

# VIGFace: Virtual Identity Generation for Privacy-Free Face Recognition dataset

Minsoo Kim<sup>\*1,2</sup> Min-Cheol Sagong<sup>\*1</sup> Gi Pyo Nam<sup>1,2</sup> Junghyun Cho<sup>1,2,3</sup> Ig-Jae Kim<sup>1,2</sup>

<sup>1</sup>Korea Institute of Science and Technology(KIST)

<sup>2</sup>Korea National University of Science and Technology

<sup>3</sup>Yonsei-KIST Convergence Research Institute

{kim1102, mcsagong, gpnam, jhcho, drjay}@kist.re.kr

## Abstract

*Deep learning-based face recognition continues to face challenges due to its reliance on huge datasets obtained from web crawling, which can be costly to gather and raise significant real-world privacy concerns. To address this issue, we propose **VIGFace**, a novel framework capable of generating synthetic facial images. Our idea originates from pre-assigning virtual identities in the feature space. Initially, we train the face recognition model using a real face dataset and create a feature space for both real and virtual identities, where virtual prototypes are orthogonal to other prototypes. Subsequently, we train the diffusion model based on the established feature space, enabling it to generate authentic human face images from real prototypes and synthesize virtual face images from virtual prototypes. Our proposed framework provides two significant benefits. Firstly, it shows clear separability between existing individuals and virtual face images, allowing one to create synthetic images with confidence and without concerns about privacy and portrait rights. Secondly, it ensures improved performance through data augmentation by incorporating real existing images. Extensive experiments demonstrate the superiority of our virtual face dataset and framework, outperforming the previous state-of-the-art on various face recognition benchmarks. <https://github.com/kim1102/VIGFace>*

## 1. Introduction

Deep learning-based Face Recognition (FR) models have significantly improved their performance due to recent advances in network architectures [17, 18, 21, 26, 45, 47] and enhancements in loss functions [9, 28, 32, 42, 43, 54]. The latest FR models utilize the softmax-variant loss for training to reduce intra-class variance and increase inter-class separability in the embedding space. This necessitates a very large dataset with numerous distinct individuals, signifi-

cant variations within each individual for intra-class variance, and precise labels of subject identities (IDs). However, datasets are typically collected through web-crawling and then refined using automatic techniques that employ FR logits [11, 60]. Although this approach is successful in eliminating mislabeled data, it still struggles with persistent issues of small intra-class variance. Moreover, in the face recognition field, unlike other ordinary image classifications, there are practical issues such as portrait rights, making it even more difficult to collect training data. For example, data sets such as those referenced in [16, 56, 60] consist of images of celebrities collected from the internet without consent. Furthermore, the datasets mentioned in [27, 38] include facial images of the general population, including children. The privacy sensitivity of such data poses significant challenges for face recognition research.

Synthetic datasets have been used to address the limitations caused by the scarcity of real datasets [8, 25, 49] and biases present in the available real datasets [20, 50]. In the realm of face recognition, artificial faces show promise in addressing the aforementioned issues related to real face datasets. Generated faces are at low risk for label noise due to conditional generation. Bias problems, such as long-tailed distributions, which lead to class imbalances, can be mitigated by data augmentation. Importantly, there are no privacy concerns with virtual identities facial images if the method is effective. Therefore, when creating artificial datasets, it is essential to consider: 1) ensuring that the generated data mirror the real data distribution, 2) the ability to generate new subjects separated from real data, and 3) maintaining ID consistency for each subject.

Previous attempts to create artificial face datasets have addressed some of the three aspects individually, but to the best of our knowledge, none have simultaneously taken into account all three aspects [2, 5, 29, 40]. SynFace [40] introduces a high-quality synthetic face image dataset, closely reminiscent of real face images, using DiscoFaceGAN [13]. However, DiscoFaceGAN is only capable of generating a limited number of unique subjects, fewer than 500 [29]. On

\*These authors contributed equally to this work

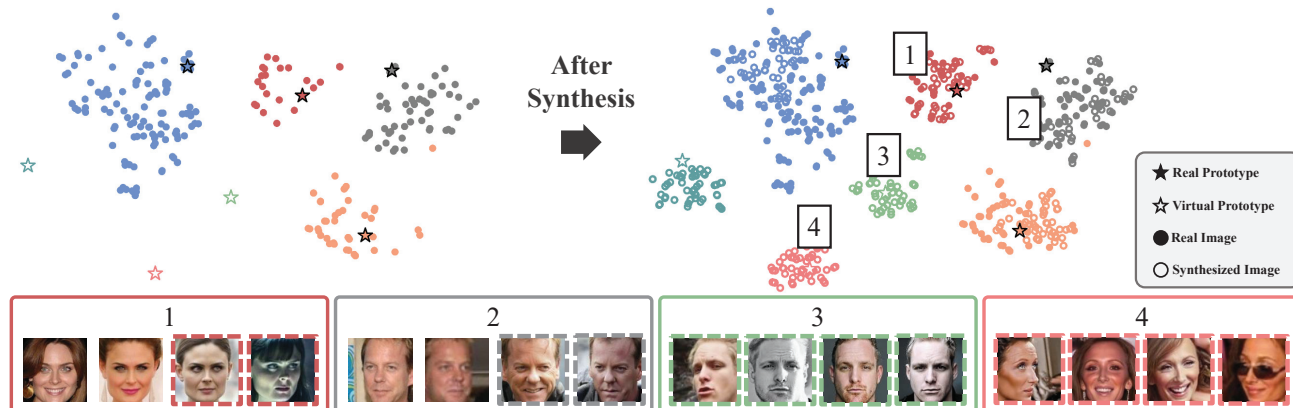


Figure 1. T-distributed Stochastic Neighbor Embedding (T-SNE) [51] plot of embeddings from real and synthesized images. The filled and lined stars represent the real and virtual prototypes, while filled and lined circles indicate the embeddings of real and synthesized images, respectively. The bottom of the figure shows the face images included in the cluster, and the dotted outlined images represent the face images generated using our method.

the other hand, DigiFace [2] utilizes 3D parametric modeling to create synthetic face images of various subjects. However, it faces difficulties in accurately replicating the quality and style distribution observed in real face images. DCFace [29] proposes a diffusion-based method to generate data that maintains the style of real datasets and ensures label consistency. However, DCFace primarily focuses on creating synthetic data rather than enhancing the capabilities of data augmentation. IDiffFace [5] utilizes identity-conditioned latent diffusion to synthesize facial images from FR feature representations. However, it fails to guarantee the uniqueness of the synthesized identities. CemiFace [48] and HSFace [55] demonstrate high performance on the various FR benchmarks. However, we found identity leakage of the training dataset from both CemiFace and HSFace, which is critical from a privacy standpoint.

The virtual data generated by our method possesses all three essential properties of virtual data mentioned above. The main concept of our paper is to incorporate virtual prototypes into the FR model. Virtual prototypes are trained simultaneously with real prototypes so that they pre-assign on the same feature space. The diffusion model is trained to preserve the identity of the individuals when generating face images based on the FR embedding. As virtual prototypes are allocated within the same embedding space as real prototypes, the produced face images are ensured to reflect the true data distribution. In addition, since prototypes are orthogonal to others, virtual subjects are guaranteed to be distinguished from the original data. Fig. 1 shows a toy example that visualizes the embeddings of real and virtual subjects to demonstrate the effectiveness of the suggested method. The virtual embedding can be distinct from the real individual clusters, whereas the images generated from virtual prototypes form unique clusters. The visualization

demonstrates that our framework generates distinctive virtual human faces with high consistency in ID. The proposed method effectively addresses privacy concerns by creating datasets of non-existent individuals and achieving state-of-the-art performance compared to models trained with previous virtual face generation methods. Additionally, the FR model, which was trained using a combination of real and synthetic images together, achieves better performance than the model trained using only real images. This shows that the proposed method is also superior from the perspective of a data augmentation method.

Our contribution can be summarized as follows:

- Introducing VIGFace, a method designed to generate virtual identities with realistic appearance, guided by three main criteria: generating novel subjects that are not found in the real-world, consistently preserving the identity, but ensuring diverse characteristics for each individual.
- Showing that the synthetic data generated by VIGFace can leverage intra-class diversity and inter-class variance, achieving SOTA performance in face recognition.
- Releasing the virtual-only face dataset that can fully substitute the real dataset, helping alleviate privacy concerns.

## 2. Related work

### 2.1. Face Recognition Models

Current state-of-the-arts (SOTA) FR methods, such as [3, 9, 12, 24, 28, 32, 36, 53], are designed based on softmax loss, but particularly utilize the angular/cosine distance instead of the Euclidean distance. The goal of these methods is to maximize the similarity between the embeddings and the ground-truth (GT) prototype, while minimizing the similarity between the embedding and prototypes of other classes. As a result, all embeddings in the same class, including prototypes, converge while maintaining their maxi-

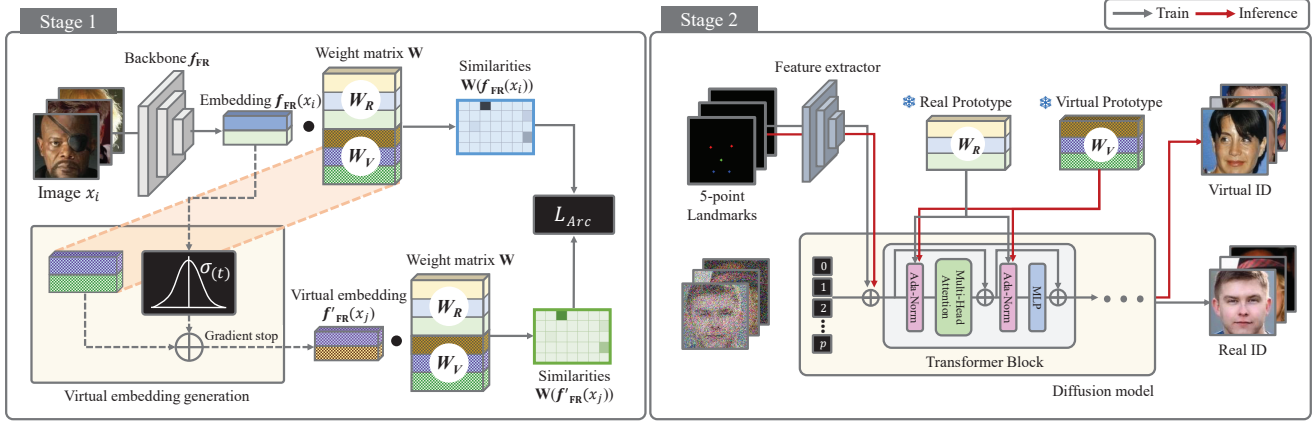


Figure 2. Pipeline for the proposed method. Conventional FR training includes prototypes for only real individuals, indicated as  $W_R$ . We add  $k$  prototypes for virtual IDs, denoted as  $W_V$ . The virtual embedding  $f'_{FR}(x_j)$  corresponding to the virtual person ID:  $j$  is generated to follow distribution of the real embeddings. To synthesize the facial image from virtual prototypes, we adopt the DiT architecture [39], following the design approach of the Vision Transformer (ViT) [14]. Additionally, we adjust the DiT model to utilize 5-point landmark images to handle pose variations.

mum distance from other classes. Therefore, if the embedding dimension is large enough, the clusters of each class will be nearly orthogonal [6, 9]. Naturally, a larger dataset that contains a greater variety of images allows for the training of better-performing FR models. Therefore, over time, larger and more diverse datasets [1, 16, 22, 60] have been collected and published in the academic field. However, issues such as portrait rights and the high cost of building large-scale data remain unresolved.

## 2.2. Synthetic Face Image Generation

SynFace [40] proposed a method that uses DiscoFaceGAN[13] to synthesize virtual faces. DigiFace-1M [2] proposes a 3D model-based face rendering method to generate virtual face data. DCFace [29] proposes a diffusion-based face generator combining subject appearance (ID) and external factor (style) conditions. GanDiffFace [35] uses GANs to generate identity features. Vec2Face [55] introduces a masked autoencoder to control the identity of face images and their attributes. CemiFace [48] produces facial samples with various levels of similarity to the center of the subject. IDiffFace [5] uses latent diffusion conditioned on FR identity to generate synthetic images. Although these methods succeed in generating separable identities, some still require additional networks or thresholding to sample synthetic identities. Furthermore, unlike [5], we propose a novel method for the pre-assignment of FR representations to maintain latent space alignment and introduce pose control of facial images using facial landmarks.

## 3. Methods

Our framework comprises two stages. First, we train the FR model using the real face dataset and design the feature

space for both real and virtual IDs. Second, synthetic images are generated using the diffusion model based on the feature space of the pre-trained FR model that was trained in the first stage. This section provides a detailed explanation of each stage of the proposed framework.

### 3.1. Stage 1: FR Model Training

The proposed framework begins with training the FR model using real face images. This stage serves two purposes: 1) Training the FR model to achieve prototype features from face images, which are necessary for training the diffusion model in the second stage, and 2) Simultaneously assigning the positions of both the real ID and the virtual ID on the feature space. We choose ArcFace [9] to train the FR model in this stage. The ArcFace loss used to train the FR backbone  $f_{FR}$  and the prototype  $W = [w_1, w_2, \dots, w_n]$  can be described as follows:

$$L_{arc} = -\log \frac{e^{s \cos(\theta_{y_i+m})}}{e^{s \cos(\theta_{y_i+m})} + \sum_{j=1, j \neq y_i}^n e^{s \cos \theta_j}}, \quad (1)$$

$$\cos \theta_j = \frac{w_j^T f_{FR}(x_i)}{\|w_j\| \|f_{FR}(x_i)\|}, \quad (2)$$

where  $n$  denotes the total number of real IDs, while  $m$  and  $s$  represent the margin and scale hyperparameters, respectively. With conventional FR training methods, only prototypes for real individuals, which can be denoted as  $W_R = [w_r^1, w_r^2, \dots, w_r^n]$ , are necessary for training. In contrast, we include additional  $k$  prototypes for virtual IDs, denoted as  $W_V = [w_v^1, w_v^2, \dots, w_v^k]$ , which are used to generate facial images of nonexistent individuals in the real world. As a result, the prototype is defined as a linear transformation matrix  $W \in \mathbb{R}^{(n+k) \times D}$ , where  $D$  refers to the dimension of the embedding features.

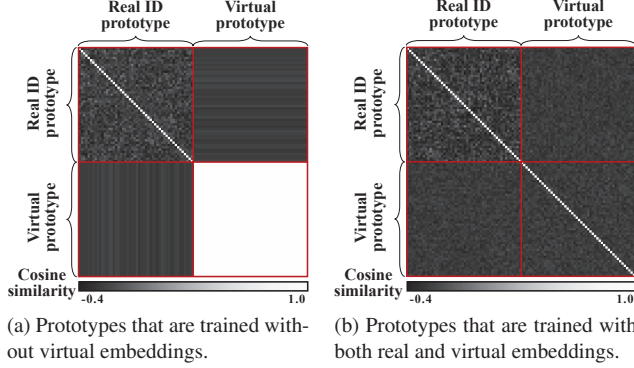


Figure 3. Changes in the similarity matrix of the prototypes from our method. Similarity values were min-max normalized.

However, due to the absence of a face image for virtual IDs, their prototypes cannot be updated to maximize the distance from each other with  $L_{arc}$ . Consequently, all virtual prototypes converge to a single point in the feature space when trained with the original  $L_{arc}$  as shown in Fig. 3a. To address this problem, we propose to use virtual feature embedding  $f'_{FR}(x_j)$  to update the virtual prototypes  $w'_v$ . The virtual embedding  $f'_{FR}(x_j)$  corresponding to the virtual person ID:  $j$  was generated as follows:

$$f'_{FR}(x_j) = w'_v + \mathcal{N}(0, 1) \cdot \sigma, \quad (3)$$

$$\sigma^2 = \frac{1}{b} \sum_{i=1}^b (f_{FR}(x_i) - w'_r)^2, \quad (4)$$

where  $b$  refers the mini-batch size. As can be seen from the equations, the virtual embedding  $f'_{FR}(x_j)$  follows a distribution in which the standard deviation matches that of the real embeddings. The aggregate loss  $L_{arc}$  is calculated simultaneously using both the real embedding  $f_{FR}(x)$  and the virtual embedding  $f'_{FR}(x)$ , allowing the virtual prototype  $w'_v$  to maintain minimal similarities with other prototypes, as illustrated in Fig. 3. Since batch configuration affects the calculated standard deviation, we utilize the exponential moving average (EMA) to reduce this influence. The corrected standard deviation  $\sigma$  for the current iteration  $t$  is calculated as follows:

$$\sigma = \sigma_{(t)} \cdot \alpha + \sigma_{(t-1)} \cdot (1 - \alpha). \quad (5)$$

The hyperparameter  $\alpha$  is set to 0.9. The number of virtual embedding,  $b_v$  is determined based on the mini-batch size  $b_r$ , the number of virtual ID prototypes  $k$ , and the number of real ID prototypes  $n$ . In our study, we set  $b_v = (k \times b_r) / n$  so that both real and virtual prototypes can be updated evenly.

The overall pipeline of this stage is illustrated in Fig. 2. Note that, as can be seen in the figure, gradients for virtual embeddings have no effect on the backbone. We confirm

that the similarity distribution between the virtual prototypes closely resembles the similarity distribution between the real prototypes, as shown in Fig. 3b.

### 3.2. Stage 2: Face Generation with Diffusion Model

The next step is to synthesize face images using the diffusion model. To obtain the training dataset for the diffusion model, we utilize the FR model, which involves collecting pairs of images  $x_0$  and their corresponding prototypes  $w_r$ . The input to our diffusion model includes the timestep  $t$ , the FR prototype vector  $w_r$ , the five facial landmark image  $y$ , and the noisy image  $x_t$ . In line with the approach proposed in the previous method [31], our model predicts velocity  $v_t$  rather than noise  $\epsilon$  injected into  $x_t$ . We adopt the DiT architecture [39] using the design approach of the visual transformer (ViT) [14]. We modify the DiT model by incorporating the five facial landmarks (including the left eye, the right eye, the nose, the left mouth corner, and the right mouth corner) [57] image. The five facial landmark images are acquired using RetinaFace [10], and employed as conditions to account for pose variations. Fig. 4 shows the synthesized face images conditioned by five facial landmarks. The first column lists input landmark conditions in different pose environments. As illustrated in the figure, the proposed model is capable of producing a wide range of styles (like hair, glasses, and lighting) while also delivering various pose variations, all with a high degree of consistency. In order to achieve both the generation of facial images and the synchronization of their feature space on the FR model, our diffusion model incorporates a constraint which aims to minimize the feature distance between the original image and the input prototypes as follows:

$$\min_{\theta} \mathbb{E}_{\epsilon, t} \|f_{FR}(\hat{x}_{\theta}(x_t, t, w_r, y)) - w_r\|_2^2. \quad (6)$$

We adopt classifier-free guidance [19] by randomly assigning zero values to condition embeddings,  $w_r$ , 10% of time. Sampling is performed as follows:

$$\tilde{x}_{\theta}(x_t, t, W, y) = g \cdot x_{\theta}(x_t, t, W, y) + (1 - g) \cdot x_{\theta}(x_t, t, y), \quad (7)$$

where  $x_{\theta}(x_t, t, w_r, y)$  and  $x_{\theta}(x_t, t, y)$  are conditional and unconditional  $x_0$ -prediction, respectively and  $g$  is *guidance weight*.

## 4. Experiments

Our implementation details for the full framework are described in Appendix A. We analyze our framework from two perspectives. Firstly, by training the FR model exclusively with facial images generated for virtual IDs, we

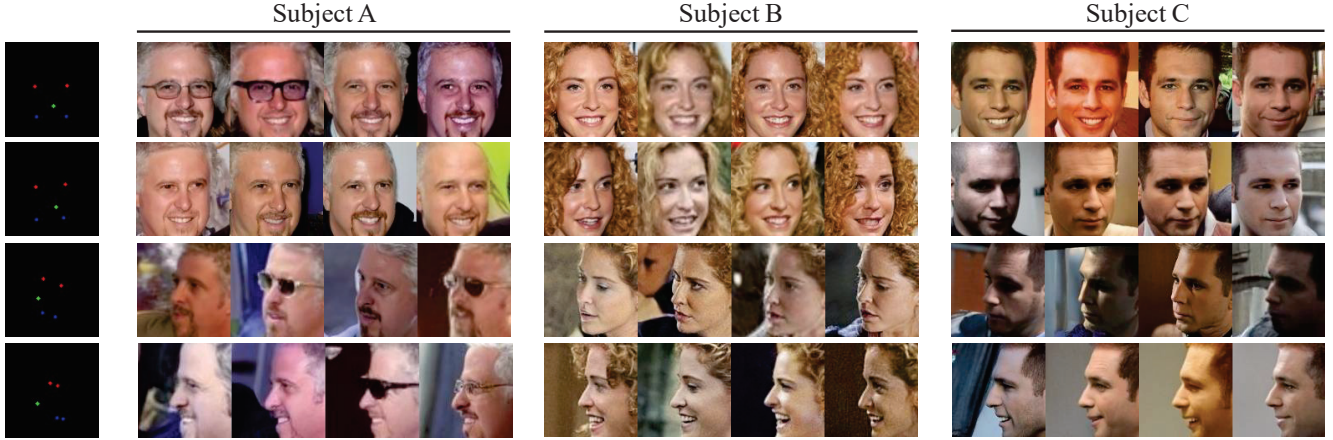


Figure 4. Virtual face generated in VIGFace (B). Each row lists facial images in different pose environments created using five facial landmarks. Our method can generate various conditions of face images, such as illumination, occlusion by accessories, and facial expressions, while controlling the pose variations of the face images.

demonstrated the capability of our model to serve as a viable alternative to real face datasets, addressing concerns such as label noise or privacy. Secondly, we train the FR model using both real images and generated face images simultaneously, showcasing the potential of our model as a data augmentation framework for face recognition.

#### 4.1. Virtual Identity Generation

Since diffusion models focus on optimizing the visual aspect with input conditions, they tend to produce images that are high in consistency but lack of variance. One simple way to achieve a balance between consistency and variance is by adjusting the classifier-free guidance weight scale. Higher guidance enforces stronger conditioning of input labels. In other words, it results in more constant images but slightly similar samples. We observe that the proposed model performs best as the scale becomes  $w = 4.0$ .

We construct a toy example visualizing the feature space consists of three real persons and five virtual persons to demonstrate the efficacy of the synthesized images in Fig. 1. As seen in the figure on the left, the virtual embedding optimization obviously provides separable virtual prototypes from real individual clusters. The figure on the right illustrates that our synthesized images enforce high variance to the dataset. Synthesized images of real individuals filling the gap in the cluster provide intra-class variance to the subjects. Additionally, the cluster of generated images from virtual prototypes, while staying separate from other subjects, demonstrates the effectiveness of our framework for generating unique virtual human faces in high consistency.

Fig. 5 shows the qualitative comparison of conventional synthetic datasets and the proposed VIGFace. As shown in the figure, the facial images generated using the VIGFace model show remarkable uniformity in generating consistent

virtual individuals, while also incorporating variations such as hair styles, accessories, makeup, and expressions. It is important to emphasize that the key characteristic of a training dataset to achieve a high-performance FR model is not the high quality of the images themselves, but rather the high consistency and variance of the generated facial images belonging to the same person.

**Property of generated Dataset** To compare the properties of the synthetic face datasets, we measured the 1) class consistency, 2) class separability, and 3) intra-class diversity of generated images. The class consistency reflects the uniformity of the samples in the same label condition. Consequently, the consistency of class  $k$  ( $C_k$ ) was measured as follows:

$$C_k = \frac{1}{N^2} \sum_{j=1}^N \sum_{i=1}^N \frac{f(x_i) \cdot f(x_j)}{\|f(x_i)\| \|f(x_j)\|}, \quad (8)$$

where  $N$  is the number of images in a class  $k$ . Higher class consistency means that the samples are more uniform under the same label.

We also measured the class separability to assess the integrity of the dataset; in other words, to ensure that all subjects in the dataset are unique. The class separability for a class  $k$  ( $S_k$ ) is measured as the average distance between the center of class  $k$  and the centers of the negative classes, as follows:

$$S_k = \frac{1}{K-1} \sum_{i=1, i \neq k}^K 1 - \frac{\bar{f}_k \cdot \bar{f}_i}{\|\bar{f}_k\| \|\bar{f}_i\|}, \quad (9)$$

where  $K$  is the number of classes.  $\bar{f}_k$  represents the class center obtained by averaging the embedding vectors of the

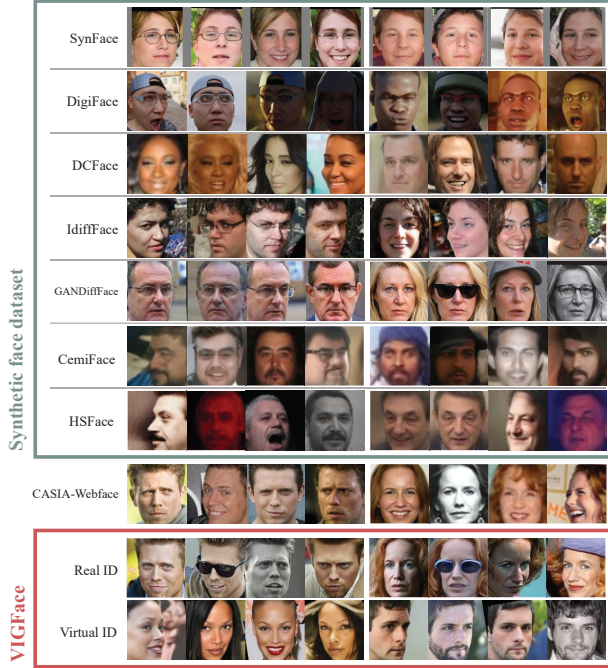


Figure 5. Comparison of randomly selected virtual ID images generated by conventional methods [2, 5, 29, 35, 40, 48, 55] and by our method, all trained on CASIA-WebFace. For each synthetic dataset, we present two subjects in a single row.

images that belong to class  $k$ . As shown in the formula, a high  $S_k$  indicates that the images are distinct from the negative classes.

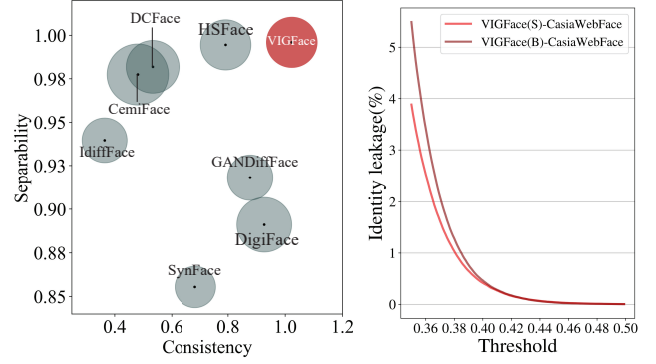
The intra-class diversity measures how various the conditions of samples are under the same label. In particular, we focus on challenging scenarios that directly impact FR performance, such as pose variations, occlusions, or lighting conditions, rather than image style. Higher diversity indicates that the dataset covers a broad spectrum of cases, from easy to hard. Motivated by this observation, we calculated the intra-class diversity of a class  $k$  ( $D_k$ ) using the variance of Face Image Quality Assessment (FIQA) scores as follows:

$$\overline{\text{FIQA}}_k = \frac{1}{N} \sum_{i=1}^N \text{FIQA}(x_i), \quad (10)$$

$$D_k = \frac{1}{N} \sum_{i=1}^N (\text{FIQA}(x_i) - \overline{\text{FIQA}}_k)^2, \quad (11)$$

where  $\text{FIQA}(x_i)$  indicates the normalized CR-FIQA [4] score of the generated image  $x_i$ .

The scores  $C_k$  and  $S_k$  are derived from a pre-trained ArcFace model that was trained using the Glint-360K dataset. Fig. 6a shows the average values of the properties for all classes, normalized by the average values achieved with CASIA-WebFace. VIGFace achieves remarkable scores in



(a) The normalized properties of various synthetic face dataset and its circles indicate the intra-class diversity. (b) Identity leakage test on the synthetic face dataset and its corresponding training dataset.

Figure 6. Properties of VIGFace compared with those of synthetic datasets generated by previous methods.

aspects of class consistency and class separability compared to other methods. SynFace exhibits the lowest separability due to its mix-up generation method. This fact indicates that SynFace can generate only a limited number of subjects as mentioned in the introduction. DigiFace achieves strong consistency, due to its 3D rendering methods, but its image style may differ significantly from real datasets, leading to lower accuracy in real-world applications. GANDiffFace demonstrates inadequate FIQA diversity, indicating limited variations in factors such as pose, lighting, and occlusion. DCFace, which uses a diffusion model, shows separability similar to that of VIGFace but with lower consistency.

In comparisons of SOTA models, HSFace, despite using a more refined dataset, comparatively lacks consistency. CemifFace, trained on the same dataset as VIGFace, shows low consistency and separability. Both methods also exhibit inferior qualitative uniformity within the same class, as illustrated in Fig. 5. Our approach uses prototypes instead of FR embedding vectors to disentangle identity features from other characteristics, allowing VIGFace to generate subjects with greater consistency by focusing on identity features. Moreover, it enhances separability by leveraging the pre-computation of feature orthogonality. We included further analysis on the properties of the datasets and provided examples in Appendix H.

**Identity leakage from real human** To claim complete privacy-free, it is necessary to prove that no real IDs or training images are included in the synthesized dataset. In this reason, we demonstrate that the generated face images represent non-existent humans by querying the most similar face in the CASIA-WebFace dataset. For comparison, we also present the similarity values of the nearest negative class in CASIA-WebFace itself. In Fig. 6b, we count the number of synthetic face images whose similarity with

Method	Training Dataset	# of Images (classes $\times$ variations)	Verification Benchmarks					
			LFW	CFP-FP	CPLFW	AgeDB	CALFW	Avg.
CASIA-Webface (Real)	-	0.49M ( $\approx 10.5K \times 47$ )	99.40	96.63	90.23	94.68	93.70	94.93
SynFace	FFHQ	0.5M ( $10K \times 50$ )	91.93	75.03	70.43	61.63	74.73	74.75
DigiFace	3D modeling	0.5M ( $10K \times 50$ )	95.40	87.40	78.87	76.97	78.62	83.45
DCFace	FFHQ+CASIA	0.5M ( $10K \times 50$ )	98.55	85.33	82.62	89.70	91.60	89.56
IDiffFace	FFHQ	0.5M ( $10K \times 50$ )	98.00	85.47	80.45	86.43	90.65	88.20
GANDiffFace	FFHQ	0.5M ( $10K \times 50$ )	90.77	73.27	72.32	66.35	74.68	75.48
ID <sup>3</sup>	CASIA	0.5M ( $10K \times 50$ )	97.68	86.84	82.77	91.00	90.73	89.80
CemiFace	CASIA+WebFace4M	0.5M ( $10K \times 50$ )	<b>99.03</b>	91.06	87.62	91.33	92.42	<u>92.30</u>
Arc2Face	WebFace42M	0.5M ( $10K \times 50$ )	98.81	<u>91.87</u>	85.16	90.18	<u>92.63</u>	91.73
HyperFace	WebFace42M	0.5M ( $10K \times 50$ )	98.50	88.83	84.23	86.53	89.40	89.50
HSFace10K	WebFace4M	0.5M ( $10K \times 50$ )	98.87	88.97	85.47	<b>93.12</b>	<b>93.57</b>	92.00
VIGFace(S), Ours	CASIA	0.5M ( $10K \times 50$ )	<u>99.02</u>	<b>95.09</b>	<b>87.72</b>	90.95	90.00	<b>92.56</b>
DCFace	FFHQ+CASIA	1.2M ( $20K \times 50 + 40K \times 5$ )	98.58	88.61	85.07	90.97	92.82	91.21
CemiFace	CASIA+WebFace4M	1.2M ( $20K \times 50 + 40K \times 5$ )	99.22	92.84	88.86	92.13	93.03	93.22
Arc2Face	WebFace42M	1.2M ( $20K \times 50 + 40K \times 5$ )	98.92	94.58	86.45	92.45	93.33	93.14
HSFace20K	WebFace4M	1.0M ( $20K \times 50$ )	98.87	89.87	86.13	<u>93.85</u>	<u>93.65</u>	92.47
HSFace300K	WebFace4M	15M ( $300K \times 50$ )	99.30	91.54	87.70	<b>94.45</b>	<b>94.58</b>	93.52
VIGFace(B), Ours	CASIA	1.2M ( $60K \times 20$ )	<b>99.48</b>	<u>97.07</u>	90.15	93.62	92.88	<u>94.64</u>
VIGFace(L), Ours	CASIA	6.0M ( $120K \times 50$ )	<u>99.33</u>	<b>97.31</b>	<b>91.12</b>	93.82	92.95	<b>94.91</b>

Table 1. FR benchmark results trained with various virtual face datasets. All results except for CASIA-WebFace and VIGFace are obtained from the original paper. Our method is specified by the number of identities as small (S), base (B) and large (L). FR backbone is IR-SE50 + AdaFace [28]. **Bold** and underline indicates the best and the second best, respectively.

the average feature vector of each class in the training data exceeds a given threshold. As shown in the figure, face images in VIGFace have low similarity to the subjects in CASIA-WebFace. This supports the claim that each synthesized image depicts a non-existent human face and there is no identity leakage from the training data. In contrast, we observed that some of the conventional SOTA methods exhibit identity leakage. As CemiFace [48] samples identity embeddings from the WebFace4M dataset, it generates individuals that exhibit resemblance to those in WebFace4M. Vec2Face [55] fails in creating novel identities, so that it produces almost same persons in WebFace4M. Further analysis and examples are reported in Appendix I. In this paper, we avoid thresholding or selective sampling, which could harm fair comparisons with previous methods.

## 4.2. Evaluation

In this section, we train the FR model with VIGFace to compare it to conventional methods. Detailed hyperparameters for training the FR network can be found in Appendix A.

**VIGFace as Virtual Dataset** We compare the performance of the FR model trained with generated facial images for virtual IDs using VIGFace with conventional methods. For the experiment, we set the number of virtual IDs to 10K, 60K and 120K. To ensure a fair comparison, we established image counts of 0.5M, 1.2M, and 6.0M, which reflects the scale of the real CASIA-WebFace dataset and previous synthetic approaches, respectively. In Tab. 1, we present the 1:1 verification accuracy (%) on five bench-

marks [23, 37, 44, 58, 59].

As shown in the table, VIGFace outperforms conventional methods in the average accuracy of five verification benchmarks. In particular, VIGFace achieves equal or even better performance on CFP-FP and CPLFW compared to the model trained with the real dataset, CASIA-WebFace. This indicates that our method benefits from its pose-variable generation capabilities. CemiFace shows notable performance on the AgeDB and CALFW benchmarks, but they sampled identity embeddings from the WebFace4M [60] dataset. Consequently, the individuals included in the CemiFace dataset may still have privacy concerns. HSFace10K also delivers outstanding results among conventional methods. However, it is important to note that HSFace benefits from training on the extensive WebFace4M dataset, which boasts twice as many images and quadruples the number of identities when compared to CASIA-WebFace. VIGFace(L) outperforms HSFace300K even with a smaller number of images. As a result, when the FR model was trained using VIGFace, it achieved performance comparable to a model trained using a real dataset in terms of average accuracy without any external identity sampling. This suggests that our virtual dataset can serve as a full replacement for the CASIA-WebFace dataset to train the FR model, while avoiding privacy concerns.

**VIGFace as Data Augmentation** To demonstrate the efficacy of VIGFace as an augmentation framework, we evaluate the accuracy of the FR model trained on real and synthetic images. We utilized the dataset that was uploaded

Condition				Verification Benchmarks						
Method	Real Image	Synthetic Image		LFW	CFP-FP	CPLFW	AgeDB	CALFW	Avg.	$\Delta$
		Real ID	Virtual ID							
CASIA-WebFace	✓			99.40	96.63	90.23	94.68	93.70	94.93	-
DigiFace	✓		✓	99.37	97.51	90.92	94.95	93.77	95.30	+0.37
DCFace (1.2M)	✓		✓	99.43	96.97	90.33	95.20	94.38	95.26	+0.33
iDiffFace	✓		✓	99.58	97.04	90.40	94.78	94.00	95.16	+0.23
GANDiffFace	✓		✓	99.52	96.61	90.30	93.98	93.57	94.80	-0.13
HSFace20K	✓		✓	99.60	97.41	91.07	95.48	<b>94.40</b>	95.59	+0.66
VIGFace(B)	✓	✓		99.45	97.23	90.78	95.25	93.22	95.19	+0.26
	✓		✓	99.55	98.03	<b>91.80</b>	95.73	94.12	95.85	+0.92
	✓	✓	✓	<b>99.70</b>	<b>98.10</b>	91.57	<b>95.85</b>	94.38	<b>95.92</b>	<b>+0.99</b>

Table 2. Augmented FR performance results for various condition of synthetic dataset. The accuracies (%) for LFW, CFP-FP, CPLFW, AgeDB-30, CALFW, and the average benchmark accuracies are reported. FR backbone is IR-SE50 + AdaFace [28].

to the official repository. Tab. 2 shows the performance change of the trained FR model using various combinations of datasets. DigiFace, which utilizes unique 3D modeling technology, has demonstrated improved benchmark performance and offers advantages from a data-augmentation perspective. However, DigiFace struggles to perform well on its own, devoid of real data [2]. This is due to their failure to precisely replicate the appearance of real-world facial images, making it impractical to present a privacy-free synthetic dataset. HSFace shows a notable improvement among conventional methods. However, it utilized WebFace4M for training, which may make the comparison with others unfair. The boosted accuracy using VIGFace, which outperforms conventional methods, demonstrates that VIGFace can accurately mirror real facial data and achieve synergy by blending them. In particular, augmented images of real IDs improve the results of the CFP-FP and CPLFW benchmarks. This observation suggests that the use of five-point facial landmarks in the conditioning method can create a variety of posed facial images, substantially improving the FR model’s ability to understand pose variations. Consequently, VIGFace not only effectively solves the privacy issue, but can also be used with real data as part of the augmentation process. When trained with the VIGFace dataset, the FR backbone achieves better generalization and higher performance, in contrast to conventional methods that achieved only marginal performance improvements.

Fig. 5 illustrates the results that synthesize variations of real and virtual IDs. Since the training process is significantly biased not just by a lack of subject but also by inner-class variance, increasing the variance within the class is crucial. The results demonstrate that the proposed approach can generate a variety of conditions for the existing real ID without any help from external data. Since VIGFace can generate varied images of real individuals, we performed experiments on each type of augmentation, *i.e.* synthetic images of real and virtual ID. In particular, we extend the long-tailed real-ID class, which contains fewer than 50 im-

ages, to 50 images. With this method, additional 0.15M images were added as real ID synthesis. As can be seen in Tab. 2, the augmented dataset for real ID shows improved FR performance on most benchmarks. This indicates that our method can increase the intra-class diversity of dataset, a critical factor in achieving high FR performance. As a result, the FR model that utilizes the entire set of augmented data exhibits the best performance, benefiting from enhanced intra-class diversity and inter-class variance.

## 5. Conclusions

This paper presents an effective method for creating a synthetic face dataset that guarantees unique virtual identities. Synthesized facial data can serve as a solution for various challenges faced by traditional real datasets, such as high expenses, inaccuracies and biases in labeling, and concerns regarding privacy. To this end, we propose the Virtual Identity Generation framework and demonstrate that it can generate not only realistic but also diverse facial images of virtual individuals, significantly narrowing the performance gap with FR models trained on real data. Furthermore, the model exhibits superior performance when trained on a combination of VIGFace and existing real data compared to models trained solely on real data. This confirms that our proposed method has potential as an effective augmentation technique. We will publicly distribute the virtual face dataset created by VIGFace and believe that these new virtual data will contribute to resolving the privacy issues inherent in face recognition training dataset.

**Acknowledgment** This work was partially supported by the Institutional Programs of the Korea Institute of Science and Technology (KIST) (Project No. 2E33612, 50%) and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2023-00227592, Development of Robust 3D Object Identification Technology for Viewpoint Changes, 50%)

## References

- [1] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1445–1449, 2021. 3, 2
- [2] 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. 1, 2, 3, 6, 8, 4
- [3] 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. 2
- [4] Fadi Boutros, Meiling Fang, Marcel Klemm, Biying Fu, and Naser Damer. Cr-fiq: face image quality assessment by learning sample relative classifiability. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5836–5845, 2023. 6, 3
- [5] Fadi Boutros, Jonas Henry Grebe, Arjan Kuijper, and Naser Damer. Idiff-face: Synthetic-based face recognition through fuzzy identity-conditioned diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19650–19661, 2023. 1, 2, 3, 6, 4
- [6] Johann S Brauchart, Alexander B Reznikov, Edward B Saff, Ian H Sloan, Yu Guang Wang, and Robert S Womersley. Random point sets on the sphere—hole radii, covering, and separation. *Experimental Mathematics*, 27(1):62–81, 2018. 3, 2
- [7] Ting Chen. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*, 2023. 1
- [8] Hyunwoo Cho, Haesol Park, Ig-Jae Kim, and Junghyun Cho. Data augmentation of backscatter x-ray images for deep learning-based automatic cargo inspection. *Sensors*, 2021. 1
- [9] 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. 1, 2, 3
- [10] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019. 4
- [11] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 741–757. Springer, 2020. 1
- [12] 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. 2
- [13] 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. 1, 3
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020. *arXiv preprint arXiv:2010.11929*, 2010. 3, 4
- [15] Papantoniou et al. Arc2face: A foundation model for id-consistent human faces. In *ECCV*, 2024. 1
- [16] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016. 1, 3
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 1
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1
- [19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 4, 1
- [20] Je Hyeong Hong, Hanjo Kim, Minsoo Kim, Gi Pyo Nam, Junghyun Cho, Hyeong-Seok Ko, and Ig-Jae Kim. A 3d model-based approach for fitting masks to faces in the wild. In *2021 IEEE International Conference on Image Processing (ICIP)*, 2021. 1
- [21] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 1
- [22] Gary Huang, Marwan Mattar, Honglak Lee, and Erik Learned-Miller. Learning to align from scratch. *Advances in neural information processing systems*, 25, 2012. 3, 1
- [23] 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. 7
- [24] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020. 2
- [25] Hochul Hwang, Cheongjae Jang, Geonwoo Park, Junghyun Cho, and Ig-Jae Kim. Eldersim: A synthetic data generation

- platform for human action recognition in eldercare applications. *IEEE Access*, 2023. 1
- [26] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 1
- [27] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4873–4882, 2016. 1
- [28] 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. 1, 2, 7, 8
- [29] Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. Dc-face: Synthetic face generation with dual condition diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12715–12725, 2023. 1, 2, 3, 6, 4
- [30] D Kinga, Jimmy Ba Adam, et al. A method for stochastic optimization. In *International conference on learning representations (ICLR)*, page 6. San Diego, California, 2015. 1
- [31] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5404–5411, 2024. 4, 1
- [32] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphreface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 1, 2
- [33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1
- [34] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 international conference on biometrics (ICB)*, pages 158–165. IEEE, 2018. 2
- [35] Pietro Melzi, Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, Dominik Lawatsch, Florian Domin, and Maxim Schaubert. Gandifface: Controllable generation of synthetic datasets for face recognition with realistic variations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3086–3095, 2023. 3, 6, 4
- [36] 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. 2
- [37] 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. 7
- [38] Aaron Nech and Ira Kemelmacher-Shlizerman. Level playing field for million scale face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7044–7053, 2017. 1
- [39] 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. 3, 4
- [40] Haibo Qiu, Baosheng Yu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao. Synface: Face recognition with synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10880–10890, 2021. 1, 3, 6, 4
- [41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2
- [42] Swami Sankaranarayanan, Azadeh Alavi, Carlos D Castillo, and Rama Chellappa. Triplet probabilistic embedding for face verification and clustering. In *2016 IEEE 8th international conference on biometrics theory, applications and systems (BTAS)*, pages 1–8. IEEE, 2016. 1
- [43] 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. 1
- [44] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016. 7
- [45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1
- [46] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1
- [47] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 1
- [48] Zhonglin Sun, Siyang Song, Ioannis Patras, and Georgios Tzimiropoulos. Cemiface: Center-based semi-hard synthetic face generation for face recognition. *Advances in Neural Information Processing Systems*, 37:35612–35638, 2025. 2, 3, 6, 7, 4
- [49] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 969–977, 2018. 1

- [50] Boris van Breugel, Trent Kyono, Jeroen Berrevoets, and Michaela van der Schaar. Decaf: Generating fair synthetic data using causally-aware generative networks. *Advances in Neural Information Processing Systems*, 34:22221–22233, 2021. [1](#)
- [51] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. [2](#)
- [52] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 765–780, 2018. [1](#)
- [53] 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](#)
- [54] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 499–515. Springer, 2016. [1](#)
- [55] Haiyu Wu, Jaskirat Singh, Sicong Tian, Liang Zheng, and Kevin W Bowyer. Vec2face: Scaling face dataset generation with loosely constrained vectors. *arXiv preprint arXiv:2409.02979*, 2024. [2](#), [3](#), [6](#), [7](#), [1](#), [4](#)
- [56] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. [1](#)
- [57] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016. [4](#)
- [58] 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. [7](#)
- [59] 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. [7](#)
- [60] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021. [1](#), [3](#), [7](#)