Data Synthesis with Diverse Styles for Face Recognition via 3DMM-Guided Diffusion

Yuxi Mi[1] Zhizhou Zhong[1] Yuge Huang[2†] Qiuyang Yuan[1] Xuan Zhao[1] Jianqing Xu[2]
Shouhong Ding[2] Shaoming Wang[3] Rizen Guo[3] Shuigeng Zhou[1†]
[1] Fudan University [2] Youtu Lab, Tencent [3] WeChat Pay Lab33, Tencent
{yxmi20, sgzhou}@fudan.edu.cn, {zzzhong22, qyyuan23, xzhao23}@m.fudan.edu.cn
{yugehuang, joejqxu, ericshding}@tencent.com
{mangosmwang, rizenguo}@tencent.com

Abstract

Identity-preserving face synthesis aims to generate synthetic face images of virtual subjects that can substitute real-world data for training face recognition models. While prior arts strive to create images with consistent identities and diverse styles, they face a trade-off between them. Identifying their limitation of treating style variation as subject-agnostic and observing that real-world persons actually have distinct, subject-specific styles, this paper introduces MorphFace, a diffusion-based face generator. The generator learns fine-grained facial styles, e.g.*, shape, pose and expression, from the renderings of a 3D morphable model (3DMM). It also learns identities from an off-the-shelf recognition model. To create virtual faces, the generator is conditioned on novel identities of unlabeled synthetic faces, and novel styles that are statistically sampled from a real-world prior distribution. The sampling especially accounts for both intra-subject variation and subject distinctiveness. A context blending strategy is employed to enhance the generator's responsiveness to identity and style conditions. Extensive experiments show that MorphFace outperforms the best prior arts in face recognition efficacy*[*].

1. Introduction

Face recognition (FR) is among the most successful computer vision applications, where persons are identified by model-extracted facial features. FR models are well known for being data-hungry. Their efficacy is built upon large-scale face image training datasets [11, 26, 92] that contain rich identities and diverse styles, *e.g.*, appearance variations in age, expression and pose. Contemporarily, open-source face image datasets are primarily collected by crawling from the web. The images are potentially enrolled without the informed consent of individuals, which yields serious legal and ethical issues regarding data privacy.

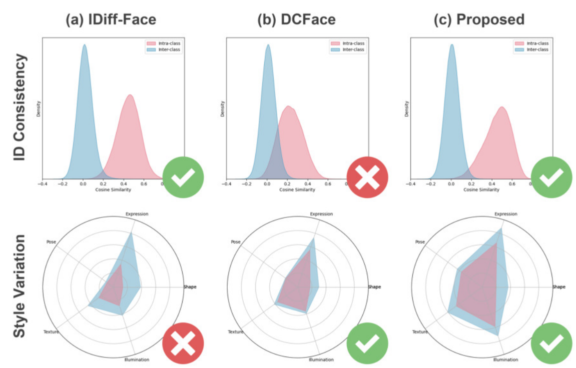

Figure 1. Analyses for identity consistency and style variation across prior arts and our proposed MorphFace. Identity consistency is measured by pairwise cosine similarity and style variation by variances of DECA attributes. Intra-class and inter-class results are represented in red and blue, respectively. Separated curves and a larger shaded area indicate better consistency and variation. Prior arts bear inadequacies in either (a) style variation or (b) identity retention, while (c) MorphFace achieves both goals simultaneously.

Identity-preserving face synthesis (IPFS) offers a remedy to the privacy issue. Its objective is to generate face images of virtual subjects and replicate the distribution of real face images so that FR models can be trained on these synthetic faces to effectively recognize real persons. Among previous efforts, early works [5, 8, 10, 43, 63] are mainly based on generative adversarial networks (GAN) that yet produce face images with limited quality. Recent studies [7, 42, 58] employ diffusion models (DM) to generate faces of massive unique subjects with fine-grained details.

The primary challenge of IPFS was to generate multiple faces for the same person. It is recently realized by conditioning a DM's denoising on the person's identity context.

[†] Corresponding authors.
[*] Code will be available at https://github.com/Tencent/TFace/.

We examine the synthetic faces of a related prior art, IDiff-Face [7], in Fig. 1(a). We measure the cosine similarity between their FR-extracted embeddings and find high *identity consistency* within each subject. Nonetheless, these images are found analogous and lack *style variation* that could help FR generalize. Recent works [42, 47] consider style as an additional DM condition that can be uniformly sampled from external sources, *e.g.*, style banks or pre-trained models. In Fig. 1(b), we use a DECA [21] 3DMM model to extract style variances of images from DCFace [42] and observe more varied styles. However, we infer from the similarity metric that their style control negatively impacts identity retention. We refer to this phenomenon as the trade-off between intra-class identity consistency and style variation.

We advocate a paradigm change to create synthetic datasets with both consistent identities and diverse styles. Prior works treat style as a subject-agnostic factor, applying uniform style control across the entire dataset. However, we observe a key divergence from reality in their approach, as they overlook the *distinctiveness of subjects*. In real-world datasets [11, 26, 92], images from different subject classes often exhibit distinct styles. For example, individuals from different gender groups typically display different facial shape variations [48]. We propose to promote subject distinctiveness in our synthetic faces, which offers two advantages: (1) This enriches dataset variability by combining intra-class style variation with subject-specific styles, without compromising identity consistency; (2) This helps mitigate overfitting to potentially biased styles, allowing FR models to focus on learning identity.

Concretely, we first present a more fine-grained and realistic approach to style control. We use DECA [21] 3DMM to parameterize 3D geometry and facial appearance from an image into attribute sets, and render them into style feature maps. To generate synthetic faces with designated identities and styles, we employ FR-extracted identity embeddings and style feature maps as a DM's context. We employ 3DMM for two reasons: (1) It effectively expresses style in synthetic images; (2) It provides precise, fully parametric control over facial style by adjusting the style attributes. To generate novel faces, we sample style attributes from real-world prior distributions through a *subject-aware strategy*, which explicitly accounts for both intra-class variation and subject distinctiveness. Since we incorporate both identity and style controls during face generation, another key challenge is the effective integration of these two contexts. Based on observations of the DM's denoising process, where styles are primarily established before identity, we propose *context blending* that reweights the style and identity contexts at appropriate denoising timesteps.

We concretize our findings into a novel IPFS generator, MorphFace, named for its ability to morph facial styles through 3DMM renderings. Experimentally, we find that MorphFace achieves a Pareto improvement in balancing intra-class consistency and variation, as shown in Fig. 1(c). It also significantly enhances FR efficacy, outperforming the best prior methods across all test benchmarks.

This paper presents three-fold contributions:

- We present a novel IPFS generator that creates synthetic faces with consistent identities and rich styles. It provides fine-grained style control via 3DMM renderings.
- We propose subject-aware sampling that promotes intra-class style variation and subject distinctiveness, and context blending that enhances context expressiveness.
- We conduct extensive experiments that demonstrate the state-of-the-art (SOTA) efficacy of our approach.

2. Related Work

Face recognition aims to match queried face images to an enrolled database. SOTA FR is established on deep neural networks [6, 27, 31], trained using margin-based softmax losses [4, 18, 34, 41, 77] on large-scale datasets [11, 26, 32, 39, 92]. Despite the datasets' vital contribution, they often face legal and ethical disputes for being web-crawled without consent [26]. They also exhibit quality problems such as noisy labels and long-tail distributions [85]. FR's performance is measured on benchmark datasets [55, 69, 88, 89] that capture real-world variations, *e.g.*, pose and age.

Face image synthesis is a long-standing task that has yielded numerous impressive results. Pioneering works use style-based GANs [35, 37, 38, 51, 56], 3D priors [17, 25, 33, 40, 51, 56, 61, 81], or semantic attribute annotations [19, 70, 71, 75] to generate images with specific facial attributes [24] or to manipulate existing reference images [72]. Recent approaches primarily leverage diffusion models [29, 65, 73] to generate subject-conditioned images. Among these, tuning-based methods personalize a pre-trained DM (*e.g.*, Stable Diffusion [65]) on a few images [20, 22, 68], extracted features [30, 78, 86], or textual descriptions [23, 91] of a specific subject, to produce images that reflect that subject's identity. Other methods, in contrast, train DMs typically conditioned on subject-descriptive features [12, 50, 80] from CLIP [64], FR-extracted identity embeddings [13, 60, 76], or them combined [82]. These methods have promoted not only data creation [14, 46] but also related tasks [54, 83, 84, 90]. However, they prioritize high image fidelity over the distinctiveness of subjects. They are less suitable for producing FR training data due to ambiguity in identity retention.

Face recognition with synthetic images offers benefits in both privacy and quality for FR training [15, 16, 52]. Closest to our study, recent works aim to generate multiple synthetic face images for each subject, unseen in real datasets, to replace real images in FR training. We refer to these methods as *identity-preserving face synthesis*. Specifically, SynFace [63], SFace [5], SFace2 [10], ID-

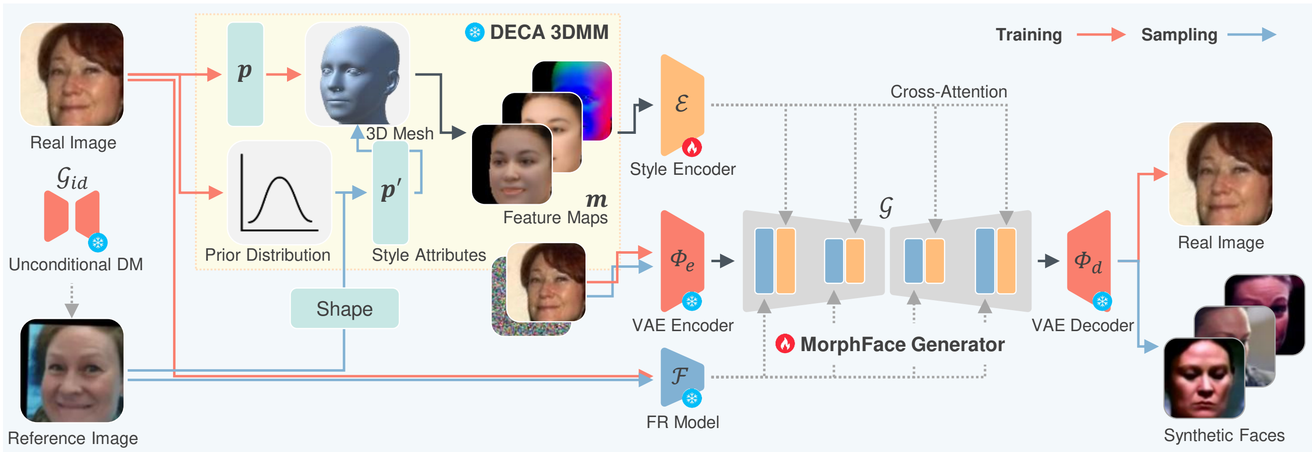

Figure 2. Pipeline of MorphFace. It uses a pair of style and identity contexts to generate faces with designated identity and diverse style. Style is extracted using DECA 3DMM to provides fine-grained, entirely parametric control. To sample virtual faces, unlabeled synthetic images are used as subject reference, and style is sampled statistically for real-world prior distribution.

Net [43] and ExFaceGAN [8] use varied GAN architectures [19, 36] in subject-conditioned settings, while USynthFace [9] uses unlabeled images to improve FR training. DigiFace [1] utilizes a 3D parametric model to produce distinctive yet less realistic faces. IDiff-Face [7], DCFace [42], Arc2Face [58], CemiFace [74] and ID3 [47] are diffusion-based latest works. Most of them [42, 47, 58, 74] explicitly promote style variations during DM's sampling process to improve FR generalization. This paper outperforms them largely by offering more precise and realistic style control.

3. Proposed Approach

Overview. We introduce MorphFace, a face generator that produces synthetic face images with consistent identities and varied styles. Our approach is fueled by a latent diffusion model (LDM) [65]. To preserve identity, we condition the LDM on FR-extracted identity embeddings. To vary styles, while prior arts have employed style banks [42], similarity metrics [74], and attribute predicates [47] to coarsely promote style variation, they are unable to control specific style attributes. In contrast, we use 3DMM renderings as the LDM's style contexts. We gain more precise control over style since the renderings provide entirely parametric style descriptions.

To generate unseen face images, we are required to sample novel identities and style contexts. For identity, we obtain reference images of virtual subjects using unlabeled faces from an additional pre-trained DM. For style, we sample 3DMM style attributes in a manner that considers both intra-class style variation and subject distinctiveness, to better mimic real-world style variations. This also differentiates our approach from prior arts [42, 47, 74] which typically apply uniform style control. Experimentally, we find our subject-aware style sampling significantly enhances FR efficacy. We further augment the style and identity contexts during the LDM's certain denoising phases to improve their expressiveness. Figure 2 illustrates our pipeline.

3.1. Preliminary

Latent diffusion models [65] are generative models trained to predict the latent representations $\mathbf{z}$ of input images $\mathbf{x}$ via a gradual denoising process. Let ϕ_e, ϕ_d be a pair of pre-trained encoder and decoder. The image $\mathbf{x}$ is mapped into a latent space as $\mathbf{z}{=}\phi_e(\mathbf{x})$, then is corrupted by variance-controlled Gaussian noise ϵ over $0{\leq}t{\leq}T$ timesteps,

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{1}$$

where $\mathbf{z}_0$ stands for the clean latent representation, α_i is from a linear variance schedule, and $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$. In the denoising process, the model attempts to recover $\mathbf{z}_{t-1}$ iteratively through following transition,

$$\mathbf{z}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{z}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})\right) + \sqrt{1-\alpha_t}\epsilon, \tag{2}$$

where $\mathbf{c}$ is a context condition such as identity or style. The image is recovered as $\mathbf{x}{=}\phi_d(\mathbf{z}_0)$. The transition is parameterized by a noise estimator ϵ_θ (*e.g.*, U-Net [66]) trained with the minimization of an l_2 objective,

$$\mathcal{L} = \mathbb{E}_{z_t, t, \epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{1})}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})\|_2^2\right]. \tag{3}$$

3D morphable face models [2] are parametric models that represent faces in a compact latent space. Among them, FLAME [48] uses linear blend skinning to create a 3D mesh of vertices that describes facial geometry, including *shape*, *pose*, and *expression*. DECA [21] incorporates FLAME with additional encoders to further provide facial

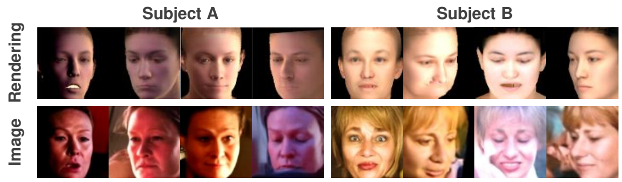

Figure 3. Sample 3DMM feature maps (here, Lambertian renderings) and their synthetic images. Sec. 3.2: Precise style control and more fine-grained detail can be observed in generated images. Sec. 3.3: Sampling subject-aware styles create renderings and images with subjective distinctiveness (*e.g.*, illumination).

appearance descriptions, including *texture* and *illumination*, through Lambertian reflectance and spherical harmonics lighting. It produces a set of numerical parameters that determinantly model them as style attributes, which can be rendered into feature maps such as surface normals, albedo, and Lambertian rendering. We use DECA, *wlog.*, as our 3DMM foundation model. For further details, we refer the reader to the latest 3DMM survey paper [49].

3.2. 3DMM-Guided Face Synthesis

LDM is by design capable of unlabeled face generation. We first condition an LDM $\mathcal{G}$ on identity embeddings to let it generate faces of specific subjects. Concretely, let $\mathbf{X}$ denote the real face image dataset on which we train the LDM. We extract its images' identity embeddings via a pretrained FR model [4] $\mathcal{F}$ as $\mathbf{c}_{id}{=}\mathcal{F}(\mathbf{x})$, and incorporate $\mathbf{c}_{id}$ into the LDM's training process, Eq. (3), as context through cross-attention. Notably, this approach is conceptually similar to IDiff-Face [7]. Figure 1(a) has shown that such generated faces bear insufficiency in intra-class style variation. We consider this as a baseline to compare with following approach.

We further condition the LDM on 3DMM renderings to promote style variation. 3DMM provides fully parametric descriptions for multiple attributes of facial styles, including shape, expression, pose, texture and illumination. This enables us to precisely control the style of specific face images based on 3DMM's parameters, an unachieved goal of prior arts [42, 47, 74].

Specifically, given input images $\mathbf{x}$, we employ an open-source DECA [21] 3DMM model $\mathcal{M}$ to infer their style attributes, $\mathbf{p}{=}\mathcal{M}(\mathbf{x})$. The style attributes are 100,50,9,50,27-dim numerical parameters with human-interpretable meanings for image-wise shape, expression, pose, texture and illumination, respectively. We can concatenate them into a 236-dim vector. Using Lambertian reflectance as part of DECA's integration, we render three feature maps $\mathbf{m}$ entirely parameterized by style attributes $\mathbf{p}$—surface normals, albedo, and Lambertian rendering. The parametric nature will facilitate the sampling of novel styles, illustrated later in Sec. 3.3. From Fig. 2, we find that the feature maps provide pixel-aligned style descriptions of the input images yet are absence of facial details. We use them to condition the LDM to produce real-looking faces: We concatenate $\mathbf{m}$ along channels and pass them through a simple encoder $\mathcal{E}$ trained end-to-end with the LDM to obtain style embeddings $\mathbf{c}_{sty}{=}\mathcal{E}(\mathbf{m})$, and optimize the LDM using both identity and style embeddings as contexts,

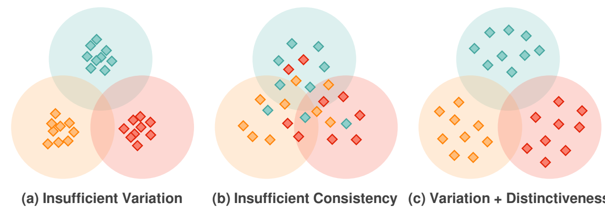

Figure 4. Illustration of style distribution. Regions represent real-world style distributions and diamonds represent samples. (a) Insufficient style variation impairs FR generality. (b) Uniformly sampling styles yields a "mixed" distribution that obscure identity consistency. (c) In our proposed approach, style and identity are both promoted by considering the distinctiveness of subjects.

$$\mathcal{L} = \mathbb{E}_{z_t,t,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{1})}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c}_{id}, \mathbf{c}_{sty})\|_2^2\right]. \quad (4)$$

To demonstrate our generator's context control, Fig. 3 shows sample synthetic images based on their 3DMM renderings. These images are of high quality and effectively preserve the renderings' style. Unlike prior works, our approach provides explicit, image-wise style control.

We further distinguish our approach from two close prior arts: DigiFace [1] also employs 3DMM for IPFS. However, it directly outputs coarse 3DMM renderings as face images, whereas we incorporate the LDM to generate more realistic faces. DiffusionRig [20] performs face editing that includes 3DMM as style control. It yet requires burdened subject-wise fine-tuning, and its identity retention is easily nullified upon changing style. It is hence less suitable for IPFS.

3.3. Synthetic Face Generation

We discuss how to sample novel identities and styles for synthetic face image generation using our trained LDM.
Novel identities. We employ an unconditional DM $\mathcal{G}_{id}$ to produce unlabeled face images. To improve the images' diversity, we filter them by a cosine similarity threshold of 0.3 on their FR-extracted identity embeddings [4] and by image quality assessed via SDD-FIQA [57]. We use the cleaned images as references for novel subject classes.
Novel styles. Since the feature maps $\mathbf{m}$ are entirely parameterized by style attributes $\mathbf{p}$, we can produce novel styles by sampling new style attributes $\mathbf{p}'$. To mimic real-world style variations, we propose to sample $\mathbf{p}'$ statistically from the prior style distribution of LDM training dataset. Formally, let $\mathbf{P}{=}\mathcal{M}(\mathbf{X})$ be the style attribute set of $\mathbf{X}$, and $\mathbb{D}(\mathbf{P})$ be its distribution. The general form of sampling $\mathbf{p}'$ is as

$$\mathbf{p}' \in \mathbf{P}', \quad \mathbf{P}' \sim \mathbb{D}(\mathbf{P}). \tag{5}$$

We note that $\mathbb{D}(\mathbf{P})$ can be approximated as a multiplicative Gaussian distribution, *i.e.*, $\mathbb{D}(\mathbf{P}){\sim}\mathcal{N}(\mu, \mathbf{\Sigma})$, where μ and $\mathbf{\Sigma}$ represent the mean and covariance matrix of $\mathbf{P}$. This approximation is grounded by the nature of 3DMM [2] and prior studies' findings [3, 59], and is empirically validated. We leave further discussion to the supplementary material.

Equation (5) does not specify how each $\mathbf{p}'$ is sampled from $\mathbf{P}'$. Prior arts [42, 47, 74] mainly offer uniform sampling, *i.e.*, providing subject-agnostic style context to each synthetic image. Similarly, we can uniformly sample styles by rewriting Eq. (5) as $\mathbf{p}'{\sim}\mathcal{N}(\mu, \mathbf{\Sigma})$. However, in Sec. 4.3, we find this means yields suboptimal FR efficacy.

We propose an improved strategy to better replicate real-world style variations by considering both *intra-class style variation* and *style distinctiveness of subjects*. Intra-class style variation imposes a seeming dilemma: Its insufficiency may impair FR generality [7], yet its excessiveness also reduces FR efficacy since this may obscure the retention of identities [42], as illustrated in Fig. 4.

While prior works advocate uniform style variations, our key observation from real-world datasets [26, 85, 92] reveals that each subject actually exhibits style distinctiveness that should be considered. For instance, women and men often possess different facial shapes [48]; A juvenile may have more youthful photos enrolled in a dataset than an elderly individual, creating age-related distinctions. We believe that such subject-specific distinctiveness plays a crucial role in dataset quality: It enhances dataset variability with less negative impacts on identity consistency, and helps FR models mitigate potential overfitting on biased styles.

We propose *subject-aware style sampling*, concretized from Eq. (5), based on the observation. To address subject distinctiveness, we first sample class-wise distribution from the style attribute set $\mathbf{P}'$. Then, we sample image style from its class distribution to allow intra-class variation. Formally, let $\mathbf{P}'{=}\bigcup_{i=1}^{m}\mathbf{P}'_i$ be a division of $\mathbf{P}'$, where m is the number of unique subjects. We sample $\{\mathbf{P}'_i\}_{i\in[m]}$ as

$$\mathbf{P}'_i \sim \mathcal{N}(\mu_i, \mathbf{\Sigma}_i), \quad \sum_{i=1}^{m}\gamma_i\mathcal{N}(\mu_i, \mathbf{\Sigma}_i) = \mathcal{N}(\mu, \mathbf{\Sigma}), \tag{6}$$

where $\sum_i \gamma_i{=}1$ if each class contains an equal number of images. Each class's μ_i and $\mathbf{\Sigma}_i$ vary by real-world distributions of class means and covariances. We then sample $\mathbf{p}'$ from class-wise distribution,

$$\mathbf{p}' \in \mathbf{P}'_i, \quad \mathbf{p}' \sim \mathcal{N}(\mu_i, \mathbf{\Sigma}_i). \tag{7}$$

Additionally, we find that using facial geometry similar to the subject's reference image can improve identity consistency. It is achieved by replacing the intra-class mean μ_i of facial shape attributes with the reference image's ground truth. In Fig. 3, we produce feature maps that vary adequately within each subject and more significantly across subjects, better reflecting real-world scenarios.

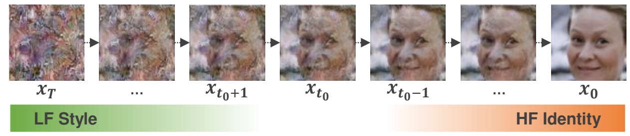

Figure 5. During denoising, LF styles (*e.g.*, pose and shape) are earlier established than HF identity details by the nature of DM. We augment style and identity contexts before and after a shifting timestep t_0 via CFG, respectively, to improve their expressiveness.

3.4. Context Blending

As $\mathcal{G}$ is conditioned on both identity and style contexts, we discuss their effective integration. We empirically find that the guidance of $\mathbf{c}_{id}$ and $\mathbf{c}_{sty}$ can slightly contradict each other during image generation, due to the inherent tension between identity and style. To demonstrate, later in Sec. 4.4, we separately strengthen $\mathbf{c}_{id}$ or $\mathbf{c}_{sty}$ via classifier-free guidance [28] (CFG), an inference-time method for context augmentation, and find the generated images exhibit reduced style variation and identity consistency, respectively.

To improve the contexts' expressiveness, we investigate DM's denoising process from a frequency perspective. DMs are known to favor specific frequency components at certain denoising timesteps: Low-frequency (LF) components are emphasized in early timesteps, while high-frequency (HF) details are progressively refined [62, 67]. In our generator, identity and style contexts align with HF and LF features, respectively: Prior works indicate that facial identity $\mathbf{c}_{id}$ is largely captured with HF details [53, 79], while $\mathbf{c}_{sty}$ mainly consists coarse LF features from 3DMM rendering. As shown in Fig. 5, step-wise denoising reveals that styles (*e.g.*, pose and illumination) are established very early, while facial identity emerges in later steps.

Based on the observation, we propose *context blending* to enhance the guidance of either context at its appropriate denoising timesteps. Specifically, we strengthen $\mathbf{c}_{sty}$ in earlier timesteps and $\mathbf{c}_{id}$ in later timesteps to improve the LDM's responsiveness to these contexts. Formally, during training, we first probabilistically replace $\mathbf{c}_{id}$ and $\mathbf{c}_{sty}$ with learnable empty contexts $\mathbf{c}_{id}^{\emptyset}$ and $\mathbf{c}_{sty}^{\emptyset}$; during inference time, we employ CFG for context augmentation. We rewrite Eq. (2) in CFG-form as

$$\mathbf{z}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{z}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_{cfg}\right) + \sqrt{1-\alpha_t}\epsilon, \tag{8}$$

where ϵ_{cfg} is weighted by w as

$$\epsilon_{cfg} = (1+w)\epsilon_\theta(\mathbf{z}_t, t, \mathbf{c}_{id}, \mathbf{c}_{sty}) - w\epsilon_t. \tag{9}$$

Method	Venue	Volume (IDs × imgs)	LFW	CFP-FP	AgeDB	CPLFW	CALFW	Avg.
CASIA	(real)	0.49M (10.5K × 47)	99.38	96.91	94.50	89.78	93.35	94.79
SynFace	ICCV 21	0.5M (10K × 50)	91.93	75.03	61.63	70.43	74.73	74.75
SFace	IJCB 22	0.6M (10K × 60)	91.87	73.86	71.68	77.93	73.20	77.71
DigiFace	WACV 23	0.5M (10K × 50)	95.40	87.40	76.97	78.87	78.62	83.45
IDnet	CVPR 23	0.5M (10K × 50)	84.83	70.43	63.58	67.35	71.50	71.54
DCFace	CVPR 23	0.5M (10K × 50)	98.55	85.33	89.70	82.62	91.60	89.56
IDiff-Face	ICCV 23	0.5M (10K × 50)	98.00	85.47	86.43	80.45	90.65	88.20
ExFaceGAN	IJCB 23	0.5M (10K × 50)	93.50	73.84	78.92	71.60	82.98	80.17
SFace2	BIOM 24	0.6M (10K × 60)	94.62	76.24	74.37	81.57	72.18	79.80
Arc2Face	ECCV 24	0.5M (10K × 50)	98.81	91.87	90.18	85.16	92.63	91.73
ID3	NeurIPS 24	0.5M (10K × 50)	97.68	86.84	91.00	82.77	90.73	89.80
CemiFace	NeurIPS 24	0.5M (10K × 50)	99.03	91.06	91.33	87.65	92.42	92.30
MorphFace	(ours)	0.5M (10K × 50)	**99.25**	**94.11**	**91.80**	**88.73**	**92.73**	**93.32**
DigiFace	WACV 23	1.2M (10K × 72, 100K × 5)	96.17	89.81	81.10	82.23	82.55	86.37
DCFace	CVPR 23	1.2M (20K × 50, 40K × 5)	98.58	88.61	90.07	85.07	92.82	91.21
Arc2Face	ECCV 24	1.2M (20K × 50, 40K × 5)	98.92	94.58	92.45	86.45	93.33	93.15
MorphFace	(ours)	1.2M (24K × 50)	**99.35**	**94.77**	**93.27**	**90.07**	**93.40**	**94.17**

Table 1. Comparsion with SOTAs by FR recognition accuracy. Our proposed MorphFace outperforms SOTAs on all benchmarks.

We choose a time-varying ϵ_t as $\epsilon_\theta(\mathbf{z}_t, t, \mathbf{c}_{id}, \mathbf{c}_{sty}^{\emptyset})$ for $t \in (t_0, T]$ to augment style, and as $\epsilon_\theta(\mathbf{z}_t, t, \mathbf{c}_{id}^{\emptyset}, \mathbf{c}_{sty})$ for $t \in [0, t_0]$ to augment identity, where t_0 is a "shifting" timestep. Section 4.4 shows that context blending improves identity consistency and style variation, and enhances FR efficacy.

4. Experiments

4.1. Experimental Setup

Datasets. We train our LDM $\mathcal{G}$ on CASIA-WebFace [85], a dataset that consists of 490k quality-varying face images from 10575 identities. We benchmark our FR model $\mathcal{F}_{syn}$ on 5 widely used test datasets, LFW [45], CFP-FP [69], AgeDB [55], CPLFW [88], and CALFW [89]. CFP-FP and CPLFW are designed to measure the FR in cross-pose variations, and AgeDB and CALFW are for cross-age variations.

4.2. Comparison with SOTAs

Recognition accuracy. We generate synthetic datasets using trained $\mathcal{G}$. We synthesize 2 data volumes: 0.5M/1.2M face images from 10K/24K subjects with 50 images for each subject. We train an IR-50 FR model $\mathcal{F}_{syn}$ on our synthetic datasets and compare IPFS SOTAs [5, 7, 8, 10, 42, 43, 47, 58, 63, 74] discussed in Sec. 2. We benchmark them on 5 widely used test datasets by FR recognition accuracy in Tab. 1. Note that some SOTAs may have larger datasets for the generator [58], larger FR backbones [74], and real-world reference images [42] that could benefit their results.

We highlight several key points: (1) MorphFace outperforms *all* SOTAs on *all* test datasets for both 0.5M/1.2M volumes. Notably, we outperform the best SOTA for 2.24 on CFP-FP, 1.08 on CPLFW, and 1.02 on average. As CFP-FP and CPLFW are both pose-varying datasets, this suggests MorphFace could be especially beneficial for cross-pose settings. (2) Our average result of 0.5M outperforms the 1.2M results of SOTAs, demonstrating our approach's high capability. (3) Our 1.2M result achieves on-par performance on CPLFW and CALFW with the real-world CASIA [85]. (4) DM-based methods [7, 42, 47, 58, 74] all ex-

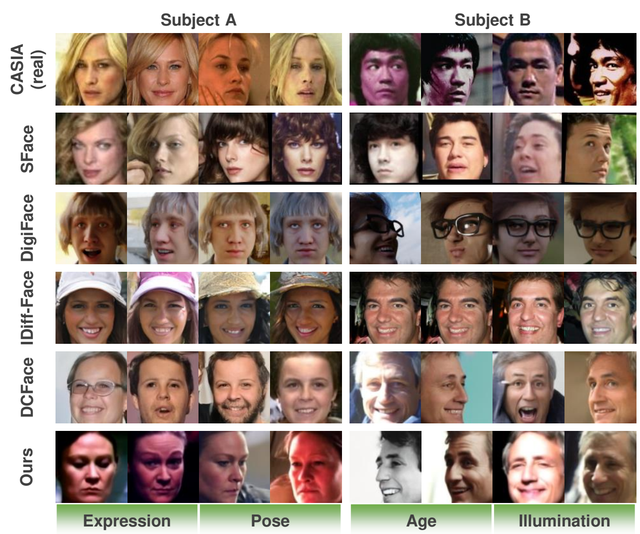

Figure 6. Image visualization for MorphFace and SOTAs. Our approach produces faces with intra-class variation and subject distinctiveness of style. It better replicates real-world style variations.

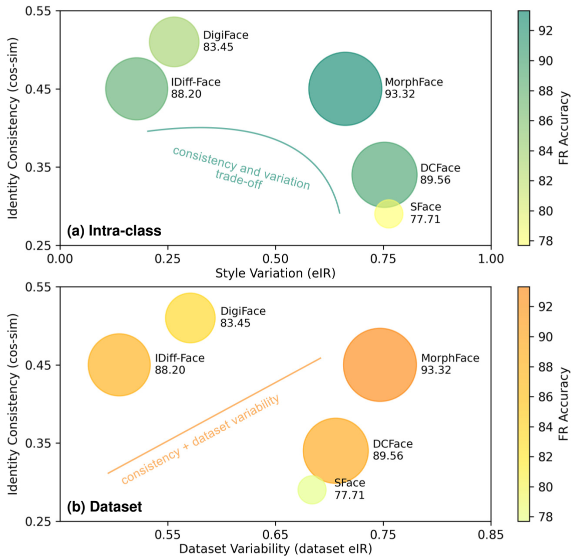

Figure 7. Comparison among 5 methods by consistency and variation metrics. Circle colors and sizes depict FR accuracy. (a) MorphFace outperforms SOTAs in consistency and variation trade-off. (b) It promotes both consistency and datasets' overall variability.

hibit quite satisfactory FR efficacy, which may be attributed to better generations of identity-reflecting HF facial details.

Visualization. We compare CASIA, several SOTAs that released their datasets, and MorphFace. In Fig. 6, we sample 8 images of 2 subjects from each dataset. We highlight: (1) **SFace** [5] preserves less consistent identities; (2) **DigiFace** [1] directly uses 3DMM renderings as images, producing less realistic faces. It yet better represents accessories (*e.g.*, glasses); (3) **IDiff-Face** [7] lacks intra-class variation, producing mainly frontal faces; (4) **DCFace** [42] largely promotes style variation. However, some attributes (*e.g.*, expression) are replicated across its subjects, suggesting less distinctiveness and overfitting to biased styles. It also occasionally creates artifacts (*e.g.*, gender transition) due to degraded identity consistency; (5) **MorphFace** promotes intra-class style variations including expression, pose, age and illumination, and also creates more distinctive subjects. It better mimics the style variations of real-world datasets.

Consistency *vs.* Variation. We quantitatively investigate the balance between intra-class identity consistency and style variation. We calculate the extended Improved Recall [44] (eIR) metric from [42] on intra-class images to measure style variation. It captures the sparseness of style space manifolds where larger eIR stands for more diverse styles. We measure identity consistency by the average cosine similarity between identity embedding pairs. In Fig. 7(a), we compare the similarity and eIR among MorphFace and SOTAs [1, 5, 7, 42], where the FR accuracy is depicted by the color and size of circles. We observe a clear trade-off between consistency and variation. While SOTAs either prioritize identity or style, our approach seeks a balance that improves FR efficacy.

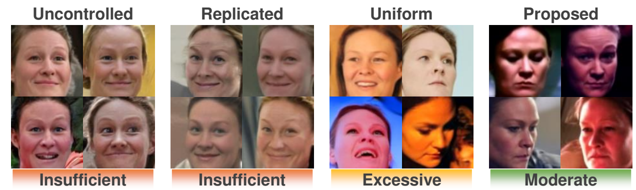

Figure 8. Sample synthetic faces from 4 different style sampling strategies. While others provide insufficiently or excessively varied styles, our proposed approach offers moderate style variation.

	Strategy	eIR	cos-sim	FR Avg.
(a) Style	Uncontrolled	0.475	0.41	91.59
	Replicated	0.178	**0.61**	82.60
	Uniform	**0.720**	0.33	92.41
	Proposed	0.642	0.45	**93.32**
(b) Context	W/o blending	0.608	0.37	93.11
	W/ identity	0.575	**0.51**	92.75
	W/ style	**0.687**	0.35	92.83
	W/ blending	0.642	0.45	**93.32**

Table 2. Analyses of identity consistency, style variation and FR efficacy for style sampling and context blending strategies.

Consistency & Dataset variability. Dataset's overall variability is a combined effort of intra-class variation and subject distinctiveness. By promoting distinctiveness, we can improve variability with less impact on consistency. In Fig. 7(b), we measure variability by dataset-wise eIR. MorphFace manages to create both consistent identities and diverse styles from a dataset perspective. This explains its better performance as both factors are vital for FR efficacy.

4.3. Effect of Style Sampling Strategy

How does the style sampling strategy affect identity consistency, style variation, and FR efficacy? We compare 4 settings: (1) **Uncontrolled**, which we condition the generator $\mathcal{G}$ solely on $\mathbf{c}_{id}$, similar to [7]; (2) **Replicated**, which we reuse the style feature maps of the reference image, instead of sampling novel styles; (3) **Uniform**, which we sample styles uniformly like [42] as $\mathbf{p}' \sim \mathcal{N}(\mu, \mathbf{\Sigma})$; (4) **Proposed**, our subject-aware style sampling discussed in Sec. 3.3.

Visualization. Figure 8 shows sample synthetic images based on the same reference image from 4 settings. We observe: (1) Uncontrolling yields insufficient style variation; (2) Replicating the reference image's style results in even less variation as the style is negatively controlled by the same $\mathbf{c}_{sty}$; (3) Though uniform sampling promotes more diverse styles, its variation is sometimes excessive for the same subject and could affect identity retention; (4) Our subject-aware setting offers moderate style variation.

Quantitative analysis. In Tab. 2(a), we present results on

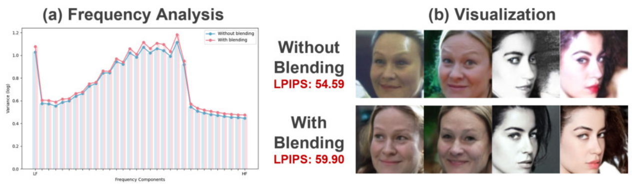

Figure 9. Effects of context blending. It produces images with (a) higher frequency variances and (b) better quality and details.

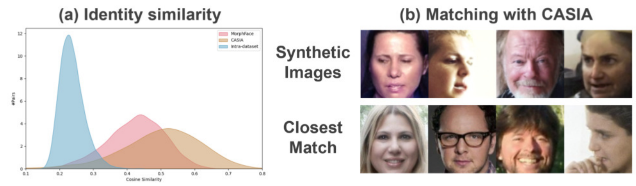

Figure 10. Privacy analyses. (a) Inter-dataset similarity is far lower than the intra-dataset similarities of CASIA and synthetic faces. (b) Synthetic faces are dissimilar to their closest CASIA matches.

eIR, cosine similarity, and average FR accuracy. The low eIR of uncontrolled settings suggests insufficient style variation. We observe a significant trade-off between replicated and uniform settings, which both yield suboptimal performance. The subject-aware setting offers the best FR efficacy due to balanced consistency and variation.

4.4. Effect of Context Blending

How does context blending affect performance? We demonstrate that it mutually benefits intra-class identity consistency and style variation. We compare 4 settings: (1) **Without blending**, where the generator's denoising is not adjusted with CFG; (2) **With identity**, where CFG is applied only to the identity context during $[0, t_0]$ timesteps; (3) **With style**, where CFG only promotes style during $(t_0, T]$; (4) **With blending**, our advocated setting.

Quantitative analysis. Comparisons of eIR, cosine similarity, and FR efficacy are shown in Tab. 2(b). We observe: (1) Context blending improves both eIR and cosine similarity, suggesting our approach's effectiveness; (2) Applying CFG to just one context results in either degraded eIR or cosine similarity, and both settings perform slightly worse than without blending, revealing the inherent trade-off between consistency and variation.

Frequency analysis. We further inspect the frequency components of synthetic images. We convert images into the frequency domain using the fast Fourier transform (FFT) and partition the spectrum into components with different frequencies. Figure 9(a) shows the dataset-average variances of components. Our proposed setting achieves both higher LF and HF variances compared to without blending, suggesting (though not definitively) more informative styles and identities, respectively.

Visualization. We compare synthetic images with and without blending in Fig. 9(b). We find better diversity (by learned perceptual image patch similarity, LPIPS [87]) and more facial details (*e.g.*, wrinkles) in with-blending images.

4.5. Privacy Analysis

The primary purpose of IPFS is to create *unseen* faces that address privacy concerns in real-world datasets. Our generator, $\mathcal{G}$, is trained on CASIA-WebFace [85], raising the natural question of how similar our synthetic faces are to those in CASIA. High similarity could lead to privacy breaches.

Source	CFP-FP	AgeDB	CPLFW	CALFW
Real-world	94.79	92.37	89.73	92.82
Learned	91.61	91.52	85.97	92.32
Proposed	94.11	91.80	88.73	92.73

Table 3. Performance on alternative sources of style attributes.

Identity similarity. CASIA and our synthetic dataset consist of 10.5K and 10K subjects, respectively. We sample one image from each subject and compare the pairwise similarity between subjects from the two datasets. In Fig. 10(a), we compare the inter-dataset similarity with intra-class similarities within each dataset. Both CASIA and our dataset show good intra-class similarities (*i.e.*, 0.52 and 0.45). However, the similarity between them is relatively low (*i.e.*, 0.24). This suggests that our synthetic faces represent virtual subjects, not directly from the training dataset.

Visualization. In Fig. 10(b), we show sample images from our dataset alongside their closest matches from CASIA. The visual dissimilarity further demonstrates the privacy-preserving nature of our synthetic dataset.

4.6. Ablation Studies

Alternative sources of style attributes. We proposed using statistically sampled style attributes. We further compare it with (1) **Real-world** attributes sampled from CASIA, which represent the theoretical upper-bound performance of our style control; (2) **Learned** attributes, where we train a VAE on $\mathbf{P}$ to predict $\mathbf{P}'$. From Tab. 3, we infer that: (1) Real-world attributes achieve better performance (partly due to its nature as $\mathcal{G}$'s training data), suggesting potential for future improvements. We note this setting is aligned with [42] that uses a real style bank; (2) Model-learned attributes perform less well, as they fail to capture the vital statistical details and subject distinctiveness.

5. Conclusion

We have presented MorphFace, a diffusion-based generator that synthesizes faces with both consistent ideneities and diverse styles. Its advancements are three-fold: (1) Achieving fine-grained, parametric control of facial styles; (2) Creating more realistic style varations that promotes FR efficacy; (3) Enhancing expressiveness of identity and style contexts.

References

[1] Gwangbin Bae, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen. Digiface-1m: 1 million digital face images for face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3526–3535, 2023. 3, 4, 7

[2] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 157–164, 2023. 3, 5

[3] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2): 233–254, 2018. 5

[4] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1578–1587, 2022. 2, 4

[5] Fadi Boutros, Marco Huber, Patrick Siebke, Tim Rieber, and Naser Damer. Sface: Privacy-friendly and accurate face recognition using synthetic data. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11. IEEE, 2022. 1, 2, 6, 7

[6] Fadi Boutros, Patrick Siebke, Marcel Klemt, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Pocketnet: Extreme lightweight face recognition network using neural architecture search and multistep knowledge distillation. *IEEE Access*, 10:46823–46833, 2022. 2

[7] Fadi Boutros, Jonas Henry Grebe, Arjan Kuijper, and Naser Damer. Idiff-face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19650–19661, 2023. 1, 2, 3, 4, 5, 6, 7

[8] Fadi Boutros, Marcel Klemt, Meiling Fang, Arjan Kuijper, and Naser Damer. Exfacegan: Exploring identity directions in gan's learned latent space for synthetic identity generation. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2023. 1, 3, 6

[9] Fadi Boutros, Marcel Klemt, Meiling Fang, Arjan Kuijper, and Naser Damer. Unsupervised face recognition using unlabeled synthetic data. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2023. 3

[10] Fadi Boutros, Marco Huber, Anh Thi Luu, Patrick Siebke, and Naser Damer. Sface2: Synthetic-based face recognition with w-space identity-driven sampling. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024. 1, 2, 6

[11] 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. 1, 2

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

[13] Zhuowei Chen, Shancheng Fang, Wei Liu, Qian He, Mengqi Huang, Yongdong Zhang, and Zhendong Mao. Dreamidentity: Improved editability for efficient face-identity preserved image generation. *arXiv preprint arXiv:2307.00300*, 2023. 2

[14] Zhuangzhuang Chen, Ronghao Lu, Jie Chen, Houbing Herbert Song, and Jianqiang Li. Implicit gradient-modulated semantic data augmentation for deep crack recognition. *IEEE Transactions on Intelligent Transportation Systems*, 2024. 2

[15] Ivan DeAndres-Tame, Ruben Tolosana, Pietro Melzi, Ruben Vera-Rodriguez, Minchul Kim, Christian Rathgeb, Xiaoming Liu, Aythami Morales, Julian Fierrez, Javier Ortega-Garcia, et al. Frcsyn challenge at cvpr 2024: Face recognition challenge in the era of synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3173–3183, 2024. 2

[16] Ivan DeAndres-Tame, Ruben Tolosana, Pietro Melzi, Ruben Vera-Rodriguez, Minchul Kim, Christian Rathgeb, Xiaoming Liu, Luis F. Gomez, Aythami Morales, Julian Fierrez, Javier Ortega-Garcia, Zhizhou Zhong, Yuge Huang, Yuxi Mi, Shouhong Ding, Shuigeng Zhou, Shuai He, Lingzhi Fu, Heng Cong, Rongyu Zhang, Zhihong Xiao, Evgeny Smirnov, Anton Pimenov, Aleksei Grigorev, Denis Timoshenko, Kaleb Mesfin Asfaw, Cheng Yaw Low, Hao Liu, Chuyi Wang, Qing Zuo, Zhixiang He, Hatef Otroshi Shahreza, Anjith George, Alexander Unnervik, Parsa Rahimi, Sébastien Marcel, Pedro C. Neto, Marco Huber, Jan Niklas Kolf, Naser Damer, Fadi Boutros, Jaime S. Cardoso, Ana F. Sequeira, Andrea Atzori, Gianni Fenu, Mirko Marras, Vitomir Štruc, Jiang Yu, Zhangjie Li, Jichun Li, Weisong Zhao, Zhen Lei, Xiangyu Zhu, Xiao-Yu Zhang, Bernardo Biesseck, Pedro Vidal, Luiz Coelho, Roger Granada, and David Menotti. Second frcsyn-ongoing: Winning solutions and post-challenge analysis to improve face recognition with synthetic data. *Information Fusion*, 120: 103099, 2025. 2

[17] Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, and Stefanos Zafeiriou. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7093–7102, 2018. 2

[18] 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. 2

[19] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5154–5163, 2020. 2, 3

[20] Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12736–12746, 2023. 2, 4

[21] 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. 2, 3, 4

[22] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[23] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4):1–13, 2023. 2

[24] Baris Gecer, Binod Bhattarai, Josef Kittler, and Tae-Kyun Kim. Semi-supervised adversarial learning to generate photorealistic face images of new identities from 3d morphable model. In *Proceedings of the European conference on computer vision (ECCV)*, pages 217–234, 2018. 2

[25] Zhenglin Geng, Chen Cao, and Sergey Tulyakov. 3d guided fine-grained face manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9821–9830, 2019. 2

[26] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016. 1, 2, 5

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

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

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[30] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[31] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[32] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 2

[33] Xianliang Huang, Yining Lang, Ying Guo, Yuan He, Hui Xue, Li Zhao, and Shuigeng Zhou. Dr-net: A multi-view face synthesis network driven by dual representation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1751–1756. IEEE, 2023. 2

[34] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020. 2

[35] 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. 2

[36] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020. 3

[37] 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

[38] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863, 2021. 2

[39] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4873–4882, 2016. 2

[40] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM transactions on graphics (TOG)*, 37(4): 1–14, 2018. 2

[41] Minchul Kim, Anil K. Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18750–18759, 2022. 2

[42] Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. Dcface: Synthetic face generation with dual condition diffusion model. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 12715–12725, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[43] Jan Niklas Kolf, Tim Rieber, Jurek Elliesen, Fadi Boutros, Arjan Kuijper, and Naser Damer. Identity-driven three-player generative adversarial network for synthetic-based face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 806–816, 2023. 1, 3, 6

[44] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019. 7

[45] Gary B. Huang Erik Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, 2014. 6

[46] Jianqiang Li, Zhuangzhuang Chen, Jie Chen, and Qiuzhen Lin. Diversity-sensitive generative adversarial network for terrain mapping under limited human intervention. *IEEE Transactions on Cybernetics*, 51(12):6029–6040, 2021. 2

[47] Shen Li, Jianqing Xu, Jiaying Wu, Miao Xiong, Ailin Deng, Jiazhen Ji, Yuge Huang, Wenjie Feng, Shouhong Ding, and Bryan Hooi. Id3: Identity-preserving-yet-diversified diffusion models for synthetic face recognition. *arXiv preprint arXiv:2409.17576*, 2024. 2, 3, 4, 5, 6

[48] 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. 2, 3, 5

[49] Xiaoyu Li, Qi Zhang, Di Kang, Weihao Cheng, Yiming Gao, Jingbo Zhang, Zhihao Liang, Jing Liao, Yan-Pei Cao, and Ying Shan. Advances in 3d generation: A survey. *arXiv preprint arXiv:2401.17807*, 2024. 4

[50] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8640–8650, 2024. 2

[51] Safa C Medin, Bernhard Egger, Anoop Cherian, Ye Wang, Joshua B Tenenbaum, Xiaoming Liu, and Tim K Marks. Most-gan: 3d morphable stylegan for disentangled face image manipulation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1962–1971, 2022. 2

[52] Pietro Melzi, Ruben Tolosana, Ruben Vera-Rodriguez, Minchul Kim, Christian Rathgeb, Xiaoming Liu, Ivan DeAndres-Tame, Aythami Morales, Julian Fierrez, Javier Ortega-Garcia, et al. Frcsyn challenge at wacv 2024: Face recognition challenge in the era of synthetic data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 892–901, 2024. 2

[53] Yuxi Mi, Yuge Huang, Jiazhen Ji, Hongquan Liu, Xingkun Xu, Shouhong Ding, and Shuigeng Zhou. Duetface: Collaborative privacy-preserving face recognition via channel splitting in the frequency domain. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6755–6764, 2022. 5

[54] Yuxi Mi, Zhizhou Zhong, Yuge Huang, Jiazhen Ji, Jianqing Xu, Jun Wang, Shaoming Wang, Shouhong Ding, and Shuigeng Zhou. Privacy-preserving face recognition using trainable feature subtraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 297–307, 2024. 2

[55] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017. 2, 6

[56] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019. 2

[57] Fu-Zhao Ou, Xingyu Chen, Ruixin Zhang, Yuge Huang, Shaoxin Li, Jilin Li, Yong Li, Liujuan Cao, and Yuan-Gen Wang. Sdd-fiqa: Unsupervised face image quality assessment with similarity distribution distance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7670–7679, 2021. 4

[58] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model of human faces. *arXiv preprint arXiv:2403.11641*, 2024. 1, 3, 6

[59] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009. 5

[60] Xu Peng, Junwei Zhu, Boyuan Jiang, Ying Tai, Donghao Luo, Jiangning Zhang, Wei Lin, Taisong Jin, Chengjie Wang, and Rongrong Ji. Portraitbooth: A versatile portrait model for fast identity-preserved personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27080–27090, 2024. 2

[61] Jingtan Piao, Chen Qian, and Hongsheng Li. Semi-supervised monocular 3d face reconstruction with end-to-end shape-preserved domain transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9398–9407, 2019. 2

[62] Yurui Qian, Qi Cai, Yingwei Pan, Yehao Li, Ting Yao, Qibin Sun, and Tao Mei. Boosting diffusion models with moving average sampling in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8911–8920, 2024. 5

[63] Haibo Qiu, Baosheng Yu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao. Synface: Face recognition with synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10880–10890, 2021. 1, 2, 6

[64] 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. 2

[65] 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. 2, 3

[66] 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. 3

[67] David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models. *arXiv preprint arXiv:2402.09470*, 2024. 5

[68] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2

[69] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016. 2, 6

[70] Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, and Xiaoou Tang. Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 821–830, 2018. 2

[71] Yujun Shen, Bolei Zhou, Ping Luo, and Xiaoou Tang. Facefeat-gan: a two-stage approach for identity-preserving face synthesis. *arXiv preprint arXiv:1812.01288*, 2018. 2

[72] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020. 2

[73] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[74] Zhonglin Sun, Siyang Song, Ioannis Patras, and Georgios Tzimiropoulos. Cemiface: Center-based semi-hard synthetic face generation for face recognition. *arXiv preprint arXiv:2409.18876*, 2024. 3, 4, 5, 6

[75] Luan Tran, Xi Yin, and Xiaoming Liu. Representation learning by rotating your faces. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):3007–3021, 2018. 2

[76] Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. Face0: Instantaneously conditioning a text-to-image model on a face. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10, 2023. 2

[77] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018. 2

[78] Qinghe Wang, Xu Jia, Xiaomin Li, Taiqing Li, Liqian Ma, Yunzhi Zhuge, and Huchuan Lu. Stableidentity: Inserting anybody into anywhere at first sight. *arXiv preprint arXiv:2401.15975*, 2024. 2

[79] Yinggui Wang, Jian Liu, Man Luo, Le Yang, and Li Wang. Privacy-preserving face recognition in the frequency domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2558–2566, 2022. 5

[80] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, pages 1–20, 2024. 2

[81] Zunnan Xu, Yachao Zhang, Sicheng Yang, Ronghui Li, and Xiu Li. Chain of generation: Multi-modal gesture synthesis via cascaded conditional control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6387–6395, 2024. 2

[82] Yuxuan Yan, Chi Zhang, Rui Wang, Yichao Zhou, Gege Zhang, Pei Cheng, Gang Yu, and Bin Fu. Facestudio: Put your face everywhere in seconds. *arXiv preprint arXiv:2312.02663*, 2023. 2

[83] Zhiyuan Yan, Jiangming Wang, Zhendong Wang, Peng Jin, Ke-Yue Zhang, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Effort: Efficient orthogonal modeling for generalizable ai-generated image detection. *arXiv preprint arXiv:2411.15633*, 2024. 2

[84] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. Df40: Toward next-generation deepfake detection. *arXiv preprint arXiv:2406.13495*, 2024. 2

[85] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 2, 5, 6, 8

[86] Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. Inserting anybody in diffusion models via celeb basis. *arXiv preprint arXiv:2306.00926*, 2023. 2

[87] 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. 8

[88] Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5(7):5, 2018. 2, 6

[89] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017. 2, 6

[90] Zhizhou Zhong, Yuxi Mi, Yuge Huang, Jianqing Xu, Guodong Mu, Shouhong Ding, Jingyun Zhang, Rizen Guo, Yunsheng Wu, and Shuigeng Zhou. Slerpface: face template protection via spherical linear interpolation. *arXiv preprint arXiv:2407.03043*, 2024. 2

[91] Yufan Zhou, Ruiyi Zhang, Tong Sun, and Jinhui Xu. Enhancing detail preservation for customized text-to-image generation: A regularization-free approach. *arXiv preprint arXiv:2305.13579*, 2023. 2

[92] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021. 1, 2, 5