Divide and Conquer: a Two-Step Method for High Quality Face De-identification with Model Explainability

Yunqian Wen[1], Bo Liu[2], Jingyi Cao[1], Rong Xie[1], Li Song[1, 3, ✉]
[1]Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University
[2] School of Computer Science University of Technology Sydney
[3] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[1]{wenyunqian, cjycaojingyi, xierong, song_li}@sjtu.edu.cn [2]{bo.liu}@uts.edu.au

Abstract

Face de-identification involves concealing the true identity of a face while retaining other facial characteristics. Current target-generic methods typically disentangle identity features in the latent space, using adversarial training to balance privacy and utility. However, this pattern often leads to a trade-off between privacy and utility, and the latent space remains difficult to explain. To address these issues, we propose IDeudemon, which employs a "divide and conquer" strategy to protect identity and preserve utility step by step while maintaining good explainability. In Step I, we obfuscate the 3D disentangled ID code calculated by a parametric NeRF model to protect identity. In Step II, we incorporate visual similarity assistance and train a GAN with adjusted losses to preserve image utility. Thanks to the powerful 3D prior and delicate generative designs, our approach could protect the identity naturally, produce high quality details and is robust to different poses and expressions. Extensive experiments demonstrate that the proposed IDeudemon outperforms previous state-of-the-art methods.

1. Introduction

Concerns about individual private information disclosure are growing with the development of computer vision techniques and image understanding applications. Face de-identification is a process which aims to remove all identification information of the person from an image, while maintaining as much information on the action and its context [1]. Ideally, while the identity information is protected, other identity-agnostic features (e.g., pose, expression and background) will not be affected. The de-identified images can still be used for identity-agnostic tasks, such as face detection and expression recognition. Accordingly, great efforts are paid to achieve an effective privacy utility trade-off [2–9]. Face de-identification can allow individuals to share personal portraits with confidence, while eliminating some ethical and legal restraints on facial data releasing.

Early face de-identification methods carry out various obfuscation operations on detected private area, which seriously impair the image's ornamental value and are not reliable when facing advanced face recognition tools [10]. K-same family methods [11–13] are once hot, but they are restrained by their strict using conditions. At present, there are two main types of methods. One kind uses adversarial noise to generate de-identified faces which can be visually indistinguishable from the original one [14–16]. However, they are highly dependent on the accessibility to target systems, and lack generalization ability. The other kind exploits generative adversarial networks (GANs) to disentangle, manipulate and finally protect identity features in the latent spaces [2–9]. These methods strive to strike a balance between privacy and utility through adversarial training in **a** network. The results depend heavily on the degree of latent space disentanglement, which is neither clear nor satisfactory. Besides, most existing methods cannot preserve various poses and expressions, which also need to be improved.

Unlike previous works, we aim to break away from this traditional privacy utility trade-off in face de-identification studies, and instead provide a reliable and explainable method of protecting individual identities. Our inspiration for this approach stems from the observation that wearing a human skin mask can effectively change one's identity. This realization highlights that a convincing de-identification requires substantial changes to the overall geometry of the facial features such as eyes, nose, ears, mouth, and facial bones. Such transformations are practically impossible to achieve with mere makeup or even surgical procedures (since that surpass the physical limits of the human body). In contrast, hairstyle, accessories, and skin color are examples of identity-agnostic features that can be easily altered by a stylist. However, they significantly impact the human perception of visual similarity between two faces. Thus, we contend that protecting privacy and retaining utility can be two distinct objectives that necessitate different strategies. By separating these objectives, we can focus on each objective independently to achieve better results.

Figure 1. IDeudemon for face de-identification at different resolutions. (a) 256 × 256, (b) 512 × 512, (c) 1024 × 1024. In each pair, left is the original image and right is the corresponding de-identified result. The results show that face identities are changed in a perceptually natural manner, while all other characteristics (hairstyles, accessories, backgrounds, poses, expressions, etc.) remain the same.

Our proposed solution, IDeudemon, adopts the "divide and conquer" strategy to achieve privacy protection and utility preservation in two distinct steps. In the first step, we use a 3D parametric modeling approach to estimate the facial geometry and obfuscate the face's 3D identity representation to conceal the real identity. Specifically, we begin by leveraging a monocular face reconstruction network to approximate the coarse 3D parameters of the given face. Using this initialization, we employ a neural radiance field (NeRF) model to calculate the face's accurate 3D parameters (ID code, appearance code and camera code). Subsequently, we apply a protective perturbation to the real ID code to get the protected ID code. Finally, the NeRF model renders an identity-protected fitted face, which has a significant change in the facial features' geometric structure.

In the second step, we focus on producing high quality images based on the fitted face, which is neither natural nor realistic. We first use face parsing maps to preserve the identity-agnostic features and maintain the visual similarity with the original image as much as possible. Then, we train a GAN to restore the de-identified face with realistic details by referring to generative facial priors. Finally, we can acquire high quality visual pleasing de-identified results.

Our main contributions are described as follows:

- We propose IDeudemon, a novel two-step NeRF-based method for face de-identification. Instead of achieving privacy utility trade-off in one network adversarially, for the first time, we divide privacy protection and utility preservation into two separate steps. IDeudemon can protect identity without weighing the image utility at the same time, and has good explainability [17].
- We confuse the real identity by a 3D parametric NeRF model, which modifies the facial geometry and changes the identity. Hence, our method has excellent privacy performance and this process is explainable. The definition of the identity refers to the mature 3D prior from 3DMMs, and is refined by the NeRF model. This verified disentangled identity code makes IDeudemon well preserve non-identity features, such as expression, pose and illumination.
- We propose a second step to intently restore high quality faces based on the fitted results of NeRF. We devise visual similarity assistance to retain identity-agnostic features and train a GAN to generate realistic facial details. These designs lead to good utility performance.
- Experimental results on two diverse face datasets (ethnicity, age, etc.) have shown the effectiveness of our proposed IDeudemon. In particular, our method brilliantly maintain the original poses and expressions, and can achieve face de-identification on megapixels.

2. Related Work

2.1. Face De-identification

Initial approaches are mainly obfuscation-based, such as blur and pixelization. Although simple and fast, they have been proved to be vulnerable [10]. Methods [11–13] based on K-same algorithm were then proposed to improve the protection ability and image utility. However, these methods have many harsh assumptions on the application scenario. Furthermore, their anonymous effects are not natural.

With the rapid development of deep learning, the neural network architectures have evolved and greatly flourished the face de-identification research. Nowadays, adversarial noise-based methods and GAN-based methods have become the dominant paradigm. The former seeks to generate a small but intentional worst-case disturbance to an original image, which misleads specific face recognition models without causing a noticeable difference perceptible to human eyes [14–16]. The latter achieves target-generic de-identification, i.e., they are designed to work against any recognizer. These methods typically first define the representations of identity and other facial attributes in the latent space. Then they design loss functions that aim at

disentangling identity and maintaining utility respectively. De-identification will be implemented in a neural network by striking a balance adversarially [2–9]. Although these methods have achieved promising progress, the explanation of identity and other facial representations in the latent space is still ambiguous. Besides, they only work well on frontal facial images with neutral expression.

In contrast, our de-identification method possesses several merits: (1) The identity protection process is explainable; (2) It eliminates the need for privacy utility trade-off; (3) It can adapt to different poses and expressions; (4) The results have photo-realistic details.

2.2. 3D Monocular Face Reconstruction

3D monocular face reconstruction refers to reconstructing the 3D model of a face from a 2D image. Methods [18–20] based on 3D Morphable Models (3DMMs) [21]has dominated this field. Besides, there exist some methods advocating direct model-free reconstruction [22] or based on other innovative models [23]. However, all these methods suffer from the problem that the reconstructed faces are not realistic. Recently, NeRF shows encouraging results in capturing implicitly-encoded complex scene structures and fitting 3D-consistent images with fine details [24–26]. As faces contain regular 3D structure, NeRF-based 3D face modeling researches [27–30] are now in full swing.

2.3. Blind Face Restoration

Blind face restoration (BFR) aims at recovering high quality faces from the low quality counterparts suffering from unknown degradation [31]. Current BFR methods always require facial priors, which can be coarsely categorized into three types according to the sources: geometric priors [32, 33], reference priors [34–36] and generative priors [31, 37, 38]. Among them, the third kind is not limited by the quality of corrupted faces, the accessibility of high-resolution references having the same identity, or the capacity of the references. So it is the most suitable for the restoration of fitted faces rendered by current NeRF.

3. Methodology

3.1. Overview of IDeudemon

Given an input face image X without any protection, the purpose of face de-identification is to generate a photo-realistic image X' which conceals the real identity. The de-identified face X' is visually similar to the original image X, but should be judged as a different person by recognition tools when comparing with X.

Fig. 2 illustrates the overall pipeline of the proposed IDeudemon, which protects privacy and guarantees utility in distinct steps sequentially. In the following, we discuss the two steps in detail.

3.2. Step I: Parametric Identity Protection

Coarse 3D Parameters Evaluation. 3DMMs are generative parametric models for the 3D representation of human faces. They are built from a set of 3D facial scans, coupled to each other with anatomical correspondences, and can represent any unseen faces as a linear combination of the training set [39]. Fitting 3DMMs, also known as 3D face reconstruction, facilitates the estimation of identity, pose, albedo and illumination related parameters from the face images. In order to provide a good basis for real-time NeRF-based fitting, we employ a 3DMM model [40], denoted as $\mathcal{M}_{fr}$, to initialize the 3D parameters [41] of the input face image X, which is denoted as:

$$c_{id}, c_{exp}, c_{alb}, c_{illu} = \mathcal{M}_{fr}(X). \tag{1}$$

c_* represent the coarse 3DMM parameters for four disentangled factors: identity c_{id}, expression c_{exp}, and albedo c_{alb} of the face X, and the illumination c_{illu} of the scene. These parameters are initialized by solving an inverse rendering optimization [42] based on the 3DMM model [40]. Although the initial identity parameter only describes the coarse geometry of the face area (without hair, teeth, etc.), it will be adaptively adjusted and become accurate through the NeRF model described below.

NeRF-based Identity Protection. With initialized 3DMM parameters c_*, we employ a pretrained parametric NeRF model [30], denoted as $\mathcal{M}_{nerf}$, to obtain the accurate 3D parameters and the fitted face X_f of original image X:

$$X_f, z_{id}, z_{app}, z_{cam} = \mathcal{M}_{nerf}(X, c_{id}, c_{exp}, c_{alb}, c_{illu}, C). \tag{2}$$

X_f is the fitted image. C is the camera parameter used for rendering (detailed calculation is shown in [30]). z_* represent the computed 3D codes for face image X, whose dimensionality is the same as that of the corresponding coarse 3D parameters. In particular, because our de-identification task hopes to distinguish the identity feature from all other facial features, we let z_{id} represent the identity separately, and name it as ID code. Then we let the appearance code z_{app} contain not only the expression and albedo of the face in x, but also the illumination of the whole scene. In addition, as the density field from NeRF can implicitly encode the 3D geometry of the scene, we can also acquire a camera code z_{cam}, which reflects the pose of the face in X.

To protect the real identity information, we use a noise generator to generate benign Gaussian noise n whose size equals to the fitted ID code z_{id} according to the actual requirements. Then we directly add the protective noise on z_{id} to get a perturbed ID code z'_{id}:

$$z'_{id} = z_{id} + n. \tag{3}$$

In Sec 4.2, we perform a series of perturbation analysis experiments, where we get the optimum scale range of perturbation for identity protection.

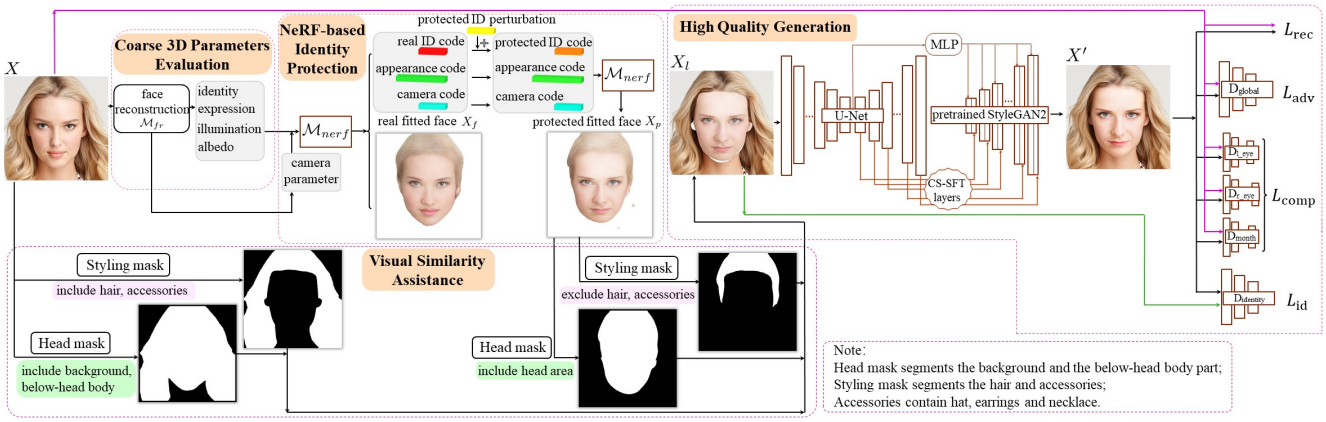

Figure 2. The architecture of IDeudemon. To protect identity, we first estimate the coarse 3D parameters of input image X as an initialization. Then a NeRF model is employed to calculate X's accurate 3D codes and fitted face X_f. After adding protective perturbation to the real ID code, the NeRF generates the de-identified fitted face X_p. To preserve utility, we design visual similarity assistance to directly retain the identity-agnostic areas and train a GAN referring to generative priors to produce the final high quality de-identified face X'.

At the end of this step, the NeRF model takes the protected identity code z'_{id}, the original appearance code z_{app} and camera code z_{cam} as input, and fits the final identity-protected fitted face X_p. It is formulated as:

$$X_p, z'_{id}, z_{app}, z_{cam} = \mathcal{M}_{nerf}(z'_{id}, z_{app}, z_{cam}). \quad (4)$$

Since our parametric NeRF model refers to the 3DMM model, the whole de-identification process has good explainability. Moreover, since the perturbation is directly added on the disentangled ID code, the result with faithful identity change still well retains identity-agnostic features (i.e., expression, albedo, illumination and pose).

3.3. Step II: Utility Preservation

Despite the promising de-identified fitted result X_p of parametric NeRF model, it has limitations in terms of realistic looks. In order to generate visual pleasing high quality faces, we take several measures as follows.

Visual Similarity Assistance. As mentioned earlier, hairstyles, accessories and background are weakly related to the identity, but may occupy a pretty large space and greatly affect human perception of visual similarity and the subsequent use. Therefore, we use face parsing maps [43, 44] to generate a head mask (which segments the background and the below-head body part) and a styling mask (which segments the hair and accessories) for X and X_p. Here we conbine the hair, accessories, background and below-head body section in the original image X with the segmented face except for the hair and accessories in the fitted image X_p. Therefore, a hybrid face image X_l is produced, which conceals the real identity and retains the identity-agnostic areas. As seen in Fig. 2, X_l has realistic identity-agnostic features, low-quality face regions, and some irregular white gaps, which still needs to be improved.

High Quality Generation. The translation from hybrid image X_l to desired high quality de-identified photo X' aims to accomplish a face restoration task, which transforms degraded image to its photo-realistic counterpart with distinct and discernible details. The domain gap is pretty large, so this task is challenging. Thanks to the leaps and bounds in BFR, here we employ a publicly available GAN model [31] that leverages rich and diverse priors encapsulated in the pretrained StyleGAN2 [45] to achieve high quality de-identified face generation. This GAN model is mainly composed of two parts: a U-Net [46] which is responsible for removing degradation and extracting "clean" features of X_l, and a pretrained StyleGAN2 that provides facial priors. They are bridged by a latent code mapping and several Channel-Split Spatial Feature Transform (CS-SFT) layers in a coarse-to-fine manner. By training this GAN model, we can obtain high quality de-identified image X'.

IDeudemon enjoys the benefits of separating the implementation of protecting privacy and preserving utility, so has the advantage of adjusting the degree of identity protection as practical need while maintaining remarkable utility performance. Our approach no longer needs to struggle with the annoying trade-off between privacy and utility.

3.4. Loss Function.

We train the GAN model with triplet of images X, X_l and X'. We inherit the validated loss functions from [31], and adjust them as the requirements of our mission.

Reconstruction Loss. The widely-used $\mathcal{L}_1$ loss and perceptual loss are summed as the reconstruction loss $\mathcal{L}_{rec}$ [47, 48], which targets at making the output X' look like the original face X:

$$L_{rec} = \lambda_{l_1} \|X' - X\|_1 + \lambda_{per} \|\phi(X') - \phi(X)\|_1, \quad (5)$$

where ϕ is the pretrained VGG-19 network [49] and we

select the $conv1, \cdots, conv5$ feature maps before activation.

Adversarial Loss. The adversarial loss $\mathcal{L}_{adv}$ is responsible for restoring realistic textures, enforcing generated faces to be indistinguishable from real images. It is formulated as:

$$L_{adv} = -\lambda_{adv}\mathbb{E}_{X'}[softplus(D(X'))], \tag{6}$$

where D denotes the discriminator and λ_{adv} represents the adversarial loss weight.

Facial Component Loss. Given that people easily detect mistakes in the appearance of a human face (uncanny valley effect), we also use the facial component loss with local discriminators for left eye, right eyes and mouth, which is defined as follows. The first term is the discriminative loss [50] and the second term is the feature style loss [51]:

$$L_{comp} = \sum_{ROI} \lambda_{local}\mathbb{E}_{X'_{ROI}}[log(1 - D_{ROI}(X'_{ROI}))] + \lambda_{f_s} \left\| Gram(\psi(X'_{ROI})) - Gram(\psi(X_{ROI})) \right\|_1, \tag{7}$$

where ROI is region of interest [52] from the component collection $\{left\_eye, right\_eye, mouth\}$. D_{ROI} is the local discriminator for each region. The feature style loss attempts to match the Gram matrix statistics [53] of real and restored patches from multiple layers of the learned local discriminators, which has been demonstrated to be conductive to generating realistic facial details and reducing unpleasant artifacts. Besides, ψ denotes the multi-resolution features from the learned discriminators. λ_{local} and λ_{f_s} represent the loss weights of local discriminative loss and feature style loss, respectively.

Identity Preserving Loss. During the process of high quality generation, the "fake" identity generated in the previous step, i.e. the identity of X_l, must remain as constant as possible. We employ a pretrained state-of-the-art (SOTA) face recognition model [55] to extract identity features. [55] is chosen because it can provide highly discriminative identity features and has a clear geometric interpretation due to the exact correspondence to the geodesic distance on the hypersphere. We use the identity preserving loss $\mathcal{L}_{id}$ to ensure that the identity of X' is the same as X_l:

$$\mathcal{L}_{id} = \lambda_{id}\left(1 - \frac{r_{id}(X') \cdot r_{id}(X_l)}{\|r_{id}(X')\|_2 \cdot \|r_{id}(X_l)\|_2}\right), \tag{8}$$

where r_{id} represents the identity feature extract by [55]. λ_{id} denotes the weight of identity preserving loss. Here we use cosine similarity rather than the original $\mathcal{L}_1$ distance in [31] because we think it better fits the angular margin based identity extractor [55] (and is proved in Sec 4.4).

The overall model objective is a combination of the above losses:

$$L_{total} = L_{rec} + L_{adv} + L_{comp} + L_{id}. \tag{9}$$

The hyper-parameters are set as follows: $\lambda_{l_1} = 0.1$, $\lambda_{per} = 2$, $\lambda_{adv} = 0.1$, $\lambda_{f_s} = 200$ and $\lambda_{id} = 5$.

4. Experiments

4.1. Experimental Setup

Datasets. We choose the FFHQ dataset [56], which contains 70K high-resolution face images with diverse demographic information like age, gender, and race, to train our GAN model in Step II. We randomly select 60K images for training and 10K for testing. All images are aligned and cropped to size 512×512 covering the whole face, as well as some background regions. Moreover, in order to compare with other methods fairly, we also test IDeudemon on the CelebA–HQ dataset [57] and show our generalization ability (see Sec 4.3 for deatils).

Evaluation Metrics. We evaluate the proposed IDeudemon in terms of two metrics, as described below.

(1) Privacy metrics. Following previous work [6], we measure the $\mathcal{L}_2$ distance of embedding vectors from the de-identified and original faces extracted by a pretrained face recognition model, denoted as **DIS**, to evaluate the quality of identity protection. For a fair comparison, we employ two models that are excluded from our training, i.e., the Face Recognition library[1] (denoted as FR), and the FaceNet [54] which is pretrained on two public datasets (CASIA-Webface [58] and VGGFace2 [59]) respectively.

(2) Utility metrics. We evaluate not only the quality of the de-identified images, but also the retention ability to pose and expression. Specifically, **PSNR**, **SSIM** and **FID** are chosen to evaluate the generation quality. PSNR and SSIM are widely-used objective methods to measure the difference between two images, while FID can measure the distance between the generated distribution and the real distribution. Besides, the $\mathcal{L}_2$ distances between pose and expression vectors from the de-identified and original faces extracted by an open-sourced pose estimator [60] and a 3D facial model [61] are calculated as pose (denoted as **POSE**) and expression (denoted as **EXP**) similarity.

Implementation Details. We implement our framework as shown in Fig. 2. Since the value range of the ID code is between [-1, 1], after Step I, the part out of the range needs to be truncated to ± 1, depending on which value is closer. The sizes of different facial codes are $c_{id}, z_{id} \in \mathbb{R}_{100}$, $c_{exp} \in \mathbb{R}_{79}$, $c_{alb} \in \mathbb{R}_{100}$, $c_{illu} \in \mathbb{R}_{27}$ and $z_{app} \in \mathbb{R}_{206}$ respectively. During the training of the GAN model in Step II, the mini-batch size is set to 6. We augment the training data with horizontal flip and color jittering. We train our model with Adam optimizer [62] for a total of 300k iterations. The learning rate was set to 2×10^{-3} and then decayed by a factor of 2 at the 220k-th, 270k-th iterations.

4.2. Protective Perturbation Analysis.

This section analyzes the performance of our IDeudemon with different levels of perturbation applied on

[1] https://github.com/ageitgey/face$_recognition$

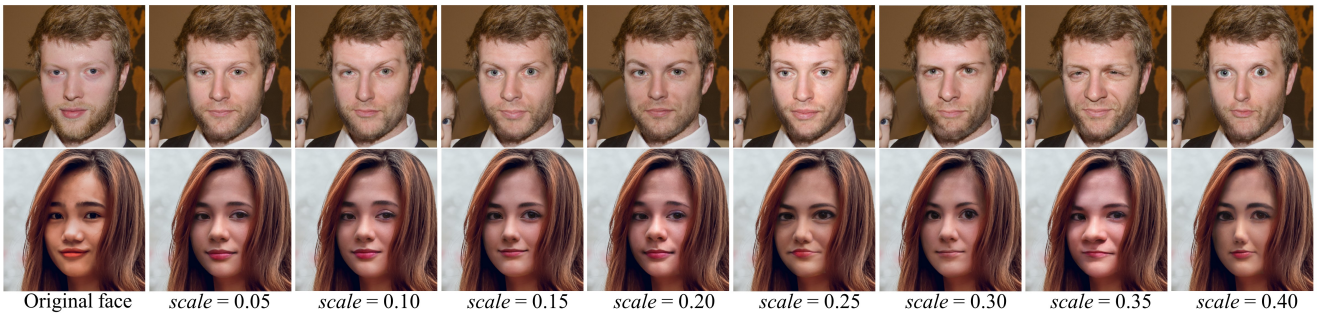

Figure 3. Qualitative results of the influence of the noise *scale* on the **FFHQ**. The first column shows the original face images. The rest columns demonstrate de-identified faces whose identity distances are closest to the mean distance under every *scale* value.

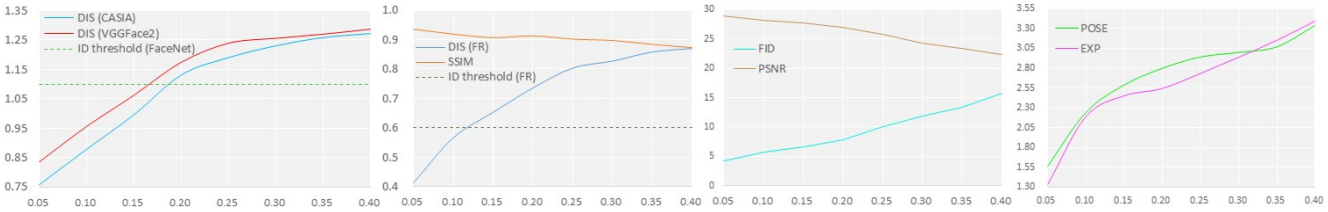

Figure 4. The de-identified performance variation with respect to the noise *scale* on the **FFHQ**. The x-axis indicates the *scale* value and the y-axis indicates different metric values. The identity judgement threshold is 0.6 for Face Recognition library [6] and 1.1 for FaceNet [54].

the original ID code in Step I. The additive Gaussian noise n is sampled from a normal distribution. The *loc* is set to 0, the value of its *scale* belongs to $\{0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40\}$ and the size equals to z_{id}. Ten de-identified faces are generated for every test face image under each *scale* value. Various statistical mean metric results are calculated at each *scale* value.

Fig. 3 shows the qualitative results. It can be observed that with the increase of the noise *scale*, the geometric difference between the de-identified and original faces expands, while the identity-agnostic attributes (hairstyle, background, etc.) are still maintained. The quality of the de-identified images is consistently good, and is almost comparable to the quality of the original images. All synthetic images have sharp details such as eyelashes, wrinkles, teeth, and lips. Quantitative results are shown in Fig. 4. One can see that the degree of identity protection can be adjusted, along with the change of utility. It is worth noting that the utility is kept at a good level (e.g., the FID values are always low). Particularly, we note that when the noise *scale* is smaller than 0.2, the results are too similar to the original faces and the ability to protect identity is not strong; when the *scale* is larger than 0.3, the geometric structure of the faces begins to become exaggerated (such as eccentric eyes, noses, wrinkles and shadows).

Based on the extensive experiments mentioned above, taking into account the visual effects and evaluation metrics comprehensively, we recommend the users to set the *scale* of the protective perturbation between 0.2 and 0.3 to obtain de-identified faces efficiently with well-preserved appearance. We no longer show the case of adding Gaussian noise with larger *scale* values, because the generated faces will be quite visually exaggerated.

4.3. Comparison with State-of-the-art Methods.

To validate the effectiveness of the proposed IDeudemon, we compare it with several SOTA de-identification methods: DeepPrivacy [2], AnonymousNet [4], CIAGAN [3], Gu *et al.* [5], Cao *et al.* [6] and AMT-GAN [16]. For fairness, the test dataset is CelebA–HQ [57] and all images are aligned and cropped to size 256×256.

To test on the dataset, we first bilinearly interpolate the input image to 512×512, and then process it according to the pipeline in Fig. 2. Because (1) the NeRF-based 3D fitting in Step I can still handle the image without photo-realistic details; (2) the GAN model in Step II is trained to process this kind of degradation, our de-identification results are still outstanding in terms of generation quality. The *scale* of protective Gaussian noise is set to 0.25. The final outputs are rescaled to 256×256.

Qualitative results are shown in Fig. 5 (a). One can see that the competing methods fail to produce photo-realistic faces, especially when the original face has a large pose (the last two rows) or expression (the second row). In contrast, our IDeudemon obfuscates the human identities in a perceptually natural manner, meanwhile, the de-identified face still shares similar appearance, as well as the same pose, expression,illumination and background with the original face. It is worth noticing that our results are high-fidelity and can retain clear lips, teeth and even eyelashes, which is superior to other methods.

Figure 5. (a) Qualitative comparison on the **CelebA-HQ** for face de-identification. Our IDeudemon conceals the real identity and produces photo-realistic details at the same time. **Zoom in for best view**. (b) User study results of different de-identification methods.

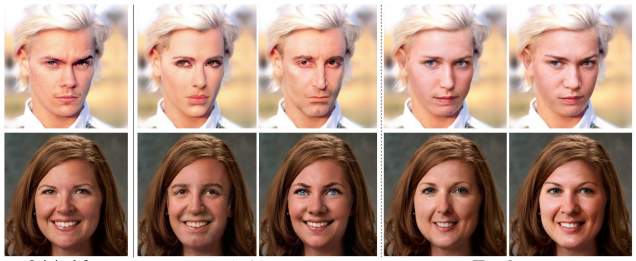

Figure 6. Comparison with StyleFace [8] at megapixel level (1024×1024, from the paper sample image).

Quantitative results are presented in Table 1. Our method obtains the best scores in privacy metrics, clearly confirming our initial motivation that manipulating the 3D parametic ID code can greatly benefit the identity protection. One can see that our IDeudemon achieves comparable PSNR and SSIM indexes to other competing methods, but achieves significantly better results on FID index, which is a better measure for the image perceptual quality. In addition, our method outperforms the other methods in retaining pose and expression. These verify the efficiency of our designs in ensuring utility and make IDeudemon have the least impact on the subsequent use of the de-identified images.

User Study. The de-identified results of comparison methods and our IDeudemon on 100 face images are presented in a random order to 10 volunteers for subjective evaluation. The volunteers are asked to rank the 7 de-identified outputs of each input image according to their perceptual quality. Finally, we collect 7k votes, and the statistics are presented in Fig. 5 (b). As can be seen, our IDeudemon receives much more rank-1 votes than other SOTA methods.

Besides, IDeudemon can conduct face de-identification at megapixel level (inherits from [45]), and we compare it with one of the first high-resolution methods published last year, StyleFace [8] (see Fig. 6). Our results are at least visually as good as the original ones of [8], despite having to run on the cropped faces extracted from the paper PDF.

Table 1. Quantitative comparison with SOTA methods on the **CelebA-HQ**. ↑ means higher is better, and ↓ means lower is better. **Red** and blue indicates the best and the second best performance.

Method	DIS↑			PSNR↑	SSIM↑	FID↓	POSE↓	EXP↓
	FR	CASIA	VGGFace2					
DeepPrivacy [2]	0.783	1.091	1.187	21.3	0.791	24.6	6.22	5.27
AnonymousNet [4]	0.497	0.875	0.936	20.4	0.803	53.7	3.69	4.02
CIAGAN [3]	0.671	0.919	1.085	18.6	0.522	28.1	8.93	5.19
Gu *et al.* [5]	0.812	1.207	1.224	23.1	0.751	39.7	3.95	3.96
Cao *et al.* [6]	0.794	1.206	1.231	24.1	**0.902**	22.6	3.04	2.81
AMT-GAN [16]	0.596	0.927	0.941	21.0	0.799	33.3	3.02	2.86
IDeudemon	**0.819**	**1.228**	**1.233**	**25.9**	0.898	**8.7**	**2.96**	**2.79**

4.4. Model Analysis and Ablation Study

3D Parametric Fitting Method Selection. In the first step of our "divide and conquer" strategy, what we need is a fast, accurate tool that can fit the disentangled facial parameters in 3D space. The NeRF model [30] created last year is the first work to accomplish this task. [30] has verified the validity of each part and its SOTA fitted effect. Therefore, we adopt it for face parametric fitting in Step I. The brilliant de-identification effects of IDeudemon have proven the correctness of this choice.

Ablation Study of Step II. In order to validate the effectiveness of our various designs in Step II, in this section we conduct an ablation study by introducing some variants of our IDeudemon and comparing their performance.

We first pick and train five SOTA face restoration models to respectively replace the GAN model [31] we used as five variants. They are denoted as BOPB [37], GPEN [38], RestoreFormer [34], CodeFormer [36] and VQFR [35]. Then w/o vsa refers to the IDeudemon model without visual similarity assistance. Additionally, we validate the necessity

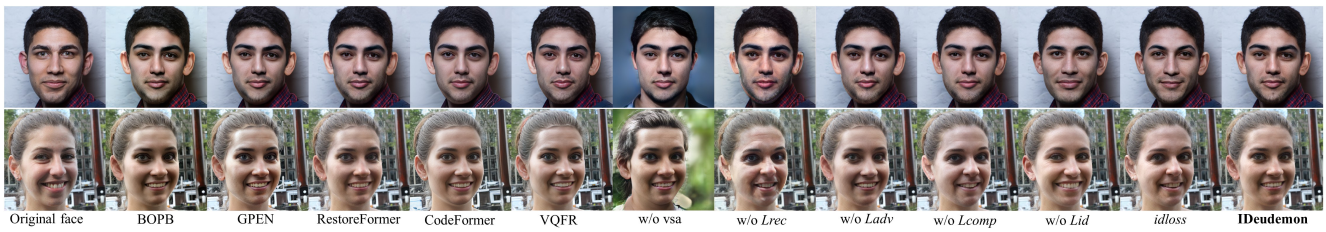

Figure 7. Ablation studies on GAN model, visual similarity assistance and identity preserving loss on the **FFHQ**. **Zoom in for best view**.

Table 2. Ablation study results of Step II on the **FFHQ**. ↑ means higher is better, and ↓ means lower is better. **Red** and blue indicates the best and the second best performance.

Method	DIS↑			PSNR↑	SSIM↑	FID↓	POSE↓	EXP↓
	FR	CASIA	VGGFace2					
BOPB [37]	0.803	1.083	1.224	26.2	0.899	17.63	2.963	2.783
GPEN [38]	0.794	1.186	1.236	24.8	0.895	11.65	2.975	2.772
RestoreFormer [34]	0.802	1.191	1.239	24.6	0.889	11.14	2.956	2.768
CodeFormer [36]	0.801	1.189	1.236	23.8	0.905	10.83	2.950	2.778
VQFR [35]	0.796	1.185	1.238	24.2	0.898	11.95	3.006	2.839
w/o vsa	**0.815**	**1.193**	**1.244**	20.4	0.728	25.78	3.084	3.854
w/o L_{rec}	0.801	1.188	1.236	24.2	0.847	10.07	3.112	2.847
w/o L_{adv}	0.799	1.191	1.231	24.6	0.863	11.53	2.993	2.788
w/o L_{comp}	0.803	1.189	1.237	25.4	0.891	10.12	2.947	2.831
w/o L_{id}	0.417	0.816	0.965	**26.3**	**0.912**	**9.613**	2.973	**2.754**
idloss	0.768	1.079	1.203	25.5	0.901	10.06	2.958	2.762
IDeudemon	0.804	1.192	1.239	25.8	0.903	9.99	**2.942**	2.761

of the loss functions, which are indicated as w/o L_{rec}, w/o L_{adv}, w/o L_{comp} and w/o L_{id}. We specifically calculate the identity preserving loss by using $\mathcal{L}_1$ distance (like [31]) rather than cosine similarity, and denote it as *idloss*.

We perform on the FFHQ dataset to evaluate IDeudemon and its seven variants. After the common Step I, except that w/o vsa takes the X_p as input, the other six variants have X_l as input. Fig. 7 and Table 2 demonstrate the qualitative and quantitative comparisons. One can see that IDeudemon achieves overall better quantitative measures than its variants of high quality generation model. Specifically, BOPB, GPEN, RestoreFormer and VQFR are weak in inpainting the irregular white gaps in X_l, BOPB alters the hue of the image, GPEN and RestoreFormer often suffer from artifacts at face contours, and VQFR sometimes produces blurry details (see the teeth). Although CodeFormer does a good job in filling in the white gaps, it tends to smooth out the whole faces and changes the clothing.

By discarding the visual similarity assistance, the results of w/o vsa cannot retain the identity-agnostic features. For instance, the background, hairstyle, accessories and the clothing. Moreover, artifacts and unnatural splotches appear randomly, which affect the visual perception. Although w/o vsa performs slightly better in identity protection, its utility performance has deteriorated significantly. These imply that visual similarity assistance plays an import role in synthesizing realistic details and preserving utility.

It can be observed that only the complete loss function combination achieves the optimal results. It proves that L_{rec} reduces artifacts and preserves visual similarity, L_{adv} enhances realism, L_{comp} improves clarity in the eyes and mouth, and L_{id} maintains the protected identity. The privacy indicators of *idloss* demonstrate that our adjustment of original identity preserving loss can better protect the human identity.

Overall, IDeudemon shows superior performance to its variants, demonstrating the effectiveness of Step II's architecture and the adjusted identity preserving loss.

5. Discussion

We want to emphasize that, while elements of IDeudemon are built on well-understood 3D reconstruction principles (dating back to Vetter and Blanz) and blind face restoration, our core contribution is new and essential. The key to making IDeudemon jump out of the annoying privacy utility trade-off is the "divide and conquer" idea that protects privacy and preserves utility in two sequential steps, the identity is protected at 3D space through a parametric NeRF model, both of which have not appeared previously in the literature. In addition, we pick the most suitable GAN model and perturbation range for our approach through sufficient experiments. We have also designed visual similarity assistance and adjusted the loss function so as to better finish the de-identification task.

6. Conclusion

In this paper, we propose a novel two-step face de-identification method that conducts "divide and conquer" strategy to solve the challenging privacy utility trade-off problem. By introducing advanced 3D parametric face fitting and obfuscating the disentangled ID code, we hide the real identity and endow the whole model with good explainability. Equipped with the visual similarity assistance and generative prior embedded GAN, our model can produce photo-realistic de-identified faces, allowing us to adjust the protection level while keeping good image utility. Extensive experiments demonstrate the superior capability of IDeudemon in face de-identification, outperforming prior arts.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities, STCSM under Grant 22DZ2229005, 111 project BP0719010.

References

[1] Prachi Agrawal and PJ Narayanan. Person de-identification in videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(3):299–310, 2011.

[2] 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.

[3] 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.

[4] Tao Li and Lei Lin. Anonymousnet: Natural face de-identification with measurable privacy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[5] Xiuye Gu, Weixin Luo, Michael S Ryoo, and Yong Jae Lee. Password-conditioned anonymization and deanonymization with face identity transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 727–743. Springer, 2020.

[6] Jingyi Cao, Bo Liu, Yunqian Wen, Rong Xie, and Li Song. Personalized and invertible face de-identification by disentangled identity information manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3334–3342, 2021.

[7] Yunqian Wen, Bo Liu, Ming Ding, Rong Xie, and Li Song. Identitydp: Differential private identification protection for face images. *Neurocomputing*, 501:197–211, 2022.

[8] Yuchen Luo, Junwei Zhu, Keke He, Wenqing Chu, Ying Tai, Chengjie Wang, and Junchi Yan. Styleface: Towards identity-disentangled face generation on megapixels. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 297–312. Springer, 2022.

[9] 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.

[10] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Faceless person recognition: Privacy implications in social media. In *European Conference on Computer Vision*, pages 19–35. Springer, 2016.

[11] Elaine M Newton, Latanya Sweeney, and Bradley Malin. Preserving privacy by de-identifying face images. *IEEE transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005.

[12] Ralph Gross, Latanya Sweeney, Fernando De la Torre, and Simon Baker. Model-based face de-identification. In *2006 Conference on computer vision and pattern recognition workshop (CVPRW'06)*, pages 161–161. IEEE, 2006.

[13] Amin Jourabloo, Xi Yin, and Xiaoming Liu. Attribute preserved face de-identification. In *2015 International conference on biometrics (ICB)*, pages 278–285. IEEE, 2015.

[14] Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu, Yuefeng Chen, and Hui Xue. Towards face encryption by generating adversarial identity masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3897–3907, 2021.

[15] Yaoyao Zhong and Weihong Deng. Opom: Customized invisible cloak towards face privacy protection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[16] Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. Protecting facial privacy: generating adversarial identity masks via style-robust makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15014–15023, 2022.

[17] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[18] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1155–1164, 2019.

[19] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos Zafeiriou. Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9269–9284, 2021.

[20] Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20333–20342, 2022.

[21] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[22] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1576–1585, 2017.

[23] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

[24] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[25] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021.

[26] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021.

[27] Pramod Rao, BR Mallikarjun, Gereon Fox, Tim Weyrich, Bernd Bickel, Hanspeter Pfister, Wojciech Matusik, Ayush Tewari, Christian Theobalt, and Mohamed Elgharib. Vorf: Volumetric relightable faces. 2022.

[28] Daoye Wang, Prashanth Chandran, Gaspard Zoss, Derek Bradley, and Paulo Gotardo. Morf: Morphable radiance fields for multiview neural head modeling. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022.

[29] Stathis Galanakis, Baris Gecer, Alexandros Lattas, and Stefanos Zafeiriou. 3dmm-rf: Convolutional radiance fields for 3d face modeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3536–3547, 2023.

[30] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022.

[31] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9168–9178, 2021.

[32] Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee K Wong. Progressive semantic-aware style transformation for blind face restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11896–11905, 2021.

[33] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2492–2501, 2018.

[34] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17512–17521, 2022.

[35] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, pages 126–143. Springer, 2022.

[36] Shangchen Zhou, Kelvin CK Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *arXiv preprint arXiv:2206.11253*, 2022.

[37] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2747–2757, 2020.

[38] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 672–681, 2021.

[39] Mohammad Rami Koujan, Michail Christos Doukas, Anastasios Roussos, and Stefanos Zafeiriou. Head2head: Video-based neural head synthesis. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 16–23. IEEE, 2020.

[40] Yudong Guo, Lin Cai, and Juyong Zhang. 3d face from x: Learning face shape from diverse sources. *IEEE Transactions on Image Processing*, 30:3815–3827, 2021.

[41] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7346–7355, 2018.

[42] 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.

[43] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[44] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, 129:3051–3068, 2021.

[45] 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.

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[47] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[48] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[49] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[50] 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.

[51] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[52] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[53] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[54] 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.

[55] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[56] 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.

[57] 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.

[58] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[59] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[60] Cho-Ying Wu, Qiangeng Xu, and Ulrich Neumann. Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry. In *2021 International Conference on 3D Vision (3DV)*, 2021.

[61] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021.

[62] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.