2026 International Conference on Automatic Face and Gesture Recognition (FG)

IDSync: Improving diffusion models through identity classification

Jernej Sabadin[1], Darian Tomašević[1], Blaž Meden[1], Peter Peer[1], and Vitomir Štruc[2]
[1] University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia
[2] University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia

***Abstract*— Effective training of face recognition models requires large-scale datasets of facial identities, yet collecting suitable data is time-consuming and raises privacy concerns. Existing deep generative models offer a promising alternative through the synthesis of high-quality images but often fail to fully preserve identity information. In this work, we propose IDSync, a novel generative diffusion-based framework designed to produce synthetic face images with more consistent identities that are better suited for training recognition models. To this end, IDSync employs a denoising network in the latent space of a frozen variational autoencoder, with identity guidance introduced via a text encoder that interprets identity embeddings from a pretrained recognition model. During training, the framework leverages a pretrained auxiliary identity classifier to define an additional cross-entropy loss, which is backpropagated to improve identity consistency. We evaluate the generated images using inter- and intra-class cosine similarity of identity features along with a variety of statistical measures between synthetic and real distributions focused on fidelity and diversity. To assess utility, we train face recognition models on the synthetic images and measure accuracy on standard verification benchmarks. Experimental results show that recognition models trained on IDSync-generated data achieve higher verification accuracies on real-world benchmarks than models trained on synthetic data produced by competing generative models. The IDSync source code is publicly available at `https://github.com/JSabadin/IDSync`.**

I. Introduction

Face recognition (FR) has become a cornerstone of modern biometrics, with applications ranging from secure authentication and surveillance to human–computer interaction and augmented reality [22], [16]. Contemporary FR systems, based on deep neural architectures, routinely achieve verification accuracies exceeding 99% on benchmarks like Labeled Faces in the Wild (LFW) [51], [10]. However, their performance critically depends on large-scale datasets of real face images, which are prohibitively time-consuming to gather when constrained by privacy regulations such as the GDPR, i.e., when explicit consent of subjects is required [23], [32]. Synthetic data generation thus emerges as a compelling alternative, enabling rapid dataset creation or augmentation without breaching individual privacy [7].

Recent advances in generative modeling, most notably diffusion models, have demonstrated the ability to produce high-fidelity images guided by text prompts across diverse domains [43], [11]. In the context of face synthesis, generative models have been leveraged to produce images of both real and synthetic identities, enabling the construction of recognition datasets with multiple samples per identity [6], [5], [36], [53]. State-of-the-art diffusion-based methods for this task typically rely either on per-identity fine-tuning [45], [48] or generalizable identity conditioning [37], [36]. While the former yields better identity consistency, it lacks scalability. In contrast, generalizable identity conditioning enables the generation of novel identities by training on large-scale datasets instead. Arc2Face [36], for instance, conditions a diffusion model with identity embeddings from a pretrained face recognition network, resulting in visually convincing faces whose identity can be interpolated in the embedding space. Despite these advances, current solutions still exhibit inconsistency in preserving identity features across multiple samples of the same subject [36]. To remedy this, several approaches have introduced custom training objectives that utilize the cosine similarity between identity embeddings of pretrained face recognition networks [37], [53]. However, cosine-based identity supervision provides only pairwise alignment, whereas cross-entropy induces proxy-based multi-class competition with implicit negatives and can outperform pairwise metric losses [4]. Existing works also demonstrate that cosine objectives can be improved with explicit negatives to form triplet objectives and prevent overfitting on unintentional cues present within identity embeddings [48].

To address these limitations, we propose IDSync, a novel identity-conditioned diffusion-based framework that incorporates an auxiliary classification loss during training to enforce identity consistency. By coupling the denoising network with a pretrained classifier that evaluates whether each generated

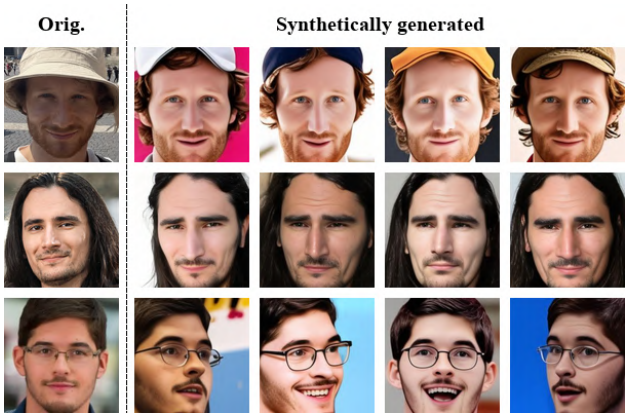

Fig. 1. **Samples generated with our IDSync framework**. The model is conditioned on identity embeddings from images outside the training set, demonstrating its ability to create realistic face images of unseen identities.

Supported in parts by the Slovenian Research and Innovation Agency (ARIS) through Research Programmes P2-0250 (B) "Metrology and Biometric Systems" and P2–0214 (A) "Computer Vision", the ARIS Project J2-50065 "DeepFake DAD", the ARIS Young Researcher Programme, and the European Union's Horizon Europe research and innovation programme through the OnMoveID project under grant agreement No. 101225635.

979-8-3315-7231-0/26/$31.00 © 2026 European Union

image is recognized as the intended identity, IDSync ensures that synthetic faces not only appear realistic but also maintain stable identity features across variations, as seen in Figure 1. The denoising network is accompanied by a variational autoencoder, which provides a low-dimensionality latent space for efficient data generation, and is conditioned on an encoded combination of text prompts and identity embeddings of the ArcFace-based recognition model [10]. In our experiments, we perform separate training of a latent diffusion model on two real-world datasets, including CASIA-WebFace [57] and a 2M-subset of the upscaled WebFace42M dataset [36], and generate datasets of $10,000$ synthetic identities, with 50 samples per identity. We compare the produced synthetic data with data generated by the state-of-the-art Arc2Face [36] framework in terms of quality, fidelity, and diversity, measured with statistical measures on distributions of features extracted with DINOv2 [35]. We also assess the utility of the produced data, by using purely synthetic data to train AdaFace-based [26] recognition models and evaluating their performance on standard verification benchmarks.

In this paper, we make the following contributions:

- We propose IDSync, a novel identity conditioned diffusion framework for generating diverse high-fidelity face images of either existing or synthetic identities.
- We introduce an auxiliary identity classification training objective for ensuring consistent and realistic identity preservation in generated samples.
- We demonstrate that training recognition models with purely synthetic data produced by IDSync yields superior accuracy on standard verification benchmarks compared to SynFace [39], DigiFace [2], IDiff-Face [5], and even Arc2Face [36] under identical settings.

II. Related Work

In this section we review prior work on generative image models and, more specifically, synthetic face image generation. We first summarize the evolution of generative models, from GANs to diffusion approaches, and then discuss methods that condition these models on facial identity.

A. Deep generative models

The introduction of Generative Adversarial Networks (GANs) [15] marks a pivotal point in the rapid development of image generation models. GANs were the first models to enable the synthesis of convincing images, by framing the task as an adversarial game between the generator and the discriminator network. Subsequent variants were designed to address crucial issues of GANs, including training instability [41] and mode collapse [1], while enabling higher-resolution synthesis [24] and style-based control of the generation process [25].

More recently, however, Denoising Diffusion Probabilistic Models (DDPMs) have reformulated image synthesis as an iterative denoising process [19], yielding better image quality than existing GAN-based approaches [11]. To predict the noise that should be removed at each denoising step DDPMs utilize a convolutional encoder-decoder architecture, which is trained to gradually synthesize images from randomly sampled noise. Latent Diffusion Models (LDMs) [43] further improved efficiency by operating in a compressed latent space of a pretrained variational autoencoder. In addition to facilitating higher-resolution synthesis, LDMs enabled prompt-based conditioning by incorporating a pretrained text encoder. These advancements along with open-source variants of LDMs, e.g. Stable Diffusion (SD) [47], have resulted in widespread use and adoption of diffusion-based image generation models. To further improve synthesis capabilities, recent developments have scaled the denoising backbone [38] and introduced bidirectional information flow between image and textual modalities [14]. Guidance of the generation process has also been extended with additional conditions including segmentation masks and depth maps [58] as well as image features extracted with pretrained models [56].

B. Synthetic face image generation

In the context of face recognition, synthetic data generated by deep generative models provides a promising avenue for addressing critical limitations of biometric datasets, including privacy and copyright issues of web-scraped images, challenges in collecting data with the consent of subjects, and demographic imbalances of available datasets [7]. To generate suitable data, face synthesis methods typically condition generative models on identity labels or identity features extracted with pretrained recognition models. Early GAN-based approaches [12], [49], [31], [42], [39] leverage explicit identity supervision to steer generation, laying the groundwork for later diffusion-based techniques.

Diffusion-based face synthesis methods generally fall into two categories, including (i) *per-identity fine-tuning* and (ii) *generalizable conditioning*. Per-identity fine-tuning approaches, e.g. DreamBooth [45], yield highly faithful reproductions for a given person, by fine-tuning pretrained diffusion model on a small set of images per subject. Additional identity-based training objectives have also been shown to further improve identity consistency [48]. However, these approaches remain unsuitable for the generation of large-scale datasets, as they require separate training for each identity and do not allow the creation of previously unseen faces. Differently, generalizable conditioning approaches, including PortraitBooth [37], FaceStudio [54] and PhotoVerse [8], train generative models on large-scale datasets by combining CLIP-based visual–textual features [40] with extracted face representations. This in turn enables the generation of diverse and high-quality images of new identities. However, a high level of consistency across faces of the same identity remains a challenge, as models often fail to consistently reproduce the same facial features due to combined conditioning. Furthermore, the CLIP model remains frozen in these methods, which limits the conditioning capability, as it was not primarily trained to work with facial features. Notable methods that generate faces from pre-computed embeddings and achieve state-of-the-art results for large-scale recognition dataset synthesis include Vec2Face [53] and Arc2Face [36]. Vec2Face employs a feature-masked

autoencoder (fMAE) that reconstructs masked identity features while back-propagating a cosine similarity loss between the embeddings of real and generated faces to boost intra-class consistency [53]. Arc2Face [36], on the other hand, injects identity vectors from an ArcFace-based recognition model [10] as token embeddings into CLIP's text encoder and conditions a latent diffusion model on those embeddings to produce diverse, high-quality face images. However, identity consistency across samples remains imperfect.

Existing face generators have made great strides in quality and diversity, but none directly enforce a stable identity loss during training. For example, Arc2Face [36] conditions on pretrained embeddings without any auxiliary loss to improve cross-sample consistency, while Vec2Face [53] and ID-Booth [48] back-propagates a cosine-similarity loss through a frozen FR network. However, this embedding-level objective can be noisy and difficult to balance. To address this gap, we introduce the IDSync framework, which embeds an identity classifier into the diffusion training loop and employs a cross-entropy loss to enforce consistent identity preservation across generated samples. By leveraging a pretrained classification model rather than an embedding-similarity metric, our Cross-Entropy-based scheme yields more stable gradients and stronger identity fidelity.

III. Methodology

In this section we present the novel IDSync framework that enables the generation of identity consistent face images with high diversity and fidelity. We start with an overview of the framework, followed by detailed descriptions of each component and the training procedure of the framework.

A. Overview of IDSync

Inspired by recent state-of-the-art generative models, such as Arc2Face [36], we present a novel diffusion-based framework, called IDSync, for identity conditioned face image generation. As depicted in Figure 2, IDSync comprises four main components: (i) a denoising diffusion probabilistic model for synthesizing face images from input noise, (ii) a Variational Autoencoder (VAE) that defines the latent space in which diffusion operates, (iii) a text encoder that maps prompts and identity features used to guide the denoising network, and (iv) a frozen identity classifier that provides an additional cross-entropy identity loss during training to align generated and target identities. After training, only components (i)–(iii) are used to produce diverse synthetic faces of either existing or novel identities. A detailed description of each component and the training procedure is given below.

B. Denoising diffusion probabilistic model

The core component of the proposed IDSync framework is the Denoising Diffusion Probabilistic Model (DDPM), whose generative capabilities stem from learning the denoising procedure. Denoising entails the reverse of the noising process, which degrades input images $\mathbf{x}_0$ with noise across a Markov chain of length T, as follows:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\big(\mathbf{x}_t; \sqrt{1-\beta_t}\,\mathbf{x}_{t-1},\, \beta_t \mathbf{I}\big), \tag{1}$$

where the variance schedule $\{\beta_t\}_{t=1}^{T}$ gradually corrupts the image with Gaussian noise. Common schedules include a linear increase or a cosine schedule, typically with $T = 1000$ timesteps. The reverse (denoising) distribution is modeled as:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\big(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t),\, \tilde{\beta}_t \mathbf{I}\big), \tag{2}$$

with $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\,\beta_t,\ \alpha_t = 1-\beta_t,\ \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. Training then minimizes the mean-squared error:

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{\mathbf{x}_0,\epsilon,t}\big\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\big\|^2, \tag{3}$$

so that at inference the model can predict the added noise at any timestep and iteratively reconstruct $\mathbf{x}_0$ from pure noise.

C. Variational Autoencoder

Instead of operating in pixel space, IDSync performs diffusion in the latent space of a pretrained Variational Autoencoder (VAE), which significantly reduces computational cost and memory usage. The VAE comprises an encoder–decoder network pair (E, D) that maps images to a compact latent representation and back. Given an input image $\mathbf{x}$, the encoder network E produces a latent code $\mathbf{z}_0$:

$$\mathbf{z}_0 = E(\mathbf{x}) \in \mathbb{R}^{c\times h\times w}, \tag{4}$$

where c, h, and w denote latent channels and dimensions. The diffusion process from Equations 1 and 2 is applied to $\mathbf{z}_0$ instead of $\mathbf{x}_0$, yielding a denoised latent $\hat{\mathbf{z}}_0$, which the decoder D reconstructs back to $\hat{\mathbf{x}}_0$ in the image space:

$$\hat{\mathbf{x}}_0 = D\left(\hat{\mathbf{z}}_0\right) \in \mathbb{R}^{c\times H\times W}. \tag{5}$$

The VAE is pretrained by minimizing the standard objective:

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[-\log p(\mathbf{x} \mid \mathbf{z})\right] + \mathrm{KL}\left(q(\mathbf{z} \mid \mathbf{x}) \parallel p(\mathbf{z})\right), \tag{6}$$

where $q(\mathbf{z} \mid \mathbf{x})$ and $p(\mathbf{x} \mid \mathbf{z})$ are the encoder and decoder distributions, respectively, and $p(\mathbf{z})$ is a Gaussian prior.

D. Text encoder network

CLIP (Contrastive Language–Image Pretraining) is a dual-encoder model, comprising a transformer-based text encoder and an image encoder, which is trained to align text and image embeddings in a shared latent space. Given a textual prompt y, the text encoder $\tau(y)$ maps the tokenized sequence into a fixed-length embedding vector.

To enable identity-conditioned image synthesis, IDSync injects identity information into the CLIP text encoder τ with a fixed input prompt (e.g., `"photo of a id person"`). After BPE tokenization of the prompt, tokens $w_1, \ldots, w_L$ ($w_i \in \{1, \ldots, V\}$) are mapped via a pretrained vocabulary embedding matrix $\mathbf{W}_{emb} \in \mathbb{R}^{V\times S}$ to token embeddings $\mathbf{e}_i = \mathbf{W}_{emb}[w_i] \in \mathbb{R}^S$. Next, we compute an identity vector $\mathbf{e}_{\text{id}} \in \mathbb{R}^S$ from the input image with a pretrained ArcFace recognition model [10] (feature ℓ_2-normalized and zero-padded to dimension S) and replace the embedding at the placeholder `"id"` position with $\mathbf{e}_{\text{id}}$. CLIP then adds the learned positional embedding $\mathbf{p}_i \in \mathbb{R}^S$ as usual, i.e., the replaced position is encoded as $\mathbf{e}_{\text{id}} + \mathbf{p}_i$.

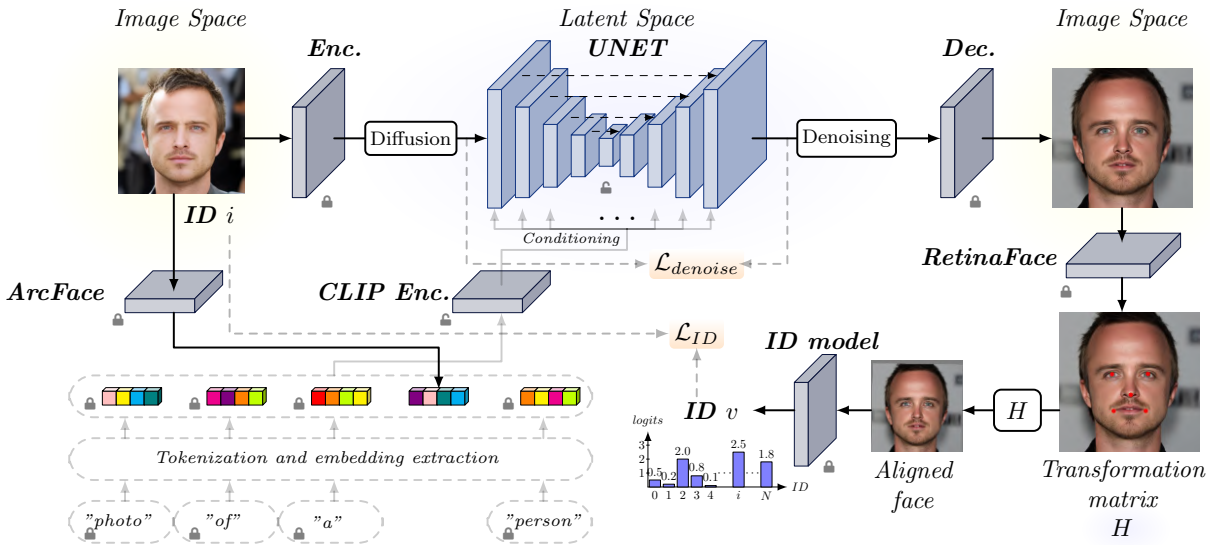

Fig. 2. **Overview of the custom-training introduced with the IDSync framework.** The diffusion model is conditioned by embedding ArcFace [10] identity features into the text-embedding space of CLIP [40] and injecting them into the denoising U-Net [44] via cross-attention blocks. Training incorporates an additional classification loss to encourage more identity-consistent outputs. The loss is computed using a frozen IR-SE-50 [13], [20] identity classification model trained beforehand on the same dataset as the diffusion model. Before being fed to the classification model, images are aligned using landmarks from the RetinaFace [9] detector. Modules marked with a lock icon are frozen during IDSync training, while those with an open lock are trainable.

During training, we fine-tune the CLIP text transformer parameters while freezing only the vocabulary token embedding matrix $\mathbf{W}_{emb}$; $\mathbf{e}_{\text{id}}$ is treated as an external input. The resulting embedding $\tau(y, e_{id})$ is thus semantically grounded while identity-aware, and continues to be updated during training in our IDSync framework. In the context of latent diffusion, this embedding $\tau(y, e_{id})$ is used to guide the generative process by injecting it into the denoising U-Net via cross-attention layers at each timestep. Formally, the noise prediction function becomes:

$$\hat{\epsilon} = \epsilon_\theta\big(z_t,\, t,\, \tau(y, e_{id})\big), \tag{7}$$

where z_t is the noisy latent at timestep t, and $\tau(y, e_{id})$ provides the conditioning signal. The cross-attention mechanism fuses this embedding into the internal activations of the denoising U-Net [44], enabling fine-grained control over the synthesis output based on the desired prompt and identity.

E. Training via identity classification

As shown in Figure 2, IDSync aims to generate a synthetic face image $\hat{x} \in \mathbb{R}^{c\times H\times W}$ that matches a given identity label $i \in \{1, \ldots, N\}$ while maintaining high visual fidelity. To accomplish this, we incorporate an auxiliary identity-classification branch into the training of the reverse diffusion process. The branch includes (i) the face detection and alignment with a pretrained RetinaFace [9] landmark detector, and (ii) classification of the identity in the image with a frozen classification, pretrained separately on the training dataset. The auxiliary identity-classification loss is defined as:

$$\mathcal{L}_{\text{ID}} = -\log p_{\text{cls}}(i \mid \hat{\mathbf{x}}'), \tag{8}$$

where, $\hat{\mathbf{x}}'$ represents the input denoised image, aligned via a transformation matrix H determined by the landmark detector [9]; if no face is detected, the unaligned image $\hat{\mathbf{x}}$ is forwarded instead. Thus, the proposed loss guides IDSync to produce that the external identity classification model, correctly recognizes as identity i. In turn, training the entire model is guided by minimizing the combined loss:

$$\mathcal{L} = \mathcal{L}_{\text{denoise}} + \lambda\, \mathcal{L}_{\text{ID}}, \tag{9}$$

where the trade-off parameter λ balances reconstruction fidelity against identity consistency. By jointly optimizing both terms, IDSync learns to produce new high-fidelity synthetic images of the desired input identity.

IV. Experiments and Results

To evaluate the proposed IDSync framework, we conduct a comprehensive series of experiments. We first outline the datasets and evaluation protocols, then perform a sensitivity analysis of the identity classification loss, assess the scalability of the framework, and finally present the ablation studies.

A. Datasets

For training IDSync, we use two real-world datasets. First, the **CASIA-Webface** [57] dataset, which contains $494,414$ images spanning $10,575$ unique identities. It is large and diverse enough to test various training configurations and analyze their impact on the quality of generated faces, yet small enough to keep training times reasonable. All images are resized to 224×224 pixels using GFPGAN 1.4 [52]. Second, for large-scale experiments, we use a random 2M-image subset of the upscaled **WebFace42M** dataset [36], comprising $65,209$ identities, with at least 10 associated images per identity, hereafter denoted as WebFace2M for brevity. Additionally, for qualitative evaluation on unseen identities, we use the **FFHQ** dataset [25]; it is employed only for visualization and does not overlap with our training identities. To evaluate performance, we generate synthetic datasets with IDSync, train face recognition models on them, and report results on the following five standard benchmarks. **LFW** [21] contains 13,233 images of 5,749 identities with varied poses, expressions, and illumination. **CPLFW** [59]

has 11,652 images of 3,930 identities with challenging frontal–profile pairs. **CALFW** [28] includes 12,174 images of 4,025 identities, emphasizing age gaps (average ~ 8.1 years). **CFP-FP** [46] comprises 7,000 images from 500 identities, with 10 frontal and 4 profile images per identity. **AgeDB** [33] provides 16,488 images of 568 identities spanning up to 30 years of age variation.

B. Implementation details

Training with IDSync. In this paper, we demonstrate the power of the IDSync framework by using it to fine-tune the state-of-the-art latent diffusion model Stable Diffusion 2.1 (SD-2.1) [43] for identity-conditioned image generation. The pretrained model is designed to generate diverse high-fidelity 512×512 images based on an input text prompt. To this end, the model performs the denoising process defined in Section III-B for 1000 timesteps following the discrete denoising scheduler with $\beta_{start} = 8.5 \times 10^{-4}$ and $\beta_{end} = 8.5 \times 0.012$ [19]. To fine-tune the model, we employ either CASIA-WebFace or WebFace2M as the training set and consider two configurations: (i) the Arc2Face baseline [36] trained solely with the denoising loss $\mathcal{L}_{\text{denoise}}$, and (ii) the baseline extended with our novel identity classification loss $\mathcal{L}_{ID}$ weighted by a range of trade-off parameters λ. For defining $\mathcal{L}_{ID}$ during training, the denoised face images are first aligned with the transformation matrix H produced by the RetinaFace [9] landmark detector, pretrained on the WIDER FACE dataset [55]. If no face is detected, the entire image is used instead. $\mathcal{L}_{ID}$ is then computed using a frozen IResNet-50 model [13] with Squeeze-and-Excitation blocks [20] and a classification head (i.e., IR-SE-50), following the procedure defined in Section III-E. Training is performed with the prompt "`photo of a id person`", where the token embedding of `id` is replaced with an identity embedding from an ArcFace-based [10] recognition model. Before substitution, extraced 512-dimensional identity embeddings are zero-padded to match the 1024-dimensional token embeddings. The model is trained on images of a resolution 224×224, normalized with $\mu = [0.5, 0.5, 0.5]$ and $\sigma = [0.5, 0.5, 0.5]$ in batches of 8 with 64 gradient accumulation steps for 8 epochs. Updates follow the AdamW optimizer [29] with a learning rate of 1.6×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, and a weight decay of 10^{-2}.

Training the auxiliary classification model. During training of IDSync we utilize the auxiliary IR-SE-50 [13], [20] classification model to form the cross-entropy identity loss. To this end, we train the classification model beforehand, either on 0.5M images from CASIA-WebFace [57] or 2M images from WebFace42M [36] depending on the experiment. This training is performed on 112×112 images, normalized with $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$, in batches of 256 with a gradient accumulation step of 4. For training we rely on the Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and $L_2 = (5 \times 10^{-5})$. The initial learning rate of 0.04 follows the cosine annealing scheduler with $T_{\max} = 100$, $\alpha_0 = 0.04$, and $\alpha_{\min} = 3{\times}10^{-5}$ for 100 epochs. Data augmentation is also applied in the form of JPEG compression, Gaussian blur, color jitter, grayscale conversion, perspective transform, as well as image rotation and flipping. Validation is stratified per identity: we hold out one image per identity (CASIA-WebFace: 10,575 images; WebFace2M: 65,209 images) and use the remainder for training. We then report top-1 accuracy on this hold-out split.

Synthetic data generation. To assess the effectiveness of IDSync for generating training data for FR, we create a 0.5M-image dataset with 10,000 synthetic identities and 50 images per identity. We extract 512-D ArcFace embeddings for all real training images and ℓ_2-normalize them, then fit PCA on centered embeddings $\mathbf{e} - \boldsymbol{\mu}$, retaining the top $k{=}400$ components which explain 99% of the variance. Here $\boldsymbol{\mu}$ is the global mean of the (unit-normalized) training embeddings, and $\mathbf{V} \in \mathbb{R}^{k \times 512}$ and $\boldsymbol{\lambda} \in \mathbb{R}^k$ denote the PCA directions and variances. For each synthetic identity we sample $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$ and generate per-image intra-class variation by adding isotropic noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{512})$ with $\sigma^2 = 0.01$, followed by re-normalization: $\mathbf{e}_{\text{syn}} = \frac{\boldsymbol{\mu} + \mathbf{V}^\top(\mathbf{z} \odot \sqrt{\boldsymbol{\lambda}}) + \boldsymbol{\epsilon}}{\left\| \boldsymbol{\mu} + \mathbf{V}^\top(\mathbf{z} \odot \sqrt{\boldsymbol{\lambda}}) + \boldsymbol{\epsilon} \right\|_2}$, where $\odot$ denotes element-wise multiplication. To accelerate image sampling, we replace the DDPM sampler with the deterministic DPM-Solver [30], utilizing 25 denoising steps with a guidance scale of 3.

Hardware and time requirements. Experiments are conducted using an NVIDIA RTX 3090 GPU with 24GB VRAM. Training the IDSync model on this hardware requires 78 hours for the CASIA-WebFace dataset, including 6 hours for the auxiliary classifier. On WebFace2M, the classifier is trained in 24 hours, and IDSync in 312 hours. Generation of each synthetic 0.5M dataset requires 24 hours.

C. Synthetic Data Evaluation Protocol

Distribution alignment. As part of our experiments, we first select a subset of $1,000$ identities to compare the distributions of synthetic and real data. To assess visual quality, we measure the Fréchet Distance [18] between real and synthetic distributions using DINOv2 features [35], along with the Kernel distance measured in terms of Maximum Mean Discrepancy (MMD) [3]. Additionally, image fidelity and diversity are measured separately with Density and Coverage [34] on the extracted features. Pose diversity is also investigated by estimating head orientation angles using the 6DRepNet model [17]. Lastly, to investigate identity fidelity, we extract embeddings from images with a pretrained ArcFace [10] recognition model and compute inter- and intra-class cosine similarities. To further assess separability and structure in the identity space, we visualize these embeddings with the t-SNE [50] method and report the Jensen–Shannon divergence between the distributions.

Downstream face recognition performance. To evaluate the practical utility of the synthetic data, we trained a face recognition model using only the generated samples and tested its verification accuracy on standard benchmarks. This served as our primary indicator of real-world applicability. For consistency with prior work [39], [2], [5], [36], we create a dataset of 0.5M synthetic images, and train the AdaFace

TABLE I

COMPARISON OF FRÉCHET DISTANCE [18] (↓), KERNEL MMD [3] (↓), DENSITY [34] (↑), AND COVERAGE [34] (↑) BETWEEN EACH SYNTHETIC SET AND THE REAL DISTRIBUTION OVER 1000 IDENTITIES (50 IMAGES EACH) FROM CASIA-WEBFACE, COMPUTED ON DINOV2 FEATURES. **BLUE** HIGHLIGHTS THE BEST VALUE AND **YELLOW** THE SECOND-BEST.

Metric	Casia-Web0.5M	Arc2Face	IDSync (0.0001)	IDSync (0.0005)	IDSync (0.001)	IDSync (0.005)	IDSync (0.01)
Fréchet ↓	45.91	746.16	689.59	750.96	700.72	758.87	707.81
MMD ↓	0.079	2.688	2.446	2.728	2.560	2.676	2.518
Density ↑	0.9051	0.1866	0.1659	0.1790	0.1759	0.2069	0.1758
Coverage ↑	0.9272	0.1764	0.1832	0.1747	0.1800	0.1843	0.1834

model (IR-SE-50 backbone) on downscaled 112×112 images. Training is performed for 40 epochs in batches off 256 and a gradient accumulation step of 2, following the SGD optimizer with a momentum of 0.9, $L2 = 5 \times 10^{-4}$, and an initial learning rate of 0.1 that is decayed by 0.1 at epochs 24, 30, 36. Images are normalized with $\mu = [0.5, 0.5, 0.5]$ and $\sigma = [0.5, 0.5, 0.5]$ and are augmented through random cropping and flipping along with color jittering.

D. Sensitivity Analysis of the Identity Loss

To understand the effect of the auxiliary cross-entropy identity loss on our synthetic data, we conduct a sensitivity analysis over six weightings of the classification-loss term, $\lambda \in \{\mathbf{0}, \mathbf{0.0001}, \mathbf{0.0005}, \mathbf{0.001}, \mathbf{0.005}, \mathbf{0.01}\}$, where $\lambda = 0$ corresponds to the baseline Arc2Face [36] training without identity supervision. For IDSync training, we first train an identity classifier on CASIA-WebFace, achieving a validation accuracy of 84% with hyperparameters from Section IV-B. We then train a separate IDSync model for λ value on the same dataset. For each model obtained at a given λ, we generate a synthetic dataset of 0.5M images (10,000 identities $\times$ 50 images per identity), following standard practice of existing works [5], [36]. The exact training and data-generation procedures are detailed in Section IV-B.

Quality, fidelity, and diversity of images. We begin our evaluation by comparing the distributions of synthetic and real-world images in terms of extracted DINOv2 features with measures defined in Section IV-C. Results reported in Table I indicate that introducing an additional criterion improves the alignment between the distributions. Specifically, a weight of $\lambda = 0.0001$ yields the lowest Fréchet and kernel distances, demonstrating highest overall quality as the synthetic and real distributions are the most similar. Conversely, $\lambda = 0.005$ produces the highest density and coverage, corresponding to maximal image fidelity and diversity compared to real-world data. Our method is primarily optimized for identity consistency; thus, the observed variations between $\lambda = 0.0001$ and $\lambda = 0.005$ simply indicate that the additional loss term does not compromise image fidelity or diversity. These observations are further supported by synthetic samples seen in Figure 3.

Identity similarity. Next, we investigate the inter- and intra- class cosine similarities of identity features extracted a pretrained ArcFace-based [10] recognition model. In Figure 4, we see that the intra-class similarity distribution for

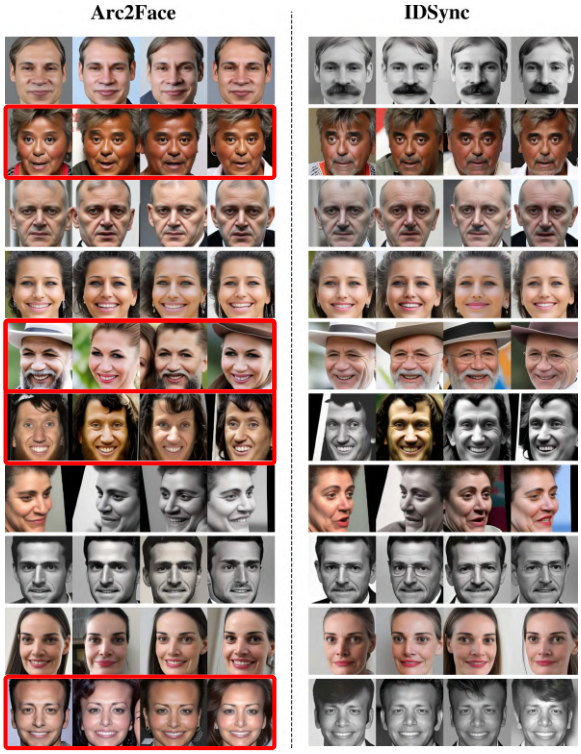

Fig. 3. **Identities generated by Arc2Face [36] and IDSync ($\lambda = 0.001$).** The models are conditioned on identical synthetic ArcFace [10] vectors to produce four images per identity. Images outlined in red mark Arc2Face failure cases; the corresponding IDSync samples in the same rows better preserve identity under the same conditioning vectors.

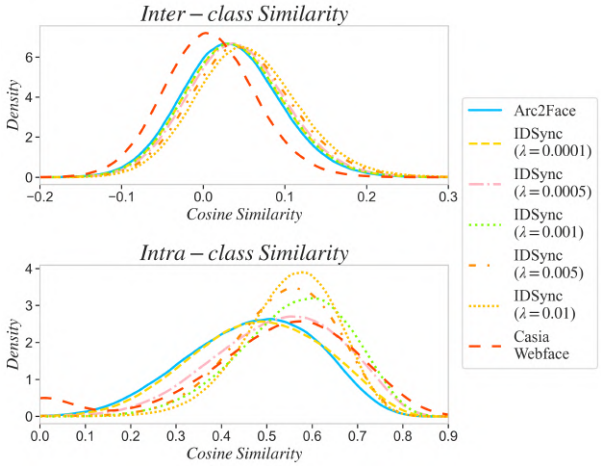

Fig. 4. **Intra-class and inter-class identity similarity.** Distributions are based on cosine similarities of ArcFace [10] embeddings extracted from synthetic images generated by the Arc2Face [36] model and IDSync with various weights λ, compared to real images of CASIA-WebFace [57].

Arc2Face exhibits lower cosine-similarity values, meaning that images of the same identity are less similar to one another compared to those produced with IDSync. We also observe a clear shift toward higher cosine similarities as the weight λ increases, with these distributions progressively converging on the distribution of the real dataset.

Pose variability. Face-angle diversity is critical for robust face-recognition training, especially on extreme-pose benchmarks like CP-LFW and CFP-FP. Although one might expect the added classification loss to push generation toward

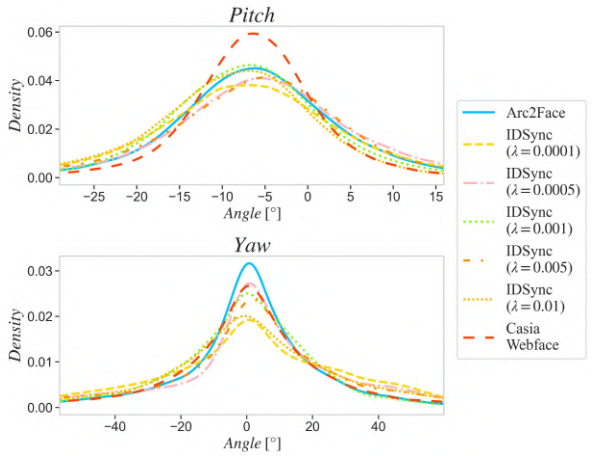

Fig. 5. **Face pose comparison of synthetic and real data.** Distributions entail face orientation angles of synthetic images generated by the Arc2Face [36] and IDSync with various weights λ, and real images of CASIA-WebFace [57]. *Roll is omitted* because standard FR preprocessing (face cropping and alignment) largely normalizes in-plane rotation.

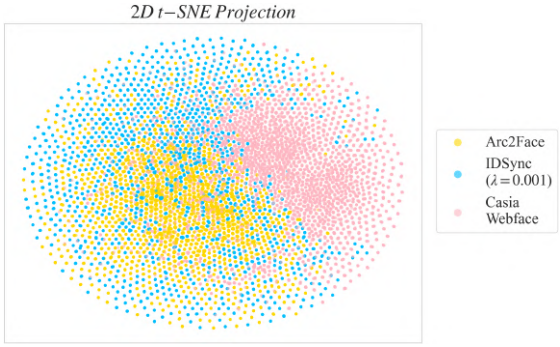

Fig. 6. **t-SNE visualization of synthetic and real identity features**. Identity features are obtained with a pretraiend ArcFace [10] model from synthetic images generated by Arc2Face and IDSync ($\lambda = 0.001$), and from real images of CASIA-WebFace. Jensen–Shannon divergence between distributions in the t-SNE space: Arc2Face vs. CASIA-WebFace = **0.494**, Arc2Face vs. IDSync = **0.101**, CASIA-WebFace vs. IDSync = **0.428**.

predominantly frontal poses with stronger semantic cues, Figure 5 shows that the full range of orientations is maintained. In fact, the yaw distribution of IDSync is even more similar to that of CASIA-WebFace compared to Arc2Face, indicating that IDSync better preserves pose variability.

Clustering behavior and sample separability. We apply t-SNE to ArcFace embeddings of 1,000 identities to visually compare the distributions of Arc2Face and IDSync in low-dimensional space, as shown in Figure 6. The visualization, together with reported pairwise Jensen–Shannon divergence values, shows that IDSync more closely matches the real CASIA-WebFace distribution than Arc2Face, with both methods clearly forming separate clusters.

Face-recognition performance. As shown in Table II, adding the cross-entropy identity loss during training consistently improves the quality of generated images, leading to better face recognition performance than without the identity loss (i.e. ($\lambda = 0$)). The best results are achieved with a weight of $\lambda = 0.001$, which improves the performance of Arc2Face by more than 1%. For larger weights, performance saturates or slightly degrades, likely because the generator overemphasizes the identity objective and produces less diverse samples per identity, as also reflected by a narrower intra-class similarity distribution in Figure 4.

TABLE II

VERIFICATION BENCHMARK ACCURACIES (%) OF THE ADAFACE MODEL TRAINED ON DATA GENERATED WITH ARC2FACE OR ID-SYNC WITH DIFFERENT WEIGHTS λ, AND REAL DATA FROM CASIA-WEBFACE. **BLUE** HIGHLIGHTS THE BEST VALUE; **YELLOW** THE SECOND-BEST.

Dataset	Casia-Web0.5M	Arc2Face (0)	IDSync (0.0001)	IDSync (0.0005)	IDSync (0.001)	IDSync (0.005)	IDSync (0.01)
AgeDB	94.08	76.23	76.49	74.81	75.48	77.57	76.50
CFP-FP	96.56	88.78	89.40	89.13	90.84	88.68	90.14
LFW	99.42	97.66	97.96	98.18	98.27	98.03	97.86
CPLFW	89.73	82.68	84.28	85.05	84.79	84.01	82.77
CALFW	93.32	82.31	84.20	84.12	84.52	84.16	83.80
Average	94.62	85.53	86.47	86.26	86.78	86.49	86.21

E. Scalability Evaluation

To evaluate the scalability of IDSync, we retrained the model on a larger dataset of 2M images from WebFace42M, as described in Section IV-A. This setup mirrors the 0.5M-image experiment but uses more training data to test whether IDSync maintains identity consistency and generation quality at scale. We also retrained the auxiliary identity classification model on the same 2M-image subset, achieving a validation accuracy of 94% using the hyperparameters from Section IV-B. After training, we synthesized new identities using both IDSync (with $\lambda = 0.001$) and Arc2Face. As shown in Figure 7, IDSync produces more visually coherent and realistic faces than Arc2Face, whose outputs frequently exhibit exaggerated facial features combined with artificial textures and unnatural lighting, across both training and unseen identities. To assess the quality and utility of generated data, we trained an AdaFace [26] recognition model on the synthetic datasets produced by both methods, following the same protocol from Section IV-C. The results, summarized in Table III, confirm that IDSync maintains strong performance on all five benchmarks when scaled to larger training data, consistently outperforming Arc2Face [36] under identical training settings. For reference, the table also reports accuracies of other state-of-the-art methods that utilized different training data, parameters, or generation procedures, including SynFace [39], DigiFace [2], DCFace [27], Vec2Face [53] and IDiff-Face [5]. Additionally, we include results obtained with the original Arc2Face model, which was trained on the full WebFace42M dataset, that entails 42M images compared to the 2M used in our setting. Afterwards, the model was also fine-tuned further on FFHQ and CelebA-HQ [36]. For Vec2Face, we report results both with and without their proposed attribute operation (AttrOP), which perturbs input identity embeddings to encourage greater intra-identity variation. For a fair comparison of generative capabilities, we also adopt the same identity features used by Vec2Face with AttrOP for conditioning IDSync - marked with †. The results in Table III demonstrate that IDSync† outperforms Vec2Face on the pose-challenging CFP-FP and CPLFW benchmarks and achieves higher average performance across all datasets.

Fig. 7. **Training and unseen identities generated with Arc2Face and IDSync.** Identities are generated with ArcFace embeddings from the training and unseen datasets, i.e. WebFace2M (Top) and FFHQ (Bottom).

TABLE III

VERIFICATION ACCURACIES (%) OF THE ADAFACE MODELS TRAINED ON SYNTHETIC DATA. *WF* DENOTES WEBFACE SUBSETS WITH 42M, 4M AND 2M IMAGES RESPECTIVELY, WHILE *C-WF* DENOTES CASIAWEBFACE0.5M. RESULTS FOR PRIOR BASELINES NOT TRAINED ON WF2M ARE TAKEN FROM THEIR CORRESPONDING PAPERS.

Model	Train set	AgeDB	CFP-FP	LFW	CPLFW	CALFW	Avg.
SynFace [39]	FFHQ	61.63	75.03	91.93	70.43	74.73	74.75
DigiFace [2]	–	76.97	87.40	95.40	78.87	78.62	83.85
IDiff-Face [5]	FFHQ	84.63	82.39	97.68	79.70	90.58	87.00
DCFace [27]	C-WF	89.70	85.33	98.55	82.62	91.60	89.56
Arc2Face [36]	WF42M	90.18	91.87	98.81	85.16	92.63	91.73
Vec2Face [53]	WF4M	90.75	76.56	98.27	81.70	92.92	88.04
+ AttrOP [53]	WF4M	93.12	88.97	98.87	85.47	93.57	92.00
Arc2Face	WF2M	77.89	90.02	97.94	85.04	85.19	87.22
IDSync (cos)	WF2M	85.33	91.56	98.15	83.88	88.45	89.47
IDSync	WF2M	87.13	91.61	98.30	84.17	88.74	89.99
IDSync†	WF2M	89.52	93.16	98.70	87.15	91.66	92.04

F. Ablation Studies

Replacing the identity loss. In addition to the cross-entropy identity classification objective, we also evaluate an embedding-level cosine alternative by replacing $\mathcal{L}_{ID}$ with $1 - \cos(\mathbf{e}_{\text{gen}}, \mathbf{e}_{\text{ref}})$, where $\mathbf{e}_{\text{gen}}$ and $\mathbf{e}_{\text{ref}}$ are ArcFace embeddings of the generated image and the target identity, respectively. Results for both loss variants are reported in Table III. Overall, the cross-entropy objective performs better than the cosine alternative. This may be due to cross-entropy providing multi-class competition with implicit negatives [4], versus pairwise cosine alignment between identity embeddings, which can lead to overfitting on unintentional cues present within them, e.g., age and pose [48].

Replacing conditioning vectors. As part of our ablation study, we first investigate how replacing ArcFace [10] conditioning vectors with deterministic Gray codes affects the training of IDSync. With this, we aim to reveal whether a structured but semantically uninformative class representation can serve as a sufficient substitute. Results in Figure 8 demonstrate that semantically uninformative Gray codes are not a suitable alternative for class conditions, as IDSync training fails to converge meaningfully.

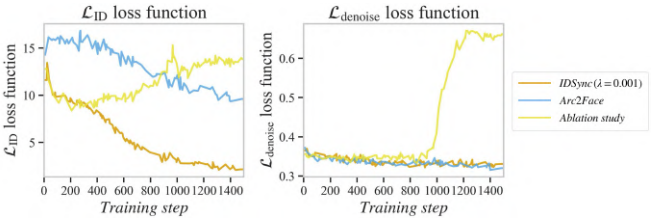

Fig. 8. **Loss function values for the ablation study.** Here, IDSync ($\lambda = 0.001$) is conditioned on semantically uninformative Gray codes, with each identity having a unique code, instead of ArcFace [10] vectors. We also report $\mathcal{L}_{\text{ID}}$ for Arc2Face, though it is only used for training with IDSync.

Fig. 9. **Results of enforcing a different identity during training.** During training $\mathcal{L}_{\text{ID}}$ enforces a different identity than the one used to condition the denoising UNet via the CLIP text encoder.

Injecting conflicting identities. In the second part, we examine the impact of the classification loss function on the learning process, where the model is conditioned during training on the ArcFace vectors of one identity, while an additional $\mathcal{L}_{\text{ID}}$ loss function is used to deliberately impose a different identity. After training, we present the results in Figure 9, which illustrate the identity-mixing effect of injecting a conflicting identity using $\mathcal{L}_{\text{ID}}$.

V. CONCLUSION

In this work, we introduce IDSync, a diffusion-based face synthesis framework that improves identity consistency across generated samples by incorporating an auxiliary identity classification objective during training. Unlike prior approaches that rely solely on embedding-level objectives (e.g., cosine similarity), our method uses a cross-entropy classification loss, enabling better alignment between generated images and intended identities. We conduct extensive experiments on mid-scale (CASIA-WebFace) and large-scale (WebFace2M) datasets, evaluating synthetic face quality via statistical measures and downstream recognition. IDSync outperforms Arc2Face [36] in identity fidelity and pose diversity, and enables training of models with higher accuracy on verification benchmarks than existing state-of-the-art methods. Ablation studies highlight the importance of conditioning on semantically meaningful identity embeddings and reveal identity-mixing effects when conflicting identity signals are injected. Overall, IDSync offers a scalable, effective solution for identity-consistent synthetic face generation, well-suited for privacy-conscious training of face recognition systems. Future work includes adaptive weighting of the identity loss and extensions to multimodal conditioning along with the analysis of potential performance disparities across different demographic groups.

Ethical Impact Statement

Intended use and potential benefits. IDSync is a generative framework designed to synthesize identity-consistent face images for research on training and evaluating FR systems. By enabling the creation of large, diverse datasets without collecting equivalent volumes of real images, our approach can help reduce reliance on web-scraped faces and lower the privacy and data-governance burden typically associated with curating real-world FR datasets.

Data sources and handling. Experiments fine-tune and evaluate models using established, publicly available research datasets for training and benchmarking. Where synthetic datasets are created, we generate new identities by sampling and perturbing identity embeddings and produce multiple images per synthetic identity; downstream FR models in our study are trained only on such synthetic images rather than newly acquired real faces. No personally identifiable images beyond those contained in standard research datasets are collected for this work, and we do not release any real-face images.

Risks and dual-use considerations. As with other generative methods, there is potential for misuse, including unauthorized impersonation, deception (e.g., deepfakes), or privacy-invasive surveillance applications. Moreover, biases and representation gaps present in training corpora (and in the identity-embedding space) may be reflected or amplified in generated imagery, potentially leading to disparate performance across demographic groups. Finally, conditioning on identity features can, in principle, be abused to approximate the appearance of real individuals.

Mitigations, safeguards, and release practices. Our experiments focus on privacy-conscious data generation for FR research and model training; we explicitly discourage use for identity spoofing or surveillance. If code and models are released, we will accompany them with clear usage guidelines and a license that prohibits deceptive, privacy-invasive, or discriminatory uses, and we recommend integrating safeguards such as provenance/watermarking, rate-limits, and human-in-the-loop review where deployment is contemplated. We will not release any real-face imagery and will document known limitations and failure modes.

Limitations and future work. We do not claim that synthetic data eliminates all ethical risks in biometrics. Future work includes (i) systematic fairness auditing (with demographically stratified reporting), (ii) bias-mitigation strategies in both conditioning and sampling, and (iii) improved misuse deterrence via stronger provenance signals and easier detection of synthetic content. We encourage practitioners to assess context-specific risks and comply with applicable laws, community norms, and institutional review requirements before deploying any FR system built with synthetic data.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, volume 70, pages 214–223, 2017.

[2] G. Bae, M. de La Gorce, T. Baltrušaitis, C. Hewitt, D. Chen, J. Valentin, R. Cipolla, and J. Shen. DigiFace-1M: 1 million digital face images for face recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3526–3535, 2023.

[3] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations (ICLR)*, pages 1–36, 2018.

[4] M. Boudiaf, J. Rony, I. M. Ziko, E. Granger, M. Pedersoli, P. Piantanida, and I. Ben Ayed. A unifying mutual information view of metric learning: Cross-entropy vs. pairwise losses. In *European Conference on Computer Vision (ECCV)*, pages 548–564, 2020.

[5] F. Boutros, J. H. Grebe, A. Kuijper, and N. Damer. IDiff-Face: Synthetic-based face recognition through fizzy identity-conditioned diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–11, 2023.

[6] F. Boutros, M. Huber, P. Siebke, T. Rieber, and N. Damer. SFace: Privacy-friendly and accurate face recognition using synthetic data. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11, 2022.

[7] F. Boutros, V. Struc, J. Fierrez, and N. Damer. Synthetic data for face recognition: Current state and future prospects. *Image and Vision Computing*, page 104688, 2023.

[8] L. Chen, M. Zhao, Y. Liu, M. Ding, Y. Song, S. Wang, X. Wang, H. Yang, J. Liu, K. Du, and M. Zheng. PhotoVerse: Tuning-free image customization with text-to-image diffusion models, 2023. arXiv preprint arXiv:2309.05793.

[9] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5203–5212, 2020.

[10] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019.

[11] P. Dhariwal and A. Q. Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8780–8794, 2021.

[12] C. N. Duong, T. D. Truong, K. Luu, K. G. Quach, H. Bui, and K. Roy. Vec2face: Unveil human faces from their blackbox features in face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6132–6141, 2020.

[13] I. C. Duta, L. Liu, F. Zhu, and L. Shao. Improved residual networks for image and video recognition. In *IEEE International Conference on Pattern Recognition (ICPR)*, pages 9415–9422, 2021.

[14] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *PMLR International Conference on Machine Learning (ICML)*, pages 12606–12633, 2024.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

[16] H. Gururaj, B. Soundarya, S. Priya, J. Shreyas, and F. Flammini. A comprehensive review of face recognition techniques, trends and challenges. *IEEE Access*, pages 1–24, 2024.

[17] T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi. 6d rotation representation for unconstrained head pose estimation. In *IEEE International Conference on Image Processing (ICIP)*, pages 2496–2500, 2022.

[18] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6626–6637, 2017.

[19] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

[20] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[21] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[22] H. Imaoka, H. Hashimoto, K. Takahashi, A. F. Ebihara, J. Liu, A. Hayasaka, Y. Morishita, and K. Sakurai. The future of biometrics technology: from face recognition to related applications. *APSIPA Transactions on Signal and Information Processing*, 10:1–13, 2021.

[23] C. Jasserand. *Massive Facial Databases and the GDPR: the New*

Data Protection Rules applicable to Research, pages 169–188. Hart Publishing / Bloomsbury Publishing Plc, 2018.

[24] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018.

[25] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019.

[26] M. Kim, A. K. Jain, and X. Liu. AdaFace: Quality adaptive margin for face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18729–18738, 2022.

[27] M. Kim, F. Liu, A. Jain, and X. Liu. DCFace: Synthetic face generation with dual condition diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12715–12725, June 2023.

[28] Z. Lei, T. Xu, X. Wang, S. Z. Li, and W. Deng. Cross-Age LFW: A database for studying cross-age face recognition in unconstrained environments. Technical Report 16-01, Chinese Academy of Sciences, 2016. Institute of Automation, Center for Biometrics and Security Research.

[29] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, pages 1–19, 2019.

[30] C. Lu, Y. Chen, B. Tzen, and J. Bao. DPM-Solver: A fast ODE solver for diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1–31, 2022.

[31] G. Mai, K. Cao, P. C. Yuen, and A. K. Jain. On the reconstruction of face images from deep face templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(5):1188–1202, 2019.

[32] B. Meden, P. Rot, P. Terhörst, N. Damer, A. Kuijper, W. J. Scheirer, A. Ross, P. Peer, and V. Štruc. Privacy–enhancing face biometrics: A comprehensive survey. *IEEE Transactions on Information Forensics and Security (TIFS)*, 16:4147–4183, 2021.

[33] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. AgeDB: The first manually collected, in-the-wild age database. In *Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[34] M. F. Naeem, S. J. Oh, Y. Uh, Y. Choi, and J. Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning (ICML)*, pages 1–10, 2020.

[35] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jégou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, pages 1–32, 2024.

[36] F. P. Papantoniou, A. Lattas, S. Moschoglou, J. Deng, B. Kainz, and S. Zafeiriou. Arc2Face: A foundation model for id-consistent human faces. In *European Conference on Computer Vision (ECCV)*, pages 1–29, 2024.

[37] X. Peng, J. Zhu, B. Jiang, Y. Tai, D. Luo, J. Zhang, W. Lin, T. Jin, C. Wang, and R. Ji. Portraitbooth: A versatile portrait model for fast identity-preserved personalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27070–27080, Seattle, WA, USA, 2024.

[38] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, pages 1–21, 2024.

[39] H. Qiu, B. Yu, D. Gong, Z. Li, W. Liu, and D. Tao. Synface: Face recognition with synthetic data. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10880–10890, 2021.

[40] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *PMLR International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

[41] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, pages 1–16, 2016.

[42] A. Razzhigaev, K. Kireev, E. Kaziakhmedov, N. Tursynbek, and A. Petiushko. Black-box face recovery from identity features. In *ECCV Workshops*, volume 12546 of *Lecture Notes in Computer Science*, pages 462–475, 2020.

[43] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022.

[44] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351, pages 234–241, 2015.

[45] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2023.

[46] S. Sengupta, J.-C. Chen, C. D. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to profile face verification in the wild. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016.

[47] Stability AI. Stable Diffusion v2.1 Release. `https://stability.ai/news/stablediffusion2-1-release7-dec-2022`, 2022. Accessed: April 27 2025.

[48] D. Tomašević, F. Boutros, C. Lin, N. Damer, V. Štruc, and P. Peer. ID-Booth: Identity-consistent face generation with diffusion models. In *International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10, 2025.

[49] T.-D. Truong, C.-N. Duong, N. Le, M. Savvides, and K. Luu. Vec2Face-v2: Unveil human faces from their blackbox features via attention-based network in face recognition, 2022. arXiv preprint arXiv:2209.04920.

[50] L. J. P. van der Maaten and G. E. Hinton. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research (JMLR)*, 9:2579–2605, 2008.

[51] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5265–5274, 2018.

[52] X. Wang, Y. Li, H. Zhang, and Y. Shan. Towards real-world blind face restoration with generative facial prior. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[53] H. Wu, J. Singh, S. Tian, L. Zheng, and K. W. Bowyer. Vec2Face: Scaling face dataset generation with loosely constrained vectors. In *International Conference on Learning Representations (ICLR)*, pages 1–25, 2025.

[54] Y. Yan, C. Zhang, R. Wang, Y. Zhou, G. Zhang, P. Cheng, G. Yu, and B. Fu. FaceStudio: Put your face everywhere in seconds. 2023. arXiv preprint arXiv:2312.02663.

[55] S. Yang, P. Luo, C.-C. Loy, and X. Tang. WIDER FACE: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5525–5533, 2016.

[56] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[57] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch, 2014. arXiv preprint arXiv:1411.7923.

[58] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023.

[59] T. Zheng and W. Deng. Cross-Pose LFW: A database for studying cross-pose face recognition in unconstrained environments. Technical Report 18-01, Beijing University of Posts and Telecommunications, februar 2018.