CemiFace: Center-based Semi-hard Synthetic Face Generation for Face Recognition

Zhonglin Sun
Queen Mary University of London
`zhonglin.sun@qmul.ac.uk`

Siyang Song[*]
University of Exeter
`ss2796@cam.ac.uk`

Ioannis Patras
Queen Mary University of London
`i.patras@qmul.ac.uk`

Georgios Tzimiropoulos
Queen Mary University of London
`g.tzimiropoulos@qmul.ac.uk`

Abstract

Privacy issue is a main concern in developing face recognition techniques. Although synthetic face images can partially mitigate potential legal risks while maintaining effective face recognition (FR) performance, FR models trained by face images synthesized by existing generative approaches frequently suffer from performance degradation problems due to the insufficient discriminative quality of these synthesized samples. In this paper, we systematically investigate what contributes to solid face recognition model training, and reveal that face images with certain degree of similarities to their identity centers show great effectiveness in the performance of trained FR models. Inspired by this, we propose a novel diffusion-based approach (namely **Ce**nter-based Se**mi**-hard Synthetic Face Generation (**CemiFace**)) which produces facial samples with various levels of similarity to the subject center, thus allowing to generate face datasets containing effective discriminative samples for training face recognition. Experimental results show that with a modest degree of similarity, training on the generated dataset can produce competitive performance compared to previous generation methods. The code will be available at:https://github.com/szlbiubiubiu/CemiFace

1 Introduction

Face Recognition (FR) has gained significant achievement in recent years owing to the combination of discriminative loss function [1–5], proprietary backbones [6–11] and large-scale face datasets [12–15]. For example, with a 4M training set, existing FR models can achieve over 99% accuracy on various academic datasets [3, 6, 2, 16]. However, in real-world industrial face recognition applications, collecting large-scale face datasets is not always available due to the related licence agreements, ethical issues and privacy policies [17].

To expand limited training samples in real-world scenarios, generative models [18–23] are widely adopted owing to their ability to generate high-quality images. However, simply adopting face images produced by those generic generative models to train face recognition models is impractical as there is ambiguity about the identities of the produced images because they are derived from random noises, i.e., the identities of these generated face images cannot be obtained without a well-trained FR model [24]. To address such issues, synthetic face dataset generation-based solutions [24–28] have been found to gain benefits in developing effective face recognition models. Existing synthetic face recognition (SFR) methods are frequently built upon recent advances of generative models

---

[*]Corresponding Author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

such as Style-GAN [25, 27], Diffusion methods [24, 27] and 3DMM rendering [26]. For instance, a style-transferring diffusion model-based method namely DCface [24] is proposed, which increases the diversity of existing face recognition datasets by generating additional discriminative face images with different styles (e.g. hair, overall lighting, which can be observed in visualization Section 4.3.2) for each subject. However, domain gap issues exists as the model is trained with paired face images belonging to the same identity, while those paired images are not available at the inference stage. It can only take samples belonging to different identities at the inference stage, which may negatively impact on the images synthesized at the inference stage. Furthermore, the definition of discriminative facial images remains unclear in this study.

To address the problems outlined above, firstly we explore the factors resulting in performance degradation for SFR and reveal that previous approaches fail to consider the properties of effective FR training–relationship/similarity between samples. Consequently, considering the facts: (a) semi-hard negative samples are crucial to train effective face recognition model for Triplet loss [29]; (b) samples close to the decision boundary contribute most to the training gradient [3]; (c) all face images belonging to the same subject can be represented by a hypersphere in the latent feature space [30] (i.e., can be measured by existing FR models, e.g., AdaFace [3]), whose distances (radius) to the identity center are negatively correlated to their similarities to the center. We hypothesize that face recognition performance is sensitive to the data with different levels of similarity to the identity center in the hypersphere, and experimentally reveal that the optimal performance is obtained with samples of mid-level similarity, which we term **center-based semi-hard samples**. Inspired by this crucial finding, we propose a novel diffusion-based synthetic face recognition approach (**CemiFace**) which generates **ce**nter-based se**mi**-hard face samples by regulating the similarity between the generated image and the inquiry image, through a similarity controlling factor condition. Figure 1 presents the overall hypothesis by showcasing samples with various similarities to the identity center. Comprehensive experiments are conducted to illustrate the effectiveness of our proposed CemiFace. Our method achieves promising performance in synthetic face recognition (SFR). The main contributions and novelties of this work are summarized as follows:

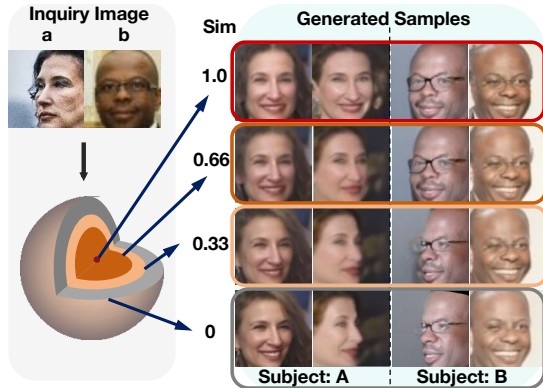



Figure 1: Visualization of the samples with different similarities. Given an inquiry image, it can form a hypersphere based on similarity to the inquiry image, where samples with the same similarity share the same radius. Samples with similarities between 0 to 1 with an interval of 0.33 are shown. With our proposed CemiFace, each inquiry image finally forms a novel subject.

- We provide the first comprehensive analysis to illustrate how FR model performance is affected by different levels of similarity of samples, particularly center-based semi-hard samples.

- We propose a novel diffusion-based model CemiFace that can generate face images with various levels of similarity to the identity center, which can be further applied to generate infinite center-based semi-hard face images for SFR.

- We demonstrate our method can be extended to use as much as the data without label supervision for training which is an advantage over the previous method [24].

- Experiments show that our CemiFace surpasses other SFR methods with a large margin, reducing the GAP-to-Real error by half.

2   Related Works and Preliminary

Synthetic Face Generation for FR: With the emergency of generative models, synthesizing facial data for various facial tasks has become a critical issue, such as applications in Face Anti-sproofing [31] and Face Recognition [25, 26, 24, 28, 32, 33]. SynFace [25] aims to mix the real

2

images with the DiscoFaceGAN-generated [19] samples. DigiFace [26] uses 3DMM for rendering facial images to construct the dataset. DCFace [24] takes diffusion models to adapt style from the style bank to the identity image and result in discriminative samples with diverse styles. IDiff-Face [28] proposes the condition latent diffusion models [23] to the feature embedding and images are synthesized by pretrained decoder.

Preliminary-DDPM: Diffusion models [21, 22] are generative models which denoise an image from a random noise image. The training pipeline for diffusion models consists of a forward process wherein noise is progressively added to a given image and a denoising process to predict the estimated noise for effective denoising. A single forward process is formulated as Markov Gaussian diffusion with timestep t:

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

Where $\mathcal{N}()$ is adding noise function. When accumulating the time step over $0 - \mathbf{T}$, the final process is given as follows:

$$q(\mathbf{x}_{1:\mathbf{T}}|\mathbf{x}_0) = \prod_{t=1}^{\mathbf{T}} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{2}$$

Then the denoising process is conducted to predict the noise for the time step t using a model σ_θ (typically a UNet [34]), the training loss is:

$$L_{\mathbf{MSE}} = E_{t,\mathbf{x_0},\epsilon}[||\sigma_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x_0} + \sqrt{1-\bar{\alpha}_t}\epsilon, t) - \epsilon||_2^2] \tag{3}$$

where β_t is the pre-set forward process variances. Then notation $\bar{\alpha}_t$ is given as: $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$ and $\alpha_t = 1 - \beta_t$. ϵ is a random Gaussian noise image $\epsilon \sim \mathcal{N}(0,1)$.

3 The proposed approach

In Section 3.1, we first investigate the relationship between sample similarity and their effectiveness in training FR models, presenting the finding that samples with certain similarities (i.e., center-based semi-hard samples) to their identity centers are more effective for training FR models on a real dataset and subsequently devise a toy experiment to validate it. Inspired by our findings, in Section 3.2 we propose a novel CemiFace, a conditional diffusion model that produces images with various levels of similarity to an inquiry image. Specifically, Section 3.2.1 introduces how we construct the similarity condition which is fed to diffusion model to guide the generation, and discusses the L_{SimMat} to require the generated sample to exhibit a certain similarity degree to the inquiry image. In Section 3.2.2, we then present how to use our diffusion model to generate a synthetic face dataset given a fixed similarity condition m and a set of inquiry images.

3.1 The Relationship between Samples Similarity and Performance Degradation

Performance Degradation for Synthetic Face Recognition: Face recognition models trained on face images synthesised by existing generative models (e.g., style-transfering [24], 3DMM rendering [26] and latent diffusion expansion [28]) frequently suffer from performance degradation [24–26]. For example, with the same data volume, the model trained on the state-of-the-art synthetic dataset DCface [24] produces 11.23% lower verification performance on CFP-FP testset than the model with the same architecture trained on the real dataset. A key reason for this issue is that these generative models only intuitively explore the properties of discriminative samples, but fail to consider the similarity levels among synthesized face images. However, previous studies [29, 2, 3] empirically reveal that *training effective FR models intrinsically relies on semi-hard negative samples in Triplet Loss [29] or samples close to the decision boundary [2, 3, 35].*

Hypothesis and Findings: Since face images belonging to the same identity/class can be aggregated within a hypersphere [30], where the location of each face image is decided by its similarity to the identity center (the center of the hypersphere) (illustrated in Fig. 1). We treat all face images of each subject as an N-1 (N=512 in AdaFace [3]) dimensional sphere with its center representing the subject-level identity center. Then, the spheres of all subjects can be combined in an N-dimensional sphere, where each subject-level sphere is a cluster.

Based on this, we hypothesize that samples of mid-level similarities to the identities center play a dominant impact on the FR performance, as they exhibit discriminative style variations (e.g. age,

3

Sim	LFW	CFP-FP	AgeDB	CALFW	CPLFW	AVG
0.85	98.43	85.67	89.43	91.08	82.78	89.48
0.81	98.91	88.8	91.03	91.71	84.58	91.01
0.76	**98.94**	90.92	**91.5**	**91.7**	85.85	**91.78**
0.70	98.66	**91.08**	90.32	90.76	**86.92**	91.55
0.53	94.63	82.12	77.63	80.11	77.3	82.36

Figure 2: Samples with different similarity groups from CASIA-WebFace dataset. From left to right are samples with lower similarity to the identity center

Table 1: Accuracy of groups with different similarities. Sim means the average similarity to the identity center. AVG is the average accuracy on the 5 evaluation datasets

pose). To validate the hypothesis, we conduct the first comprehensive investigation for the impact of different levels of similarity to the identity center on the FR performance. We first split face images in the CASIA-WebFace [36] into various levels of groups according to their similarities to their corresponding subject-level identity centers. Here, the identity center of each subject is obtained by the weight of the linear classification layer, trained using AdaFace [3]. To avoid the impacts caused by different numbers of training samples, we assign around 100k face images to each group representing close similarities to their identity centers. This results in 5 distinct similarity groups. Table 1 reports the performance of model trained on each group and test on five standard face recognition evaluation datasets [37–41]. Table 9 in *Supplementary material* Section B.4 displays the similarity range of each group. We further validate the style variation in Visualization Sec 4.3.2. We also visualize randomly selected samples of each group in Figure 2. Results reveal that groups with middle-level similarities (0.76 and 0.70) produced similar but top-performing average accuracy. This indicates that **face images of a certain low similarity to their identity centers (which we refer to as center-based semi-hard samples) are essential for learning highly accurate face recognition models.** In contrast, the group whose images have the lowest similarity (i.e., 0.53) to their identity centers obtained the worst performance, which suggests that it is difficult to train an effective face recognition model with the most challenging samples (i.e., the samples are normally hard to be distinguished by human observation).

3.2 Center-based Semi-hard Face Image Generator

Inspired by the above findings, this section proposes a novel conditional diffusion model, namely CemiFace, for synthesising effective center-based semi-hard face images given the inquiry image (identity center) $\mathbf{x}$ and a pre-defined similarity controlling factor $\mathbf{m}$, based on which a new discriminative synthetic dataset is obtained to train effective face recognition models.

Methodology overview: As illustrated in the left side of Fig 3, the training process starts with adding a noise $\epsilon \sim \mathcal{N}(0,1)$ with timestep t to the clean input image x, resulting in x_t. Meanwhile, similarity conditions m and identity condition $\mathbf{E_{id}}(x)$ are fed to the diffusion model by cross-attention as illustrated in the lower right part of Fig. 3. Consequently, the diffusion Unet σ_θ outputs the estimated noise $\epsilon' = \sigma_\theta(x_t, t, m, \mathbf{E_{id}}(x))$ for denoising the image as a clean estimated image $\hat{x}_0$. Based on the obtained estimated image $\hat{x}_0$, original x and condition C_{att}, the whole model is optimized by the combination of L_{MSE} and L_{SimMat}, defined in the following Section 3.2.1 as well as the details of constructing condition.

At the inference stage (the upper right part of Fig. 3), random noise $x_t = x_\mathbf{T} = \epsilon \sim \mathcal{N}(0,1)$ and the time step $t = \mathbf{T}$ are first fed to a CD block. This results in an estimated noise $\epsilon' = \sigma_\theta(x_t, t, \mathbf{C_{att}})$. Then, a denoise step is adopted to generate x_{t-1} from x_t for efficient interface speed. This process is repeatedly conducted on the obtained denoised latent images $(x_{t-1}, x_{t-2}, \cdots, x_0)$ until $t = 0$, where x_0 is treated as the final generated face image. Here, we assign the same identity label as x to all face images generated from the inquiry image x. To ensure high inter-class variation, our inquiry images are filtered by a pretrained FR (IR-101 trained on the WebFace4M [12] dataset by AdaFace.), which enforces the similarity between each pair of query images is lower than 0.3. The number of identities is fully decided by the number of inquiry face images. The pseudo-code for training and generation are given in *Supplementary Material* Section A.3.

4

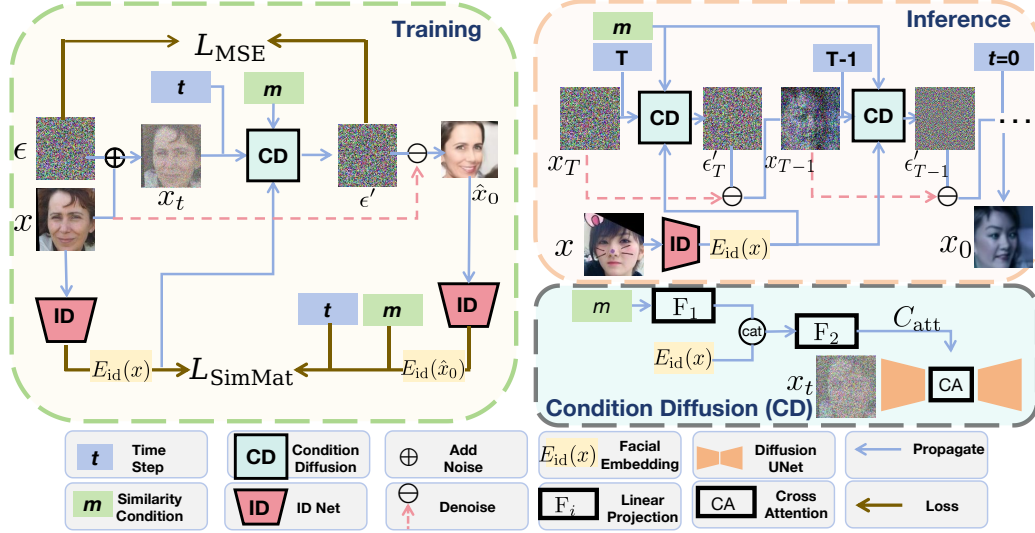



Figure 3: Illustration of our proposed method. The left part is the training framework for learning images with various levels of similarity. Firstly noise is added to the clean facial image before it is processed by the diffusion model. Then similarity controlling condition **m** ranging between [-1,1] with facial embedding is injected to guide the generation. Consequently, the model outputs the estimated noise, which is adopted to calculate the estimated image. We add similarity matching loss $L_{\textbf{SimMat}}$ between the estimated image and the input image. For generation, we gradually denoise a noising image with time step scaling from **T** to 0, conditions for identity and similarity are left fixed. The two diffusion models in the generation part mean the same diffusion model at two different time steps. The right bottom part is the details of using cross-attention to inject similarity condition and facial embedding into the diffusion models

3.2.1 Training CemiFace

To facilitate our diffusion-based CemiFace can generate diverse center-based semi-hard face images, we propose a novel diffusion model training strategy. During training, a random Gaussian noise image $\epsilon \sim \mathcal{N}(0,1)$ is firstly added to a clean face image $\mathbf{x}$ at the time step t, before feeding it to the diffusion model to generate the noise face image $\mathbf{x_t}$:

$$\mathbf{x_t} = \sqrt{\bar{\alpha}_\mathbf{t}}\mathbf{x_0} + \sqrt{1 - \bar{\alpha}_\mathbf{t}}\epsilon \tag{4}$$

Then, conditions are constructed based on the similarity controlling factor **m**, the identity condition $\mathbf{C_{id}}$ and time step t condition. Subsequently, the diffusion model outputs the estimated noise $\epsilon' = \sigma_\theta(\mathbf{x}_t, t, \mathbf{E_{id}}, \mathbf{m})$ for denoising the image.

Constructing Similarity Controlling Condition: To address the purposes of generating images at different scales of similarities, two conditions are injected into the diffusion process to guide the generation process. The first one is the identity condition $\mathbf{C_{id}}$ aiming to anchor the center of the generated facial images which can be formulated as:

$$\mathbf{C_{id}} = E_{\mathbf{id}}(\mathbf{x}) \tag{5}$$

where $E_{\mathbf{id}}$ is a pre-trained face recognition model (e.g., IResnet-50 pretrained from AdaFace [3]). $\mathbf{C_{id}}$ represents the feature embedding of the given image $\mathbf{x}$. Then the most important part is similarity controlling condition $\mathbf{C_{sim}}$ which maps the scalar similarity $\mathbf{m}$ into feature embedding. This condition serves to regulate the similarity to the inquiry image, facilitating the generation of images spanning from the most challenging samples (**m**=-1) to the most similar ones (**m**=1).

$$\mathbf{C_{sim}} = F_1(\mathbf{m}) \tag{6}$$

Where $F_i()$ is the linear projection layer. Then following DCFace [24] the two conditions are combined and projected as cross-attention conditions for sending to the DDPM process. AdaGN [42] is adopted to embed time step condition t. $\mathrm{cat}()$ is the concatenation operation.

$$\mathbf{C_{att}} = F_2(\mathrm{cat}(\mathbf{C_{id}}, \mathbf{C_{sim}})) \tag{7}$$

5

The $\mathbf{C_{att}}$ is further processed by a cross-attention operation with the intermediate latent representation of diffusion UNet σ_θ learned from the input noisy image as:

$$CA(Q, K, V, K_c, V_c) = SoftMax(\frac{QW_q([K, K_c]W_k)^T}{\sqrt{d}})W_v[V, V_c] \tag{8}$$

where $\mathbf{C_{att}}$ is treated as the key K_c and value V_c (same as DCFace) to influence the generated face images. $Q = K = V$ are the query, key and value, representing the latent feature of UNet σ_θ.

Training Loss: To ensure the similarity between the generated face $\mathbf{x_0}$ and the corresponding inquiry image (identity center) $\mathbf{x}$ adheres to the specified similarity factor $\mathbf{m}$ as given in the following equation:

$$\mathbf{m} = \text{sim}(E_{\mathbf{id}}(\mathbf{x}), E_{\mathbf{id}}(\mathbf{x_0})) \tag{9}$$

where $\text{sim}()$ denotes a similarity measurement (e.g., can be computed by Cosine Similarity or Euclidean Distance). Following DDPM [24, 21], an approximated clean sample $\mathbf{x_0}$ can be traced from $\mathbf{x}_t$ at the time step t through the following formula:

$$\mathbf{x_0} \approx \hat{\mathbf{x}}_\mathbf{0} = (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t}\epsilon')/\sqrt{\bar{\alpha}_t} \tag{10}$$

This gives a hint that the generated face image $\mathbf{x_0}$ can be controlled at the training phase by regularizing the estimated $\hat{\mathbf{x}}_\mathbf{0}$, which allows the gradient to be back-propagated to the diffusion model, e.g., controlling facial attributes [43] and styles [24]. Inspired by this, we propose a novel similarity Matching loss $L_{\mathbf{SimMat}}$ aimed at disentangling the generated face image $\mathbf{x_0}$ to exhibit a certain similarity to the inquiry image, which is determined by the similarity controlling factor $\mathbf{m}$. We employ the Time-step Dependent loss [24] with different time step t at Eq 13, specifically firstly an identity loss for recovering the identity of the original inquiry image x, which will be applied to produce original facial embedding when the time step $t \to 0$:

$$L_{\mathbf{rec}} = ||1 - \text{sim}(E_{\mathbf{id}}(\mathbf{x}), E_{\mathbf{id}}(\hat{\mathbf{x}}_0))||_2 \tag{11}$$

Then, we require the estimated $\hat{\mathbf{x}}_0$ to produce an feature embedding $E_{\mathbf{id}}(\hat{\mathbf{x}}_0)$ which matches the original $\mathbf{x}$ with $\mathbf{m}$ similarity as:

$$L_{\mathbf{sim}} = ||\mathbf{m} - \text{sim}(E_{\mathbf{id}}(\mathbf{x}), E_{\mathbf{id}}(\hat{\mathbf{x}}_0))||_2 \tag{12}$$

Consequently, the overall identity regularization loss at the time step t can be formulated as:

$$L_{\mathbf{SimMat}} = (1 - \gamma_t)L_{\mathbf{rec}} + \gamma_t L_{\mathbf{sim}} \tag{13}$$

where $\gamma_t = \frac{t}{\mathbf{T}}$ is the scaling weight for adjusting the similarity of the generated $\hat{\mathbf{x}}_0$. At the time step t=0, the model outputs an image with the same identity as the original image $\mathbf{x}$. When t scales from 0 to the maximum time step $\mathbf{T}$, the generated face image gradually shifts far away from the $\mathbf{x}$. When approaching $\mathbf{T}$, the model will output the image with $\mathbf{m}$ similarity to the original image. The proposed $L_{\mathbf{SimMat}}$ loss is inspired by the fact that facial images, with diverse styles but the same degree of similarity, are located at a circle of the hypersphere. This loss can regularize the model to learn this kind of pattern. Specifically, the similarity is guaranteed by our proposed loss, and the diversity is facilitated by the random noise ϵ of the diffusion models, which is validated in the Visualization Sec. 4.3.2. The overall training object is:

$$L = L_{\mathbf{MSE}} + \lambda L_{\mathbf{SimMat}} \tag{14}$$

where λ is a hyperparameter for balance the training focuses on noise estimation or identity-related similarity regularization.

3.2.2 Face Image Generation with Appropriate Similarity

Given a random noise ϵ and conditions(i.e. identity, similarity and time step), the well-trained model progressively denoises the noisy image ϵ with a varying time step t (from the maximum $\mathbf{T}$ to 0) and a fixed similarity factor condition $\mathbf{m}$ to generate a clean image with a specified similarity $\mathbf{m}$ to the given inquiry image $\mathbf{x}$, we adopt DDIM [22] for efficient interface speed.

We experimentally investigate the appropriate generation similarity $\mathbf{m}$ for synthetic face recognition. Specifically, we first adopt fixed similarity factors to test the best similarity. We also explore mixing the similarity around the appropriate fixed $\mathbf{m}$ (mixing semi-hard $\mathbf{m}$) and mixing appropriate fixed $\mathbf{m}$ samples with easy samples (mixing easy $\mathbf{m}$).

6

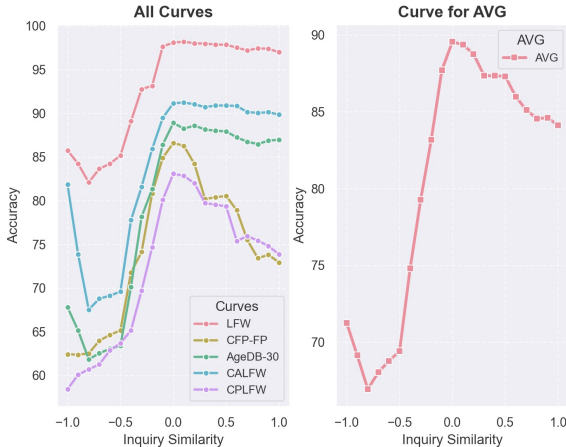
4 Experiment

4.1 Implementation Details

Evaluation Metrics: We examine the 1:1 verification accuracy trained on the dataset generated by our CemiFace on various famous testsets including LFW [37], CFP-FP [38], AgeDB-30 [39], CPLFW [40], CALFW [41] and their average verification accuracy **AVG**. Gap-to-Real is the gap to the results trained on CASIA-WebFace with CosFace loss.

Details of CemiFace Training and Generation: The condition $\mathbf{m}$ is appropriately adjusted during the training phase to facilitate better generalization across various similarities. Considering the overall cosine similarity ranges from -1 to 1, the model is enabled to discern differences in generated images under varying similarity controlling conditions when training. Specifically, in the mini-batch, we assign a randomly selected m from -1 to 1 with an interval of 0.02, allowing the model to generate corresponding images at different similarity scales. The synthetic face recognition datasets are generated in 3 volumes. Specifically in 0.5M data volume, we generate 50 images per subject and a total of 10k subjects; As for 1.0M, we keep 50 images per subject but with 20k subjects; For 1.2M, we add 5 images per subject with 40k subjects to the 1.0M settings. Oversampling method as used in DCFace is adopted which adds 5 repeated inquiry images to each subject. For more details including model, ablation studies and discussions please refer to *Supplementary material* A.1, B and C

Details of Training the Synthetic Dataset
As the training code of DCFace [24] and DigiFace [26] for training the SFR is not released. We opt for CosFace [1] with some regularizations to match the performance of DCFace [24]. Specifically, the margin of Cosface is 0.4, weight decay is 5e-4, learning rate is 1e-1 and is decayed by 10 at the 26th and 34th epoch, totally the model is trained for 40 epochs. We add random resize & crop with the scale of [0.9, 1.0], Random Erasing with the scale of [0.02,0.1], and random flip. Brightness, contrast, saturation and hue are all set to be 0.1. The backbone opted for is IR-SE50 [2]

Figure 4: Accuracy of samples with different similarity varying from -1 to 1. The left figure is the specific performance on each evaluation dataset. The right figure is the average accuracy of our CemiFace

4.2 Ablation Studies

4.2.1 Impact of Similarity m

Appropriate m for Generation: Herein we ablate how a scalar $\mathbf{m}$ influences the generation in terms of training performance. We adopt different $\mathbf{m}$ ranging from -1 to 1 with the interval of 0.1 to generate face groups of 10k identities with 10 samples per identity to match the data volume in the finding for CASIA-WebFace in Sec. 3.1. Figure 4 illustrates the accuracy curves when using those data for training face recognition (for detailed numerical results please refer to *Supplementary Material* B.4.2). Similarity $\mathbf{m}$=0 provides the best recognition performance **89.567** in terms of the AVG, then $\mathbf{m}$=0.1 has the AVG of 89.368, with $\mathbf{m}$=-0.1 obtains 87.708. It can be concluded the appropriate degree of similarity for generating discriminative samples is around 0 to 0.1, which is different from the similarity of CASIA-WebFace where the best recognition performance is obtained with the similarity of 0.7. This may be because the model for comparing similarity on CASIA-WebFace is pretrained on this dataset.

Generation with Mixing m: We conduct the experiment of including mixed $\mathbf{m}$ when generating the dataset, as is shown in the top part of Tab 2 and Tab 3. Specifically, we opt for mixing the generation m from -0.1 to 0.1, and from 0 to 0.1. We use training m varying from [0,1], and the generation interval is 0.02. The results show that mixing m with 0 to 0.1 when generating the data will bring worse performance compared to single m=0. However, mixing m with -0.1 to 0.1 obtains a similar performance compared to m=0. Additionally, progressively mixing m with easy and semi-hard

7

samples are provided in Tab 3, as observed, with more easy samples included in the training dataset, the FR performance reduced more prominent. We keep the generation m to be 0 for later discussion.

Experiment	Training m	Generation m	Interval	AVG
Mixing Generation m	[0,1]	0	0.02	**91.64**
		[0,0.1]	0.02	91.11
		[-0.1, 0.1]	0.02	91.61
Mixing Training m	[0,1]	0	0.02	91.64
	0		-	91.15
	[-1,1]		0.02	92.28
	[-1,1]		0.04	**92.30**
	[-1,1]		0.06	92.09

Table 2: Abaltion studies for mixing m in training and generation stage. The generation **m** is **mixed** with close similarity of **semi-hard** samples.

Training m	Generation m				AVG
	0	0.5	0.9	1	
[0,1]	✓				91.64
	✓	✓			90.36
	✓	✓	✓		90.12
	✓			✓	89.57

Table 3: Abaltion studies for **mixing m** in generation stage with **easy and semi-hard samples**

Training with Various m: The choice of various levels of **m** during the training stage are ablated in the bottom part of Table 2 where 3 settings are considered when training CemiFace:(a) single **m** with similarity of 0; (b) multiple discrete **m** ranging from 0 to 1 with 50 steps; (c) similar to (b) but with a range of [-1,1] and interval 0.04. Then we synthesize the data with **m**=0. As observed, setting (c) yields the best performance, indicating that with a broad range of similarity across -1 to 1, covering all the available probabilities, the CemiFace model can generalize well when adapted for generating highly discriminative samples. We also include experiments of changing the interval for (c) setting from 0.02 to 0.06 at the bottom of Tab 2, the result suggests that our approach is robust to the discrete interval but sensitive to the range of training **m**. We do not consider continuous similarity as the trained model collapses to generate the same image when given different similarities **m**.

Method	Training Data	Inquiry Data	AVG
CemiFace	CASIA	1-shot Web	**91.64**
		DDPM	91.49
		1-shot Flickr	88.97
	Flickr	1-shot Web	**90.25**
		DDPM	90.19
		1-shot Flickr	88.65
	VGGFace2	1-shot Web	**92.20**
		DDPM	92.01
		1-shot Flickr	90.586
DCFace	CASIA	1-shot Web	89.8
		DDPM	90.18

Table 4: Impact of Training and Inquiry Data. We also include results of training on DCFace for comparison

Dataset	m	AVG
WebFace	-0.1	91.27
	0	**91.64**
	0.1	90.89
DigiFace	-0.1	89.96
	0	**90.67**
	0.1	90.38
DDPM	-0.1	91.36
	0	**91.47**
	0.1	90.96

Table 5: Accuracy of the optimal **m** on different inquiry sets.

4.2.2   Ablation Study for Training and Inquiry Data

Impact of Training Data: Since our method does not require paired images for training the diffusion model, the limitation of using unlabelled data is alleviated. Consequently, we conduct experiments to see the impact of different training data. Specifically, we employ 3 datasets for training:(a) CASIA-WebFace as used in DCFace; (b) A challenging in-the-wild dataset Flickr with 1.2M images collected by us from Flickr website; (c) VGGFace2 [13] which is a large-scale dataset containing 3.3M clean images. Training **m** is set to vary within the range of [-1,1] while generation **m** is kept as 0. We do not consider data from FFHQ [18] due to restrictions on being applied for face recognition.

We can see from Table 4 using VGGFace2 as the training set produces the best performance when training a model on it, indicating that training on a large-scale dataset will bring more advance in generating discriminative dataset. However, to conduct a fair comparison with previous methods, we adopt CASIA-WebFace for the following studies. Additionally, although Flickr contains much more challenging conditions such as blurred, cartoon, and occluded faces, it results in similar performance compared to DCFace [24], which proves the effectiveness of our proposed CemiFace.

Impact of the Inquiry Data: The choice of appropriate inquiry image **x** which can be referred to as an initial point, is essential because we regard the generated group from the given **x** to be an

8

independent identity group. DCFace employs a pre-trained DDPM model [21] trained on FFHQ to generate synthetic facial images. The style bank is sampled from a real-world dataset, e.g. CASIA-WebFace [36]. Their process involves a combination of synthetic facial data and a real dataset. In contrast to DCFace, our method has fewer constraints when referring to the source data. The source data can be either synthetic or real, and we ablate the impact of using synthetic data and real data.

For taking synthetic data as the inquiry samples, we use the samples from DCface to conduct a fair comparison, noted as DDPM. As for adopting real-data, we consider two options: (a) 1-shot data randomly sampled from WebFace-4m [12] which provides a clean dataset. (b) 1-shot Flickr, a challenging dataset filtered from the one collected in Sec. 4.2.2, with fewer licence restriction. If inquiry images with high similarity, they result in overlapped groups of synthetic images in hypersphere space. Therefore, we follow DCface to filter out samples with a similarity higher than 0.3. We ablate the choice of the inquiry data source in Table 4, observing from changing the inquiry data, using 1-shot data of WebFace4M performs slightly better for our CemiFace. However, applying 1-shot WebFace4M to DCFace leads to a performance drop, as there are constraints for DCFace training and generation, e.g. frontal face and no glasses. Then using the challenging 1-shot Flickr as inquiry data brings worse results. This indicates that clean and real inquiry images are beneficial to generate discriminative datasets. Additionally, appropriate $\mathbf{m}$ for each inquiry dataset with 0.5M volume is also around 0 which can be observed in Tab 5.

Method	Data Volume	LFW	CFP-FP	AgeDB	CALFW	CPLFW	AVG	GtR
CASIA-WebFace (AdaFace)	0.49M	99.42	96.56	94.08	93.32	89.73	94.62	-
CASIA-WebFace (CosFace)†		99.3	94.87	94.35	93.15	89.65	94.26	0
SynFace		91.93	75.03	61.63	74.73	70.43	74.75	19.51
DigiFace		95.4	87.40	76.97	78.62	78.87	83.45	10.81
IDiff-Face	0.5M	98.00	85.47	86.43	90.65	80.45	88.20	6.06
DCFace		98.55	85.33	89.70	91.60	82.62	89.56	4.70
DCFace†		98.33	87.7	90.01	91.61	83.26	90.18	4.08
CemiFace, ours		**99.03**	**91.06**	**91.33**	**92.42**	**87.65**	**92.30**	**1.96**
DCFace		98.83	88.40	90.45	92.38	84.22	90.86	3.40
DCFace†	1.0M	98.88	89.71	91.25	92.15	85.2	91.44	2.82
CemiFace, ours		**99.18**	**92.75**	**91.97**	**93.01**	**88.42**	**93.07**	**1.19**
DigiFace		96.17	89.81	81.10	82.55	82.23	86.37	7.89
DCFace	1.2M	98.58	88.61	90.97	92.82	85.07	91.21	3.05
DCFace†		99.05	89.8	91.73	92.7	86.05	91.87	2.39
CemiFace, ours		**99.22**	**92.84**	**92.13**	**93.03**	**88.86**	**93.22**	**1.04**

Table 6: Comparison with the previous methods. AVG is the average accuracy of the 5 evaluation datasets. GtR is the results compared to CASIA-WebFace with CosFace. Methods with † are the results reproduced by our settings

4.3 Comparison with the State-of-Art methods

4.3.1 Quantitative Results:

We compare our CemiFace with the previous methods to demonstrate its effectiveness. The models compared are SynFace [25], DigiFace [26], IDiff-Face [28] and DCFace [24] in both 0.5M, 1M and 1.2M image volumes. The loss for training the synthetic dataset is CosFace. For the CemiFace training set, we choose CASIA-WebFace to have a fair comparison with DCFace, training $\mathbf{m}$ ranges from -1 to 1 with 50 discrete steps, and generation $\mathbf{m}$ is 0. The results are available in Table 6. In 0.5 M protocol, our method exceeds the previous state-of-art method DCface in terms of all the evaluation datasets where we achieve significant improvement on pose-sensitive dataset CFP-FP and CPLFW by 3.36 and 4.39 respectively. And in the average protocol, we get 92.30 while DCFace is 90.18. Our method still cannot exceed the model trained on the real dataset CASIA-WebFace, but we reduce the GAP-to-Real error from 4.08 to 1.96 ($\frac{4.08-1.96}{4.08} = 51.96\%$ relative error) compared to DCFace. When it refers to the 1.0M and 1.2M settings, a similar phenomenon can be observed, our method surpasses DCFace on all protocols which reduces the Gap-to-Real by half, i.e. 1.59 and 1.36. In general, CemiFace behaves well on all verification accuracy and improves pose-related performance by a large margin.

9

4.3.2 Qualitative Results

We visualize the generated results to compare with DCFace in Figure 5. Specifically, samples with different m scaling between [-1,1] with interval 0.2 are presented. For each row, we opt for the same noise to illustrate the variations across different similarities. We observe that when m is set to 1, the identity of the generated sample is very close to the inquiry image. When m is 0.4, gender and age change can be observed from the last two rows. With m scaling far away from the inquiry image, pose changes can be noticed for the first 3 rows. Another interesting phenomenon appears when similarity is -1.0 where the generated samples change significantly. Additionally, when the noise changes, the generated images exhibit different styles, aligning with our hypothesis in Sec. 3.2.1. Finally, with m=0, the group looks extremely different to the inquiry image, but can deliver highly accurate face recognition performance.

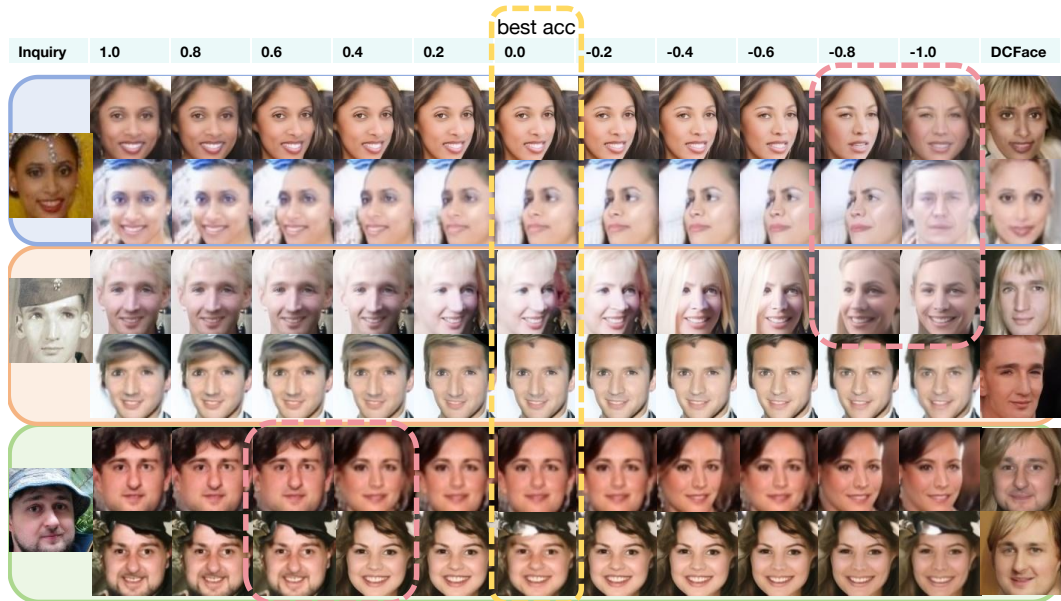



Figure 5: Sample Visualization under different similarity. From left to right are inquiry images, images with m from 1 to -1 and samples generated by DCFace. Different rows in each inquiry group represent the results produced by different noises. The first column are the inquiry images. The yellow dashed box includes samples where we obtain the best accuracy. Pink dashed boxes are samples that vary vastly.

5 Conclusion

This paper proposes a novel method to generate a discriminative dataset for training effective face recognition models with reduced privacy concerns. We investigate the factors contributing to the effective face recognition model training and re-formulate the challenge of generating discriminative samples as synthesizing center-based semi-hard samples. A similarity controlling factor condition is adopted for generating semi-hard samples. Models trained on the generated dataset with center-based semi-hard samples produce accurate face recognition performance over the previous methods. A notable advantage of CemiFace is its independence from a labelled dataset for training. However, the limitations of CemiFace include relying on the pretrained identity network's performance for conducting similarity comparisons, being sensitive to the quality of the inquiry image and privacy issues arising as the pretrained model derives from a dataset without user consent.

Acknowledgments: This work was in part funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10093336] and funded by the European Union [under EC Horizon Europe grant agreement number 101135800 (RAIDO)].

10

References

[1] 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] 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.

[3] 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*, pages 18750–18759, 2022.

[4] Yandong Wen, Weiyang Liu, Adrian Weller, Bhiksha Raj, and Rita Singh. SphereFace2: Binary classification is all you need for deep face recognition. *arXiv preprint arXiv:2108.01513*, 2021.

[5] 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*, pages 1578–1587, 2022.

[6] Zhonglin Sun and Georgios Tzimiropoulos. Part-based face recognition with vision transformers. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. URL `https://bmvc2022.mpi-inf.mpg.de/0611.pdf`.

[7] Weidi Xie, Li Shen, and Andrew Zisserman. Comparator networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 782–797, 2018.

[8] Pengyu Li. BioNet: A biologically-inspired network for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10344–10354, 2023.

[9] Jing Yang, Adrian Bulat, and Georgios Tzimiropoulos. Fan-face: a simple orthogonal improvement to deep face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12621–12628, 2020.

[10] Qiangchang Wang, Tianyi Wu, He Zheng, and Guodong Guo. Hierarchical pyramid diverse attention networks for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8326–8335, 2020.

[11] Zhonglin Sun, Chen Feng, Ioannis Patras, and Georgios Tzimiropoulos. Lafs: Landmark-based facial self-supervised learning for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1639–1649, 2024.

[12] 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.

[13] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, 2018.

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

[15] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016.

[16] Jiankang Deng, Jia Guo, Jing Yang, Alexandros Lattas, and Stefanos Zafeiriou. Variational prototype learning for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11906–11915, 2021.

[17] Protection Regulation. Regulation (eu) 2016/679 of the european parliament and of the council. *Regulation (eu)*, 679:2016, 2016.

11

[18] 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.

[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.

[20] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

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

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] Pietro Melzi, Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, Dominik Lawatsch, Florian Domin, and Maxim Schaubert. Gandiffface: Controllable generation of synthetic datasets for face recognition with realistic variations. *arXiv preprint arXiv:2305.19962*, 2023.

[28] 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.

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

[30] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[31] Yaojie Liu, Joel Stehouwer, and Xiaoming Liu. On disentangling spoof trace for generic face anti-spoofing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 406–422. Springer, 2020.

[32] 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.

[33] 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.

12

[34] 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.

[35] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14234, 2021.

[36] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

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

[38] 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 *WACV*, 2016.

[39] 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.

[40] 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), 2018.

[41] 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.

[42] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10619–10629, 2022.

[43] Bohan Zeng, Xuhui Liu, Sicheng Gao, Boyu Liu, Hong Li, Jianzhuang Liu, and Baochang Zhang. Face animation with an attribute-guided diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 628–637, 2023.

[44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[45] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

[46] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[47] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

[48] Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. *arXiv preprint arXiv:2306.13092*, 2023.

[49] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

[50] Noel Loo, Ramin Hasani, Alexander Amini, and Daniela Rus. Efficient dataset distillation using random feature approximation. *Advances in Neural Information Processing Systems*, 35: 13877–13891, 2022.

[51] Manuel Kansy, Anton Raël, Graziana Mignone, Jacek Naruniec, Christopher Schroers, Markus Gross, and Romann M Weber. Controllable inversion of black-box face recognition models via diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3167–3177, 2023.

13

This is the supplementary material for the paper **CemiFace: Center-based Semi-hard Synthetic Face Generation for Face Recognition**.

A Addition to: Implementation details

A.1 Diffusion Details

We follow most of the settings of DCFace [24]. Specifically, the model is trained on CASIA-WebFace [14] with 10 epochs. The maximum time step T for diffusion training is 1000. Then for generating the synthetic face recognition dataset, the time step for DDIM [22] is 20. The optimizer opted for is AdamW [44]. The batch size is 160 on 2 A100 GPUs. CemiFace training takes 8 hours, the generation also takes 8 hours. As a comparison, DCFace takes 10 hours for Training and 9 hours for Generation. Both DCFace and our method need around 6-7 hours to conduct FR training.

As for the diffusion UNet, we remove the identity feature in Residual Block, for more details of the Diffusion UNet please refer to DCFace [24].

A.2 High Inter-class Variations and High Intra-class Variations

(1) High Inter-class Variations: Each inquiry face image is selected to be highly independent on other inquiry images. Specifically, we follow DCFace to use a pre-trained FR model to keep samples with a threshold of lower 0.3.

(2) High Intra-class Variations: high intra-class variations are ensured by (a) changing the similarity condition m, as a small input similarity m results in the generated semi-hard images belonging to the same identity having long distances to the identity center; and (b) the face images of the same identity generated by CemiFace are distributed in all directions from the identity center, which can be observed from **supplementary material** T-SNE Fig. 7. This is guaranteed by randomly sampled Gaussian noises ϵ input to the diffusion model, which exhibits a large variation. As a result, both properties would ensure the generated face images of the same identity are almost evenly distributed in a sphere that has a relatively large radius, and thus they would have high intra-class variations.

A.3 Pseudo-code

The pseudo-code is provided below.

Algorithm 1 The training pipeline of our CemiFace

1: Initialization: Original Training Set $\mathbf{D_o}$, pretrained FR network $E_{\mathbf{id}}$, Diffusion Unet σ_θ, Maximum time step T, Maximum iteration τ, iteration $n \leftarrow 0$, similarity $m \in [-1, 1]$
2: **repeat**
3:     $n \leftarrow n + 1$
4:     Randomly sample a batch of facial images x_0 from D_o(also treated as inquiry data d), noise images ϵ from normal distribution , similarity condition from range [-1,1], single time step t
5:     construct ID & similarity condition $\mathbf{C_{att}}$ using Eq. 7.
6:     add noise $x_t \leftarrow$ use Eq. 4, given $x_0 \& t$
7:     output estimated noise $\epsilon' = \sigma_\theta(\mathbf{x}_t, t, \mathbf{C_{att}})$
8:     Update $\sigma_{\theta^{n+1}} \leftarrow \sigma_{\theta^n} - \nabla_{\sigma_\theta^n}$ Eq. 14
9: **until** converges or $n = \tau$
Output: output model σ_θ

A.4 Dataset statistics

We have also calculated the number of face images belonging to different similarity groups for CemiFace and DCFace in the Tab 7, indicating that our CemiFace tends to generate images showing lower similarities to their identity centers (i.e. all samples are semi-hard), while DCFace containing more easy samples.

14

Algorithm 2 The pipeline of CemiFace-based face dataset generation

---

1: Initialization: Inquiry Data D_I, pre trained Diffusion Unet σ_θ, Maximum time step T, fixed similarity m, Maximum Number of samples in each identity K
2: $n = 0$ is the identity index, $k = 0$ is the sample index
3: **repeat**
4:    $n \leftarrow n + 1$, $k = 0$
5:    Sample a batch of inquiry data d, construct the ID & similarity condition $\mathbf{C_{att}}$ using Eq. 7
6:    **repeat**
7:      $k \leftarrow k + 1$, $t = T$
8:      Generate noise image x_t from normal distribution $N(0, I)$
9:      **repeat**
10:        output estimated noise $\epsilon' = \sigma_\theta(\mathbf{x}_t, t, \mathbf{C_{att}})$
11:        denoise the image using following DDIM [22]$x_{t-1} \leftarrow \text{denoise}(x_t, \epsilon')$
12:        $t \leftarrow t - 1$
13:      **until** t=0
14:      assign x_0 the same label $y_d = n$ of the inquiry data, $[x_0, y_d]$
15:    **until** k=K
16: **until** $n = \text{len}(D_i)$
Output: output the generated dataset

Method	avg sim	std	Number of identites					
			0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	above 0.5
DCFace	36.24	9.14	14	231	2059	5899	1788	9
CemiFace	28.54	7.76	196	1043	3281	4539	930	7

Table 7: The statistics of the average similarity of each group. **avg sim** and **std** is the average/std similarity to the inquiry images of the whole dataset. **0-0.1** means the number of identities has a similarity of 0-0.1. CemiFace is distributed farther away from the inquiry center with less variation than DCFace.

B Further Experiments

B.1 Impact of Identity Center and Random Center

The performance of CemiFace is highly affected by the characteristics of the inquiry samples. Herein we examine how the model behaves when subjected to numerical identity conditions. Two kinds of centers are considered:(a) identity centers derived from the CASIA-WebFace dataset, and (b) random centers with a similarity range of [-0.1, 0.2] to (a). By observing from the Table 8, with random center the model results in invalid results; On the other hand, when utilizing identity centers, the model performs optimally when the similarity controlling condition $\mathbf{m}$ is set to 0 which aligns our previous finding. However, it is noteworthy that with identity center the performance is worse than the dataset inquired by 1-shot WebFace, exhibiting similar results to DCFace.

Inquiry source	sim	LFW	CFP-FP	AgeDB	CALFW	CPLFW	AVG
Random Center	1.0			Not converge			
Identity Center	1.0	96.80	71.81	86.13	89.52	71.72	83.20
	0.7	97.22	75.03	86.90	89.93	74.47	84.71
	0.5	97.50	78.96	87.12	90.38	77.62	86.32
	0.2	98.17	86.29	89.07	91.40	83.03	89.59
	0.1	**98.25**	87.30	**89.98**	91.35	83.23	90.02
	0.0	98.23	**87.49**	89.53	**91.47**	**83.73**	**90.09**
1-shot DigiFace	0.0	98.28	90.04	89.68	91.23	84.12	90.67
1-shot WebFace	0.0	99.03	91.06	91.33	92.42	87.65	**92.30**
DCFace	-	98.33	87.7	90.01	91.61	83.26	90.18

Table 8: Comparison of different inquiry centers. The results of DCFace run by our setting are copied for reference.

15

To provide deeper insights into this phenomenon, we visualize the samples generated by different inquiry centers in Figure 6. Notably, with $m=1$ the random center produces images with different identities which can simply be concluded by human observation. Conversely with the identity center, given a similarity of 1.0, the generated samples appear highly similar, except for the samples circled in red. Further investigation reveals that the number of images in that subject comprises approximately 16 images while the left subject provides approximately 50 images. Intuitively, A model trained on this dataset will focus more on the subjects with a large number of images which explains the suboptimal results obtained by identity center.

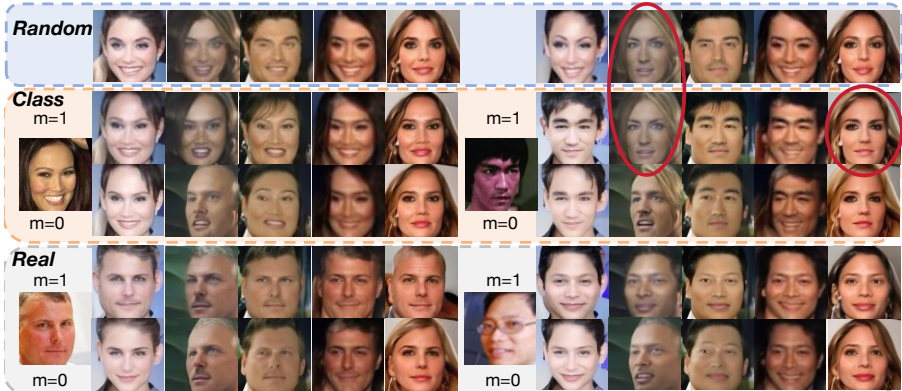



Figure 6: Comparison of different inquiry center. From top to bottom are images inquired by *Random Center*, CASIA *Identity Center* and 1-shot *Real* images. For *Identity Center* and 1-shot *Real* images, images similarity of 1 and 0 are shown. Different columns represent given different noise. Two examples are shown for each case. The inquiry images in the identity center are selected from the dataset. The red circles contain samples that look extreme different from the inquiry center.

We further visualize the T-SNE of the feature embedding in Figure 7. As shown in the upper figure, with higher similarity, the samples tend to cluster in the central region. Subsequently, by inspecting the bottom figure, it becomes apparent that with a similarity of 1.0, each subject is located in a different specific area. Consequently, a similarity of -1.0 results in each image being positioned close to other subjects in the middle area.

B.2   Addition to the Inquiry Data: Image Quality

The above discussion validates how CemiFace is affected by different centers in the aspect of numerical results. For a better understanding of the negative impact brought by challenge inquiry data such as 1-shot Flickr, we visualize the images generated from different image quality in Figure 8. Specifically, we present inquiry images subjected to *blur*, *occlusion*, *extreme pose*, *painted* and *clear* conditions, with a similarity controlling condition m set to 0. By comparing the last block with the rest of the blocks, one can conclude that extreme image quality fails to generate clean images. In conclusion, unblurred, non-occluded, appropriately posed, and real-world data are essential for our model to generate a highly clean synthetic face recognition dataset.

B.3   Further Ablation Studies

B.3.1   Impact of Different Pretrained loss

As DCFace hasn't released its AdaFace-based SFR training code and details, we were not able to reproduce it for our model training. Thus, in Tab 6 fairly compare ours with DCFace by adopting the same pre-trained AdaFace model to train our diffusion generator, and then employing the same CosFace loss for both ours and DCFace's SFR models training. Results show that our CemiFace still outperformed the SOTA DCFace. Additionally, we provide results achieved by using pre-trained model trained by CosFace. Specifically, we apply a model pre-trained by CosFace to train both our generator, and employ the same CosFace loss for their SFR models' training. The experiment shows that the model pretrained from CosFace performs better than that of AdaFace.

16

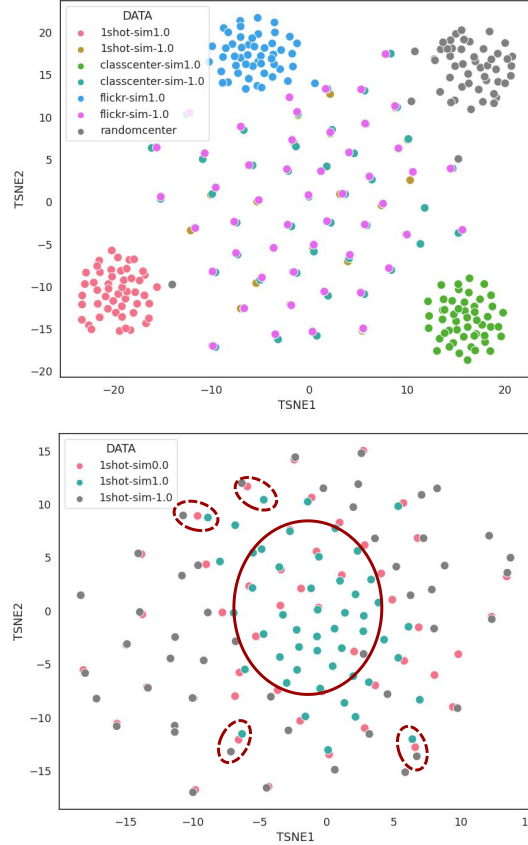



Figure 7: T-SNE visualization. The bottom figure is the T-SNE generated by 1-shot data with similarity of 1.0, 0.0 and -1.0 respectively. The upper figure is different inquiry centers with two similarities 1.0 and -1.0, the random center is also given. Red circles are samples worth noticing, with their order being green, red, and grey, positioned from center to outside

Method	Pretrained FR	SFR loss	AVG
CASIA-WebFace	-	AdaFace	94.62
CASIA-WebFace	-	CosFace	94.26
CemiFace	AdaFace	CosFace	92.30
CemiFace	CosFace	CosFace	**92.60**

B.4 Upper/Lower Bound of Different Similarity Group in CASIA-WebFace dataset

The range of each similarity group in the Section 4.2.1 is given in the following Table 9

B.4.1 Impact of Different Training Backbone

Following previous works(DCFace [24], DigiFace [26], SynFace [25]), we use the IResnet-SE-50 modified by ArcFace [2] as the default backbone. Additionally, we provide the results achieved by IResnet-18(R18), IResnet-SE-50(R50) and IResnet-SE-100(R100) in table 10 for reference.

B.4.2 Numercial Results for Different $\mathbf{m}$

Here we provide the numerical results for the impact of different similarity levels in Tab 11, $\mathbf{m} = 0$ provide the best performance.

17

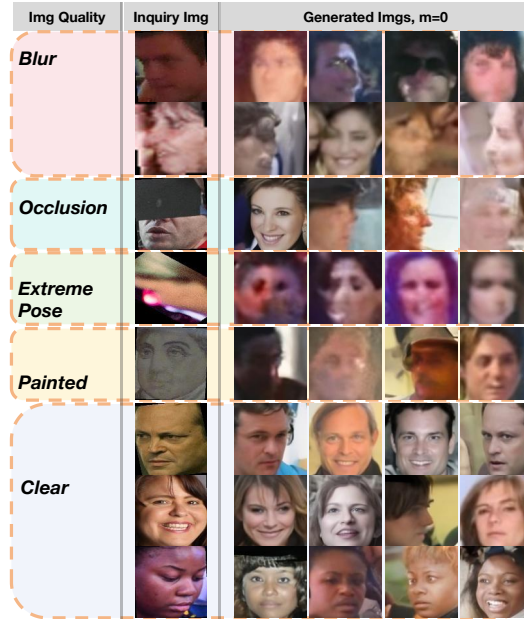



Figure 8: Examples of samples under challenging conditions, including Blur, Occlusion, Extreme Pose, and Painted conditions, are presented. Samples generated by clear images are appended for better comparison.

Avg Sim	average largest sim	average lowest sim	AVG
0.85	0.887	0.831	89.48
0.81	0.831	0.794	91.01
0.76	0.794	0.747	91.78
0.70	0.747	0.676	91.55
0.53	0.767	0.277	82.36

Table 9: **Average largest sim** represents the mean value of the largest similarity values appeared in every identity; and **Average lowest sim** represents the mean value of the lowest similarity values appeared in every identity

B.4.3   FID Image Quality

We use Fréchet Inception Distance(FID) which measures the distribution similarity of the given two datasets. Specifically, in Tab 12, FID is reported by comparing randomly selected 10k samples with randomly selected CASIA. Need to note that our method doesn't intend to generate images similar to the distribution of CASIA-WebFace, but to construct a discriminative dataset that is conducive to providing highly accurate FR performance

B.4.4   Euclidean Distance

As shown in Tab 13 using Euclidean distance leads to worse performance than cosine similarity, which might be due to the FR training loss (CosFace [1]) being carried on cosine similarity.

B.4.5   Impact of λ

We present the results using different λ in the left part of the Tab 14. Performance is sensitive to λ, and large λ results in performance degradation.

18

Backbone	R18	R50	R100
AVG	90.75	91.64	91.82

Table 10: Impact of different training backbone

Sim	LFW	CFP-FP	AgeDB-30	CALFW	CPLFW	AVG
1	97	72.94	86.98	89.85	73.86	84.126
0.9	97.38	73.81	86.88	90.13	74.82	84.604
0.8	85.75	62.42	67.8	81.85	58.43	71.25
0.7	97.2	75.5	86.75	90.15	75.95	85.11
0.6	97.52	78.91	87.25	90.84	75.39	85.982
0.5	97.85	80.55	87.93	90.9	79.35	87.316
0.4	97.88	80.39	88.01	90.89	79.55	87.344
0.3	97.98	80.19	88.15	90.72	79.73	87.354
0.2	98.02	84.21	88.6	91.03	81.99	88.77
0.1	98.2	86.29	88.25	91.25	82.85	89.368
0	98.1	86.6	88.9	91.15	83.08	**89.567**
-0.1	97.65	84.9	86.42	89.47	80.1	87.708
-0.2	93.15	80.83	81.33	85.92	74.68	83.182
-0.3	92.77	74.13	78.15	81.58	69.72	79.27
-0.4	89.11	71.78	70.13	77.78	65.17	74.794
-0.5	85.18	65.16	63.42	69.58	63.68	69.404
-0.6	84.23	64.63	63.05	69.13	62.86	68.78
-0.7	83.65	63.98	62.53	68.78	61.26	68.04
-0.8	82.1	62.51	61.85	67.53	60.7	66.938
-0.9	84.23	62.38	65.13	73.85	60.08	69.134
-1	85.75	62.42	67.8	81.85	58.43	71.25

Table 11: Numercial results for the impact of different similarities

C Privacy Concerns

In this section, we are going to discuss the privacy issues that lie in developing synthetic face generation for face recognition. The primary aim of synthetic face recognition is to mitigate concerns associated with privacy. Large-scale face recognition data are usually collected from web scrappers by searching name lists (usually celebrities), without obtaining user consent. Consequently, some of the large-scale datasets [13, 15] are abandoned by their collector to avoid Legal Risk. In addition, IDiff-Face [28] mentions European Union (EU) has come up with the General Data Protection Regulation (GDPR) [17] to regulate the application of facial data, making it harder to use face recognition data.

We notice that DCFace [24] incorporates a labelled dataset for training style transferring solution, and when they generate the new dataset, they use samples provided by DDPM [21] trained on FFHQ [18]. However, a noteworthy concern arises as the FFHQ dataset, whose derivative model is used as pretrained model in DCFace for sample generation, explicitly bans its application in face recognition. Consequently, we are not sure whether the model and synthetic face images based on FFHQ are allowed to be used. We try to avoid privacy concerns from the aspect of collecting Flickr which contains diverse licenses with reduced privacy problems. Another potential solution to avoid privacy concerns is to use samples like Digiface [26] which is rendered by 3DMM. However, DigiFace is only allowed to be adopted for non-commercial applications, but one can render images from 3DMM following the DigiFace pipeline for commercial purposes. We append the result inquired by 1-shot DigiFace in the bottom part of Table 8 for reference and example images generated by 1-shot DigiFace are shown in Figure 9. Results reveal that 1-shot DigiFace still can not surpass 1-shot WebFace but still behave better than DCFace. Finally, although 1-shot Digiface samples sometimes don't appear to be like real humans, the generated samples exhibit similar patterns to real-world images from human observation.

Our method CemiFace offers the advantage of not requiring labels during the training phase compared to DCFace. Nonetheless, both our method and DCFace adopt a pre-trained face recognition model which may counter legal issues. we hope further researchers bring steps forward to avoid using this kind of pre-trained face recognition model to alleviate legal concerns in this domain.

19

Method	Ours	DCFace [24]	DigiFace [26]
FID	18.72	15.82	65.39

Table 12: Fid score to the real dataset CASIA-WebFace.

Base	Euclidean	Interval 0.06
91.64	90.95	91.43

Table 13: Difference between Euclidean and larger similarity interval

D Discussion

D.1 Why Semi-hard samples work

We assume the benefits of the semi-hard training face images could be attributed to:

- easy training samples are typically images where the face is clear, well-lit, and faces the camera directly, and thus training on such easy samples would not allow the trained FR models to be able to generalize for face images with large pose/age/expression variations and different lighting conditions/backgrounds that are frequently happened in real-world applications. AdaFace [3] also mentioned that easy samples could be beneficial to early-stage training, while hard sample mining is needed for achieving generalized and effective FR models;

- Hard samples normally contain noise data. Specifically, FaceNet [29] demonstrates that the hardest sample mining using a large batch size leads to hard convergence and produces inferior performance. This is because training with very hard samples may not allow FR models to learn effective features but focus on cues apart from facial identities;

- Semi-hard samples generated by CemiFace mostly contain large posed faces but fewer face-unrelated noises. We also evaluate the training epochs needed to reach the highest AVG performance for easy samples ($m = 0.7$), semi-hard samples ($m = 0$) and extreme hard samples ($m = -0.5$). Easy samples take 10 epochs to reach the best AVG and 20 epochs to produce the training loss of 0; Semi-hard samples take much longer (38 epochs) to provide the highest AVG while the final training loss is around 3; and FR models training on extreme hard samples could not converge.

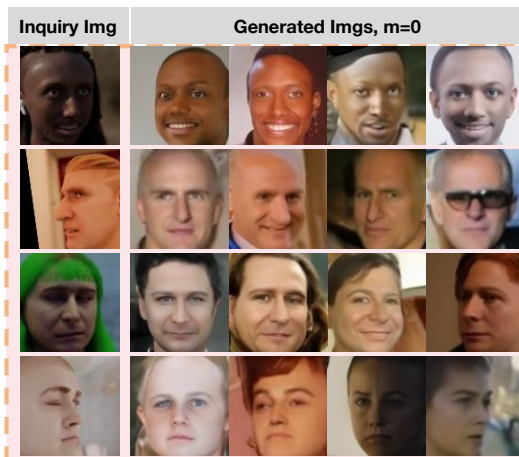



Figure 9: visualization of samples inquired by 1-shot DigiFace. Different rows are results inquired by different images. Different columns are randomly selected generated samples.

20

λ	0.01	0.05(default)	0.1	0.5
AVG	**91.72**	91.64	91.29	90.77

Table 14: Impact of different λ

The actual similarity to the inquiry center indicates that our CemiFace tends to generate images showing lower similarities to their identity centers (i.e. all samples are semi-hard), while DCFace contains more easy samples.

D.2 Different diffusion Loss

As there are some other variation diffusion losses such as Improved-DDPM [45] which has been applied in Diffusion Transformer (DIT) [46], Variational Diffusion Models (VDM)[47]. We follow the previous SOTA SFR studies (DCFace [24] and IDiffFace [28]) to choose the same generic MSE diffusion loss [21, 22] as our base model, ensuring the reproducibility of our approach and its fair comparison with DCFace [24] and IDiffFace [28].

D.3 Difference between Dataset Distillation

Dataset distillation methods [48–50] are widely adopted to create a dataset that can produce high performance when training a model on it. SRe2L [48] is a recent state-of-the-art method for dataset distillation which trains the noise image through a pretrained backbone. Their main process contains a forward process to get the classification label of the trainable noise inquiry image and train the noise inquiry image to produce a specific class prediction with BN alignment. The distinctions between our method with theirs are:

- **Embedding vs Classification Layer**: We aim to explore the feature embedding of the backbone, not the classification layer.
- **Consideration of Image Similarity**: Our method explores the similarity of the given inquiry image, which is not considered in recent dataset distillation methods.
- **Pattern Distillation**: Their approach focuses on distilling data from existing classes, while our CemiFace distils patterns from the pretrained face recognition model. This learned pattern can be applied to unseen subjects, as we utilize independent data that was not part of the pretrained model's training dataset.
- **Extra Model**: We incorporate a diffusion model to introduce parameters for producing high-quality images.

D.4 Relationship to ID3PM

Recent work, i.e. ID3PM [51] proposes to invert the Black-Box model of face recognition to generate a similar image to the inquiry image. However, our method differs from theirs in several aspects:

- **Purpose**: Their objective is to invert the black-box model without full access, whereas we aim to generate a discriminative dataset.
- **Image Similarity**: They require the generated image to be like the original image, while our goal is to ensure the generated images encompass diverse styles.
- **Evaluation Approach**: They evaluate by replacing the data of the evaluation dataset, whereas our approach involves training a model on the generated dataset.
- **Theoretical Degradation**: When $\mathbf{m}$ is set to 1, our model theoretically degrades to their model.
- **Diffusion Model Structures**: We use different diffusion model structures to conduct experiments, specifically employing cross-attention and AdaGN [42] for inserting conditions.

21

NeurIPS Paper Checklist

1. **Claims**

    Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

    Answer: [Yes]

    Justification: Contributions of this paper are included in the abstract and introduction

    Guidelines:

    - The answer NA means that the abstract and introduction do not include the claims made in the paper.
    - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
    - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
    - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

    Question: Does the paper discuss the limitations of the work performed by the authors?

    Answer: [Yes]

    Justification: We have discussed the limitation in the conclusion section

    Guidelines:

    - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
    - The authors are encouraged to create a separate "Limitations" section in their paper.
    - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
    - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
    - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
    - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
    - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
    - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

    Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

    Answer: [NA]

22

Justification: We only have experimental assumptions and they are proved in the main paper

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We will release the code and data upon acceptance. And we provide detailed information for reproducing.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

23

Answer: [No]

Justification: We will release the code and data upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All experimental details are provided

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Our experimental results show that the proposed method exceeds previous works by a large margin. And we have run the experiments multiple times to confirm the effectiveness of the proposed method.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

24

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Computational cost is included in the Supplementary Material

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have made sure the anonymity

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed the privacy issues brought by Face Recognition.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

25

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer:[Yes]

Justification: We avoid using data that is banned from being applied to Face recognition and discussed the privacy issues.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer:[Yes]

Justification: Our models and code used in the paper are licensed

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

26

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No assets

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer:[NA]

Justification: We do not collect data, but generate synthetic data

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We do not collect data, but generate synthetic data

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

27