related details, thanks to its face-swapping approach and the advantages of its model architecture.

4.4. Ablation Study

We conduct an ablation study on our anonymization approach, focusing on three key design elements related to

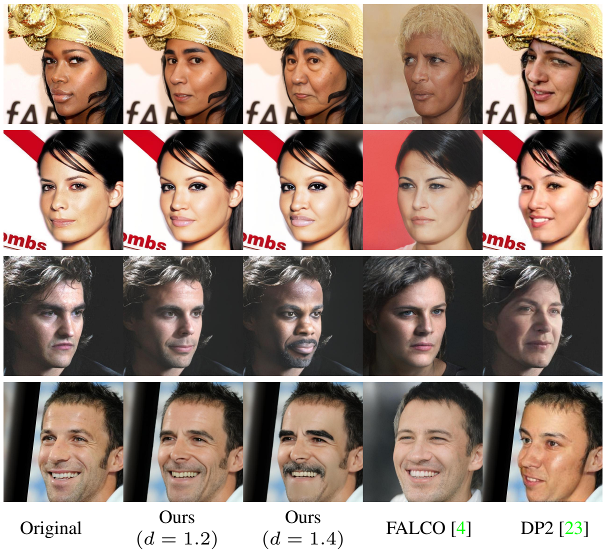

Figure 6. Qualitative results for the face anonymization task for the CelebA-HQ [26] test set.

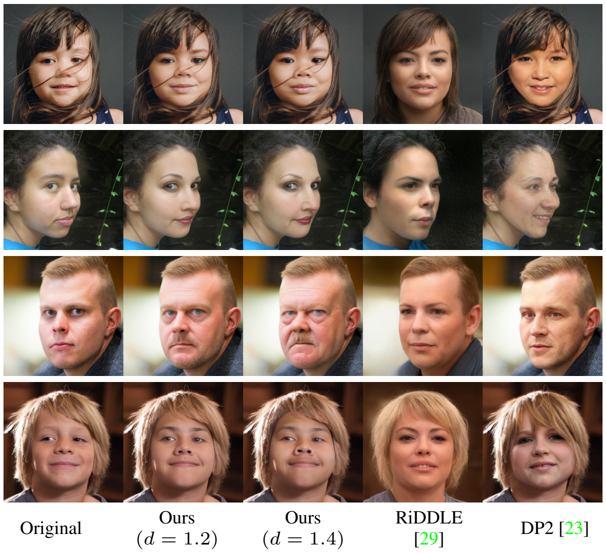

Figure 7. Qualitative results for the face anonymization task for the FFHQ [27] test set.

Eqs. (1) and (2): (1) unmodified intermediate image embeddings from the image encoder, (2) unmodified states of the source ReferenceNet model, and (3) modification limited to intermediate image embeddings from the image encoder and conditional states of the source ReferenceNet model, excluding its unconditional states.

Table 2 presents the re-identification performance and face shape distance for our full method and each individual design choice. Our analysis reveals that: (1) modifying only the intermediate image embeddings or only the ReferenceNet states is not enough to improve re-identification performance or increase face shape distinctiveness. (2) Changing both the intermediate image embeddings and the conditional states of the source ReferenceNet model, without including its unconditional states, also fails to achieve significant improvements. (3) The key to substantially enhancing re-identification performance and creating less similar face shapes lies in a combined approach—modifying both the intermediate image embeddings and the conditional states of the source ReferenceNet model, while also incorporating its unconditional states.

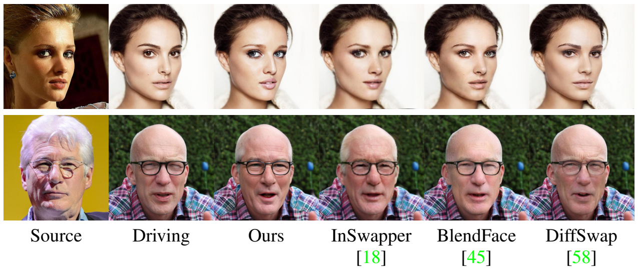

Figure 8. Qualitative results for the face swapping task for the CelebA-HQ [26] test set in the upper row and the FFHQ [27] test set in the lower row.

	Identity Distance			
	Re-ID ↓		Shape ↑	
	CelebA-HQ	FFHQ	CelebA-HQ	FFHQ
~~embeds~~ [a]	0.378	0.309	15.756	18.881
~~states~~ [b]	0.288	0.545	21.342	16.566
~~uncond states~~ [c]	0.159	0.243	17.559	18.867
Ours	**0.008**	**0.039**	**53.244**	**41.695**

[a] Ours without modifying intermediate image embeddings
[b] Ours without modifying ReferenceNet states
[c] Ours without including unconditional ReferenceNet states

Table 2. Ablation analysis of identity anonymization performance on the CelebA-HQ [26] and FFHQ [27] test sets, with the best results highlighted in bold.

The last row of Tab. 2, representing our full method, demonstrates the effectiveness of this comprehensive approach.

5. Conclusion

We have introduced our approach leveraging diffusion models for face anonymization. Our framework eliminates the need for facial keypoints and masks and relies solely on a reconstruction loss, while still generating images with detailed fine-grained features. Our results show that this method effectively anonymizes faces, preserves attributes, and produces high-quality images. Additionally, our model can use an extra facial image input to perform face swapping tasks, demonstrating its versatility and potential for various facial image processing applications.

References

[1] General Data Protection Regulation (GDPR) Compliance Guidelines. https://gdpr.eu/. 1

[2] Ahmed A Abdelrahman, Thorsten Hempel, Aly Khalifa, Ayoub Al-Hamadi, and Laslo Dinges. L2cs-net: Fine-grained gaze estimation in unconstrained environments. In *2023 8th International Conference on Frontiers of Signal Processing (ICFSP)*, pages 98–102. IEEE, 2023. 6

[3] Hadi Alzayer, Zhihao Xia, Xuaner Zhang, Eli Shechtman, Jia-Bin Huang, and Michael Gharbi. Magic fixup: Streamlining photo editing by watching dynamic videos. *arXiv preprint arXiv:2403.13044*, 2024. 4

[4] Simone Barattin, Christos Tzelepis, Ioannis Patras, and Nicu Sebe. Attribute-preserving face dataset anonymization via latent code optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8001–8010, 2023. 1, 2, 4, 6, 7, 8

[5] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 5

[6] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 157–164. Association for Computing Machinery, 2023. 6

[7] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 2024. 6

[8] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2003–2011, 2020. 3

[9] Umur Aybars Ciftci, Ali Kemal Tanriverdi, and Ilke Demir. My body my choice: Human-centric full-body anonymization. *arXiv preprint arXiv:2406.09553*, 2024. 2

[10] Umur A Ciftci, Gokturk Yuksek, and Ilke Demir. My face my choice: Privacy enhancing deepfakes for social media anonymization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1369–1379, 2023. 2

[11] Nicola Dall'Asen, Yiming Wang, Hao Tang, Luca Zanella, and Elisa Ricci. Graph-based generative face anonymisation with pose preservation. In *International Conference on Image Analysis and Processing*, pages 503–515. Springer, 2022. 2

[12] deepfakes. FaceSwap. https://github.com/deepfakes/faceswap/. 3

[13] 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. 4

[14] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 6

[15] Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. Information bottleneck disentanglement for identity swapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3404–3413, 2021. 3

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[17] Xiuye Gu, Weixin Luo, Michael S Ryoo, and Yong Jae Lee. Password-conditioned anonymization and deanonymization with face identity transformers. In *European conference on computer vision*, pages 727–743. Springer, 2020. 2

[18] Jia Guo, Jiankang Deng, Xiang An, Jack Yu, and Baris Gecer. InsightFace Swapper. https://github.com/deepinsight/insightface/tree/master/examples/in_swapper/. 5, 6, 8

[19] Majed El Helou, Doruk Cetin, Petar Stamenkovic, and Fabio Zund. Vera: Versatile anonymization fit for clinical facial images. *arXiv preprint arXiv:2312.02124*, 2023. 2

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 6

[21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5, 6

[22] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 4

[23] Håkon Hukkelås and Frank Lindseth. Deepprivacy2: Towards realistic full-body anonymization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1329–1338, 2023. 1, 2, 6, 7, 8

[24] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *International symposium on visual computing*, pages 565–578. Springer, 2019. 2, 3

[25] iperov. DeepFaceLab. https://github.com/iperov/DeepFaceLab/. 3

[26] 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. 5, 6, 7, 8

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 5, 6, 7, 8

[28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[29] Dongze Li, Wei Wang, Kang Zhao, Jing Dong, and Tieniu Tan. Riddle: Reversible and diversified de-identification with latent encryptor. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8093–8102, 2023. 1, 2, 4, 6, 7, 8

[30] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019. 3, 4

[31] Maomao Li, Ge Yuan, Cairong Wang, Zhian Liu, Yong Zhang, Yongwei Nie, Jue Wang, and Dong Xu. E4s: Fine-grained face swapping via editing with regional gan inversion. *arXiv preprint arXiv:2310.15081*, 2023. 3

[32] Xiaoming Li, Shiguang Zhang, Shangchen Zhou, Lei Zhang, and Wangmeng Zuo. Learning dual memory dictionaries for blind face restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5904–5917, 2022. 5

[33] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[34] Yuchen Luo, Junwei Zhu, Keke He, Wenqing Chu, Ying Tai, Chengjie Wang, and Junchi Yan. Styleface: Towards identity-disentangled face generation on megapixels. In *European conference on computer vision*, pages 297–312. Springer, 2022. 2

[35] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5447–5456, 2020. 2, 3

[36] 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. 3

[37] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 7

[38] 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. 4

[39] Xiaohang Ren, Xingyu Chen, Pengfei Yao, Heung-Yeung Shum, and Baoyuan Wang. Reinforced disentanglement for face swapping without skip connection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20665–20675, 2023. 3

[40] 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. 4, 5

[41] Felix Rosberg, Eren Erdal Aksoy, Fernando Alonso-Fernandez, and Cristofer Englund. Facedancer: Pose-and occlusion-aware high fidelity face swapping. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3454–3463, 2023. 3

[42] Felix Rosberg, Eren Erdal Aksoy, Cristofer Englund, and Fernando Alonso-Fernandez. Fiva: Facial image and video anonymization and anonymization defense. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 362–371, 2023. 2, 4

[43] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2074–2083, 2018. 6

[44] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 5, 6, 7

[45] Kaede Shiohara, Xingchao Yang, and Takafumi Taketomi. Blendface: Re-designing identity encoders for face-swapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7634–7644, 2023. 3, 6, 8

[46] Nadiya Shvai, Arcadi Llanza Carmona, and Amir Nakib. Adaptive image anonymization in the context of image classification with neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5074–5083, 2023. 2

[47] Shaolin Su, Hanhe Lin, Vlad Hosu, Oliver Wiedemann, Jinqiu Sun, Yu Zhu, Hantao Liu, Yanning Zhang, and Dietmar Saupe. Going the extra mile in face image quality assessment: A novel database and model. *IEEE Transactions on Multimedia*, 2023. 6

[48] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and effective obfuscation by head inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5050–5059, 2018. 2, 3

[49] 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. 2

[50] Julian Todt, Simon Hanisch, and Thorsten Strufe. Fant\^ omas: Understanding face anonymization reversibility. *arXiv preprint arXiv:2210.10651*, 2022. 2

[51] Qilin Wang, Zhengkai Jiang, Chengming Xu, Jiangning Zhang, Yabiao Wang, Xinyi Zhang, Yun Cao, Weijian Cao, Chengjie Wang, and Yanwei Fu. Vividpose: Advancing stable video diffusion for realistic human image animation. *arXiv preprint arXiv:2405.18156*, 2024. 4

[52] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hififace: 3d shape and semantic prior guided high fidelity face swapping. *arXiv preprint arXiv:2106.09965*, 2021. 3, 4

[53] Yunqian Wen, Bo Liu, Jingyi Cao, Rong Xie, and Li Song. Divide and conquer: a two-step method for high quality face de-identification with model explainability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2023. 2, 3

[54] Jingyun Xue, Hongfa Wang, Qi Tian, Yue Ma, Andong Wang, Zhiyuan Zhao, Shaobo Min, Wenzhe Zhao, Kai-

hao Zhang, Heung-Yeung Shum, et al. Follow-your-pose v2: Multiple-condition guided character image animation for stable pose control. *arXiv preprint arXiv:2406.03035*, 2024. 4

[55] Ge Yuan, Maomao Li, Yong Zhang, and Huicheng Zheng. Reliableswap: Boosting general face swapping via reliable supervision. *arXiv preprint arXiv:2306.05356*, 2023. 3

[56] Liming Zhai, Qing Guo, Xiaofei Xie, Lei Ma, Yi Estelle Wang, and Yang Liu. A3gan: Attribute-aware anonymization networks for face de-identification. In *Proceedings of the 30th ACM international conference on multimedia*, pages 5303–5313, 2022. 2, 4

[57] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[58] 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. 3, 4, 6, 8

[59] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18697–18709, 2022. 7

[60] Yixuan Zhu, Wenliang Zhao, Yansong Tang, Yongming Rao, Jie Zhou, and Jiwen Lu. Stableswap: Stable face swapping in a shared and controllable latent space. *IEEE Transactions on Multimedia*, 2024. 3