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

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

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.

Algorithm 2 The pipeline of CemiFace-based face dataset generation

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

To provide deeper insights into this phenomenon, we visualize the samples generated by different inquiry centers in Figure 6. Notably, with $\mathbf{m}$=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.

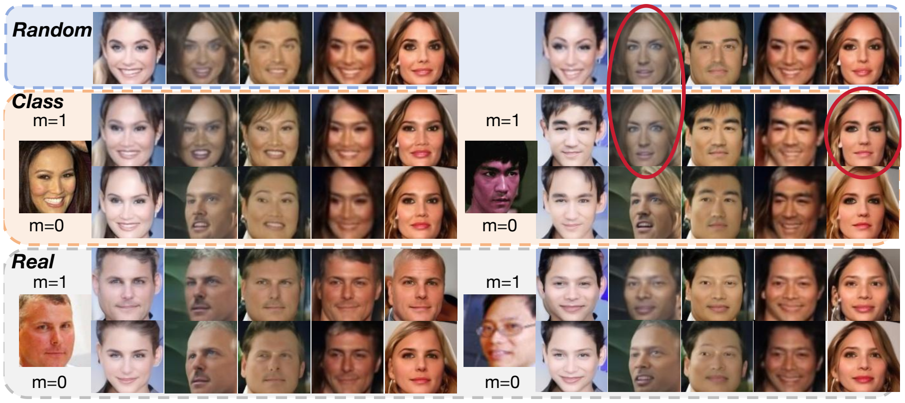

Figure 6: Comparison of different inquiry center. From top to bottom are images inquired by *Random Center*, CASIA *Identity Center* and 1-shot *Real* images. For *Identity Center* and 1-shot *Real* images, images similarity of 1 and 0 are shown. Different columns represent given different noise. Two 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 $\mathbf{m}$ 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.

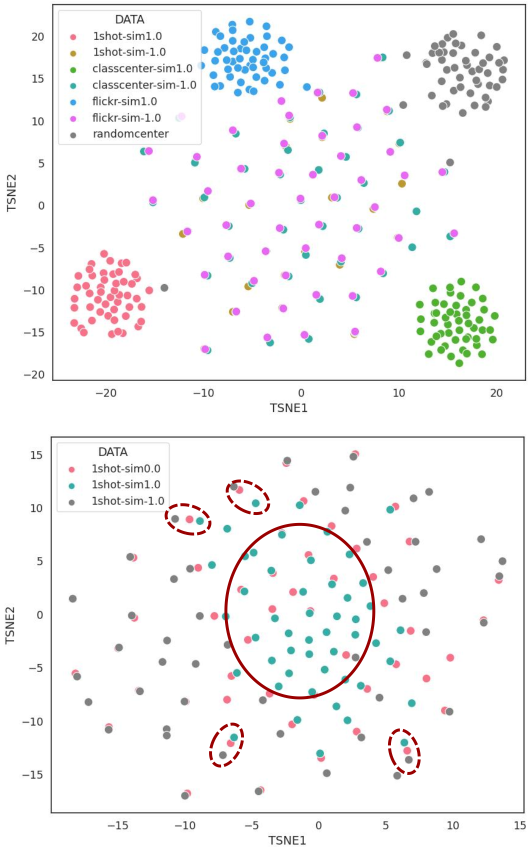

Figure 7: T-SNE visualization. The bottom figure is the T-SNE generated by 1-shot data with similarity of 1.0, 0.0 and -1.0 respectively. The upper figure is different inquiry centers with two similarities 1.0 and -1.0, the random center is also given. Red circles are samples worth noticing, with their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

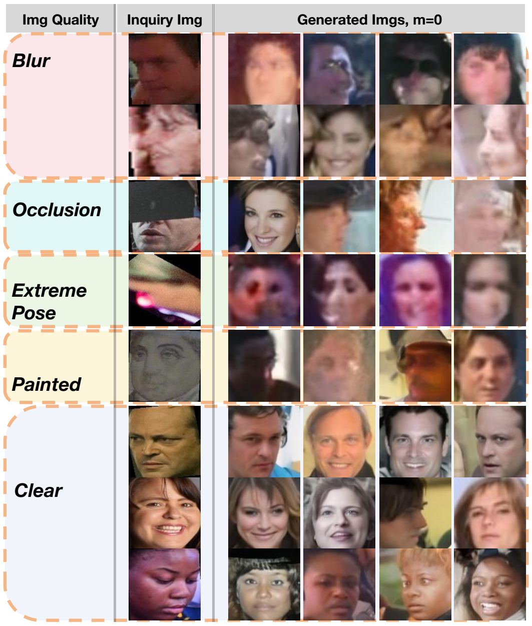

Figure 8: Examples of samples under challenging conditions, including Blur, Occlusion, Extreme Pose, and Painted conditions, are presented. Samples generated by clear images are appended for better comparison.

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.

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.

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.

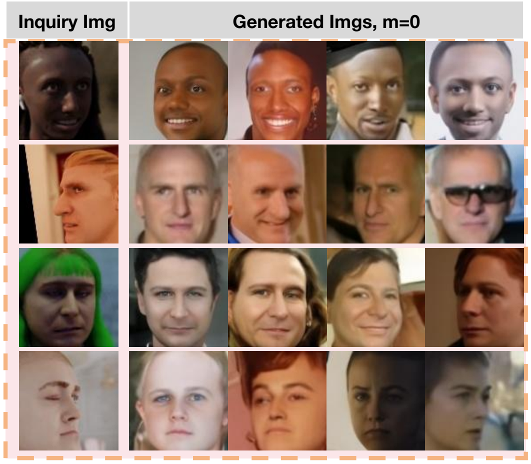

Figure 9: visualization of samples inquired by 1-shot DigiFace. Different rows are results inquired by different images. Different columns are randomly selected generated samples.

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 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.