creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

M.S. THESIS

Complete the Feature Space: Diffusion-Based Fictional ID Generation for Face Recognition

얼굴 인식 특징 공간 완성을 위한 확산 모델 기반 가상 인물 생성

BY

이 명 연

2023 년 8 월

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Complete the Feature Space: Diffusion-Based Fictional ID Generation for Face Recognition

얼굴 인식 특징 공간 완성을 위한 확산 모델 기반 가상 인물 생성

지도교수 이 상 구

이 논문을 공학석사 학위논문으로 제출함

2023 년 4 월

서울대학교 대학원

컴퓨터 공학부

이명연

이 명 연의 공학석사 학위논문을 인준함

2023 년 6 월

위 원 장	신 영 길
부위원장	이 상 구
위 원	김 건 희

Abstract

In deep face recognition (FR) tasks, the size and diversity of the training dataset are essential factors in improving performance. Unfortunately, crawled datasets suffer from issues such as label noise, the long-tailed problem, and privacy concerns. These problems can be solved if we can generate face images while preserving IDs in either real IDs or fictional IDs. However, previous face synthesizing approaches have limitations of requiring explicit control of facial attributes or exhibiting a lack of diversity, resulting in unsuccessful FR performance. In this paper, we propose DiffFR, a method that generates diverse face images for enhancing FR datasets within core fictional identities (IDs) by utilizing an ID-preserving diffusion model. We condition the diffusion model with a representative feature called the ID feature, to condense ID information which enables the diffusion model to generate face images in either real IDs or fictional IDs. Among the numerous fictional IDs, we select core IDs that fill the void space of FR feature space, specified as improving the inter-class sparsity. Furthermore, by leveraging the ID features to predict intra-class diversities, we ensure that intra-class diversity is duly reflected in the selection of core IDs. Our experiments demonstrate that DiffFR surpasses other synthesizing methods for FR dataset augmentation on FR benchmark sets, owing to its ability to generate datasets with a high degree of intra-class diversity and inter-class sparsity.

Keywords: Face Recognition, Image Generation, Diffusion Models

Student Number: 2017-21172

Contents

Abstract **i**

1 Introduction **1**

2 Related Work **5**

2.1 Face Synthesis for FR 5

2.2 Synthesizing Diverse Images with ID Preservation 6

2.3 Core Set Selection for FR 6

3 Method **8**

3.1 Preliminary: Diffusion Models 8

3.2 Preliminary: Face Recognition 9

3.3 ID-Preserving Diffusion Models 9

3.3.1 Integrating FR Features to an ID Feature 9

3.3.2 ID Feature Conditional Diffusion Models 10

3.4 Generating Fictional IDs 12

3.4.1 Variance-Based Spherical Interpolation 12

3.4.2 Diversity-Aware Non-Maximum ID Suppression 13

4 Experiments **17**

4.1 Experimental Settings 17

4.1.1 Settings for the Diffusion Model 17

4.1.2 Settings for the FR 18

4.2 Fictional ID Augmentation 18

4.2.1 Results 20

4.2.2 Effectiveness of ID-NMS 20

4.3 Supplying Images to Tail IDs 20

4.3.1 Results 22

4.4 Dataset Evaluation 25

4.4.1 Results 25

4.5 Analysis 26

4.5.1 Effect of Variance-Based Interpolation 26

4.5.2 Comparison of Real and Synthetic Datasets 28

5 Conclusion **31**

A Comparison of Real and Synthetic Datasets **32**

A.1 Qualitative Results 32

B Qualitative Results of ID-NMS **34**

B.1 Discarded and Selected Fictional IDs 37

초록 **45**

List of Figures

1.1 Samples of fictional IDs generated by DiffFR 4

3.1 Generating a fictional ID and samples of the fictional ID 11

3.2 Generating fictional IDs 12

3.3 Distribution of similarities with nearest IDs 15

4.1 FR performance trends as the dataset width and depth increase 21

4.2 Samples of IDs with few training images 24

4.3 Comparison of samples without and with variance-based interpolation 27

A.1 Comparison of real and synthetic images. 33

B.1 Samples of discarded fictional IDs 35

B.2 Samples of selected fictional IDs 36

List of Tables

4.1 FR performance comparison when enhancing width 19
4.2 FR performance comparison when enhancing depth 22
4.3 Dataset evaluation 23
4.4 FR performance comparison of real and synthetic datasets 29

Chapter 1

Introduction

A large and diverse training dataset plays a key role in deep face recognition (FR) tasks [1, 2, 3, 4, 5, 6]. To be specific, enhancing the number of identities (IDs) and the intra-class diversity directly improves FR accuracy [1, 5, 6]. With the advent of the big data era, we can obtain a million-scale face dataset from the internet without much difficulty. Despite this accessibility, crawled datasets often suffer from label noise, the long-tailed problem, and privacy concerns [7, 8, 6]. There have been proposed methods for noise-cleaning and clustering in face datasets to cope with the issue of label noise [2, 9, 4]. But still, the long-tailed problem and privacy issues remain unsolved.

However, these aforementioned problems can be resolved when we synthesize FR datasets. The synthesized dataset is free from label noise if the synthesizing method can generate face images while preserving the target ID. For the long-tailed problem, the unbalanced number of images can be alleviated by synthesizing face images for tail IDs. Furthermore, synthetic datasets offer advantages with respect to privacy concerns, as they consist of faces in fictional IDs. Additionally, synthesizing FR datasets also has the benefit of enabling the expansion of the number of IDs in a dataset by adding

fictional IDs. On the basis of these benefits, approaches that add a face image synthesis process to the FR model in order to boost FR have been proposed [10, 1, 8, 11, 6].

Thanks to disentangled latent representations of GANs, it has been able to generate diverse face images with respect to various hand-crafted facial attributes while preserving the IDs [12, 13, 14]. However, a critical downside of GANs is that they are sensitive to the quality of alignment when generating face images [15, 16]. On the other hand, diffusion models provide better coverage of the distribution and are more robust to the alignment quality compared to GANs [17, 18, 19]. Several diffusion-based approaches can be adapted to generate diverse face images while preserving the ID, which is necessary for synthesizing FR datasets [20, 21, 6]. However, all the aforementioned diffusion-based or GAN-based synthetic approaches have limitations in that they rely on explicit controls, such as facial attributes, a target image, or text guidance, to diversify images. In contrast to the earlier approaches, our method generates diverse images without explicit controls, with enough diversity to be suitable for FR datasets, benefiting from the mode coverage of diffusion models.

Another key factor for enhancing FR datasets is to expand the number of IDs [1, 2, 5, 3, 6]. When synthesizing fictional IDs through techniques such as interpolations and random latent variables, the number of fictional IDs that can be generated is countless. However, due to the limitations of training time and computational resources, it is infeasible to utilize all synthesized fictional IDs. Therefore, a process for selecting the core IDs that significantly impact FR performance across innumerable fictional IDs is needed. The fundamental objective of FR is to learn feature representations that distinguish the IDs of individuals [22, 23]. In this context, selecting core IDs to cover the feature space thoroughly with minimal redundancy can improve the FR model's understanding of the FR feature space [24].

In this paper, we propose DiffFR, a diffusion-based synthesizing method for FR datasets that enhance the coverage of FR feature space comprehensively. Using an

ID-preserving diffusion model, DiffFR generates diverse samples within core fictional IDs that can fill the void space of the FR feature space with minimal redundancy. First, to make the diffusion model generate face images of the target ID, we condition them with a representative feature called an ID feature, which integrates features within IDs. This enables the diffusion model to leverage the discriminative power of a pre-trained FR, leading to the generation of diverse images while preserving the target IDs of both real and fictional. Fig. 1.1 displays examples of diverse images within fictional IDs generated by DiffFR. Second, we introduce non-maximum suppression ID selection (ID-NMS) for selecting core IDs among the numerous fictional IDs that facilitate the coverage of unoccupied regions in the FR feature space, while taking the intra-class diversity into consideration.

In the experiments, we ultimately show that DiffFR outperforms existing synthesizing methods for FR dataset augmentation on various FR benchmark sets. The effectiveness of DiffFR is examined by supplementing the FR datasets in terms of the number of IDs and the number of images per ID. With regard to generation quality, we show that DiffFR has the capability to generate high-quality samples of IDs whose number of training images is extremely small, which is the necessary ability for supplying images to tail IDs. Furthermore, to elucidate the rationale behind improvements in FR, we demonstrate that the designated IDs through ID-NMS increase the inter-class sparsity of the FR feature space while also exhibiting high intra-class diversity.

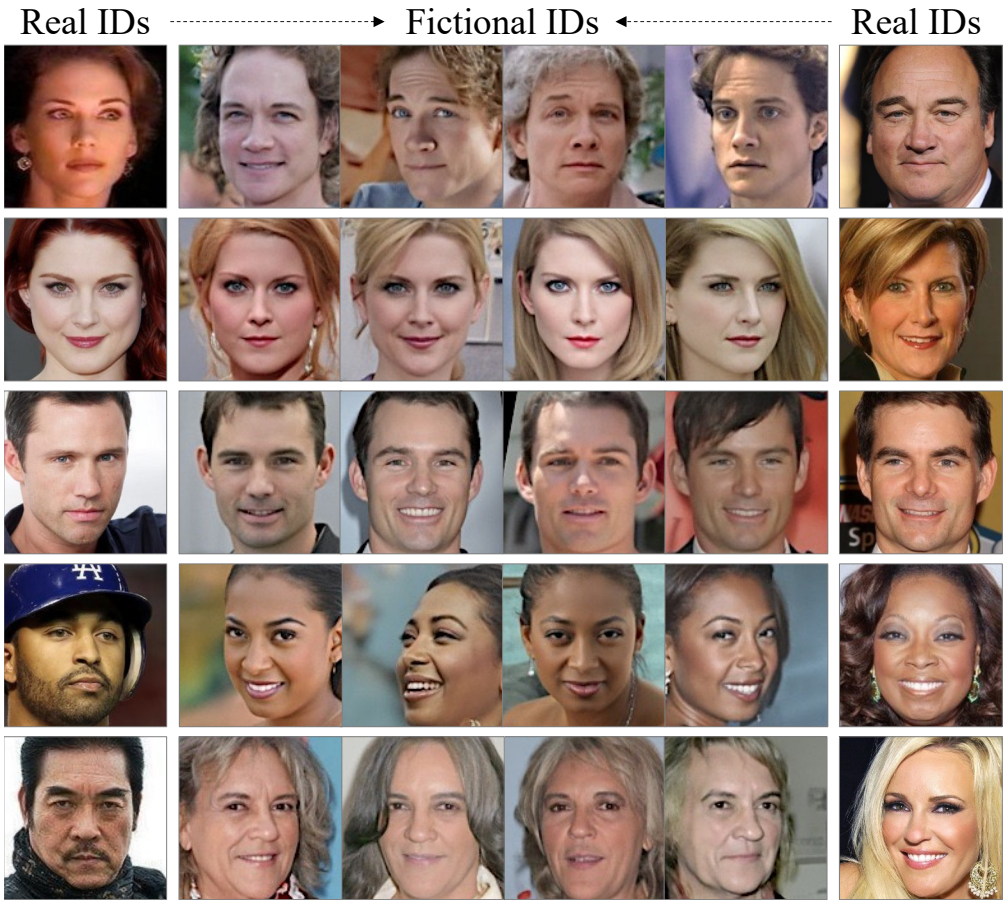

Figure 1.1 Samples of fictional IDs generated by DiffFR. Samples are generated based on the interpolations between features of the real IDs on their left and right sides. Without any explicit control of facial attributes or guidance, DiffFR generates diverse images while preserving the target ID.

Chapter 2

Related Work

2.1 Face Synthesis for FR

Attempts to enhance FR by utilizing synthesized face images have been proposed as photo-realistic face image generating techniques were invented [10, 8, 1, 25, 11, 26]. By employing face synthesis, we gain control over the properties of FR datasets, including the number of IDs or images per ID, which allows us to substantiate the influence of these properties. DCFace [6] and SynFace [1] uncover that the number of unique IDs and their intra-class diversities significantly impact FR models. In order to assess the fulfillment of desirable properties in synthetic datasets, DCFace [6] introduced three class-dependent metrics that measure the uniqueness of IDs, the preservation of IDs (intra-class consistency), and intra-class diversity. DCFace [6] and DigiFace [8] reveal that there remains a performance gap in FR between real and synthetic datasets when the datasets are of the same scale. In addition, SynFace [1] mentioned terminologies for describing the scale of FR datasets, such as “depth” for the number of images per ID and “width” for the number of IDs.

2.2 Synthesizing Diverse Images with ID Preservation

Controlling facial attributes, including pose, illumination, and expression, while preserving IDs, is a representative approach for generating diverse images to synthesize FR datasets [1, 8, 12, 25]. For instance, SynFace [1] synthesizes intra-class diverse images by utilizing DiscoFaceGAN [12] to control facial attributes. Diff-AE [20] can manipulate facial attributes of face images by conditioning diffusion models using a learnable encoder for semantics. DigiFace [8] uses a 3D rendering-based pipeline to control the facial attributes, accessories, textures, and environments of the images, enabling the generation of diverse images. However, these approaches need explicit controls of attributes in order to achieve intra-class diversity within the dataset. Without any control of hand-crafted attributes, CFSM [10] augments FR datasets to bridge the domain gap between the training dataset and the target dataset. DiffuseIT [21] also provides face image translations that guide the source images into the target domain with proper semantic changes while preserving the IDs. Likewise, DCFace [6] utilizes an external style input image to generate images with diverse styles. Nonetheless, these methods explicitly demand additional input images to extract features from the target domain or target style. Unlike the aforementioned approaches, DiffFR generates diverse images without explicit controls of facial properties or images from the target domain.

2.3 Core Set Selection for FR

As training sets for FR are getting massive, concerns about training time, excessive consumption of resources, and memory cost are raised. Face-NMS [24] mitigates this problem by resolving the redundancy problem of images in each ID. Inspired by non-maximum suppression (NMS) [27] in the detection field, Face-NMS ranks the face images by their potential contribution to the overall sparsity on the FR feature space.

This can be seen as a core selection since it condenses the number of images while minimizing the degradation of FR performance. However, distinct from our setting, Face-NMS is intended for existing datasets rather than synthetic ones, and thus it does not address an approach for selecting core IDs. Since a numerous number of fictional IDs can be generated when synthesizing an FR dataset, we introduce ID-NMS, which selectively designates the core fictional IDs to thoroughly cover the FR feature space.

Chapter 3

Method

In this section, we present DiffFR, which generates synthetic face images of both existing and fictional identities utilizing an ID feature conditional diffusion model. The generated fictional IDs are aimed to improve the coverage of the FR feature space.

3.1 Preliminary: Diffusion Models

The diffusion probabilistic model is a parameterized Markov chain whose transitions are learned to reverse a diffusion process to sample from a distribution [28]. A sampling of diffusion models starts with noise x_T and gradually removes noise until getting a clean sample x_0. A forward diffusion process $q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t I)$ adds Gaussian noise at each timestep t with variance $\beta_t \in (0, 1)$ which are hyperparameters representing the noise levels. Accordingly, the reverse process $p_\theta(x_{t-1}|x_t)$ is modeled as a diagonal Gaussian $\mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$, where there are many different ways to model the mean $\mu_\theta(x_t, t)$ and the variance $\Sigma_\theta(x_t, t)$. In case of DDPM [17], $\mu_\theta(x_t, t)$ is calculated with a parameterized model $\epsilon_\theta(x_t, t)$ that predicts

the noise of a noisy image x_t by the timestep t. The loss function for the model is a mean-squared error between the predicted noise and the actual noise ϵ which can be formulated as $\|\epsilon_\theta(x_t, t) - \epsilon\|^2$. The variance $\Sigma_\theta(x_t, t)$ is fixed to a known constant in DDPM, whereas it is parameterized with β_t and $\tilde{\beta}_t$ which are the upper and lower bounds on reverse process variances in improved DDPM (iDDPM) [18]. We employ denoising diffusion implicit model (DDIM) [29] to improve computational cost and speed of sampling. It generalizes DDPM by formulating a non-Markovian nosing process that provides the magnitude of σ_t to control stochasticity while maintaining the same marginals as the original DDPM. By erasing the term for stochasticity, the process becomes deterministic, which enables the model to produce high-quality samples much faster.

3.2 Preliminary: Face Recognition

3.3 ID-Preserving Diffusion Models

3.3.1 Integrating FR Features to an ID Feature

In order to generate a synthetic FR dataset, the diffusion model has to be conditioned to generate face images while preserving the intended ID. For the diffusion model to generate face images reliably within the corresponding ID of the given feature, it is necessary for the conditioning feature used in training to sufficiently capture the ID information. An ID centroid in the FR is a representative feature of the ID. Given that the IDs are in the FR training set, the corresponding centroids of these IDs are the weights of the fully connected (FC) layer of the FR [23, 22]. However, if the ID is not in the FR training set, an alternative way to extract a representative feature of the ID is needed. To integrate the ID information of features into a representative feature, we compute the mean feature from the features within the ID, which we call the ID feature. This enables us to obtain the ID features out of the FR training set while exploiting the

discriminative power of a pre-trained FR, which is trained on a massive FR dataset and optimized by sophisticatedly designed loss functions. When there are n images in an ID a, an ID feature of the ID a can be noted $F_a = \frac{1}{n}\sum_{i=1}^{n} f_i$ where f_i is i-th feature of the ID a.

3.3.2 ID Feature Conditional Diffusion Models

The reverse transition of our conditional diffusion model that receives ID feature F as an input can be formulated as follows,

$$p_\theta(x_{t-1}|x_t, F_{id}) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, F_{id}), \Sigma_\theta(x_t, t)). \tag{3.1}$$

Here, $\mu_\theta(x_t, t, F_{id})$ is parameterized with the function approximator ϵ_θ that is intended to predict ϵ from x_t [17] as follows,

$$\mu_\theta(x_t, t, F_{id}) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t, F_{id})\right) \tag{3.2}$$

where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t}(1-\beta_s)$. From the defined reverse transitions, the variational bound from μ, notated as L_{simple}, can be formulated as follows,

$$L_{simple}(\theta) := E_{t,x_0,\epsilon}\left[||\epsilon - \epsilon_\theta(x_t, t, F_{id})||^2\right] \tag{3.3}$$

where $\epsilon \in \mathbb{R}^{3\times h\times w} \sim \mathcal{N}(0, I)$. For reflecting the condition to the UNet, we use the scale and shift approach, which is also called adaptive group normalization (AdaGN), proposed by [19]. Conditioning with ID features enables the model to generate not only face images of an existing ID in the training set but also that of a fictional ID. The method for generating fictional ID features that enhance the FR performance is explained in the following Sec. 3.4.

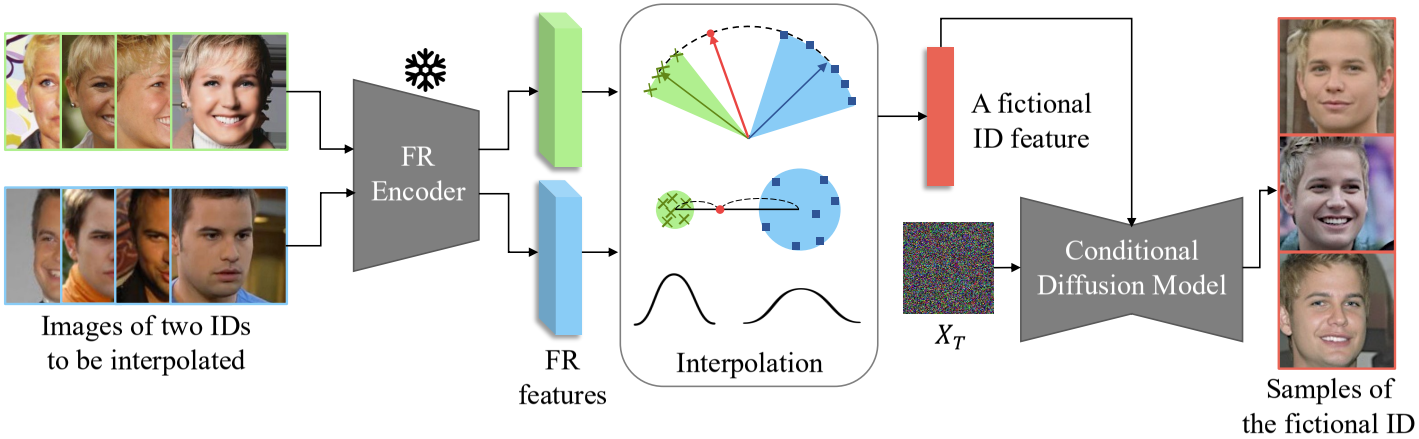

Figure 3.1 Generating a fictional ID and samples of the fictional ID. As noted in Eq. 3.4, element-wise variances of the IDs to be mixed are taken into consideration when generating a feature of a fictional ID. The generated feature of the fictional ID can be fed into the conditional diffusion without any additional process.

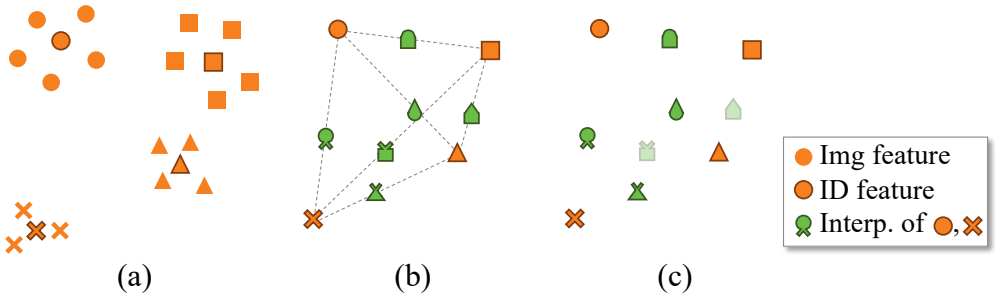

Figure 3.2 Generating fictional IDs. The process can be divided into three parts: (a) Integrating features into an ID feature for each ID, (b) Variance-based spherical interpolations, and (c) Selecting core IDs utilizing ID-NMS.

3.4 Generating Fictional IDs

As depicted in Fig. 3.2, the process of generating fictional IDs consists of three components: integrating features into an ID feature for each ID, generating fictional IDs, and selecting core fictional IDs. The strategy for integrating features is mentioned in Sec. 3.3.

3.4.1 Variance-Based Spherical Interpolation

Basically, a fictional ID is generated with an interpolation between two existing IDs from the training set. As illustrated in Fig. 3.1, we find that element-wise feature variance varies across each ID. Therefore, to locate a fictional ID sufficiently distant from the interpolation endpoint IDs, indicating its uniqueness, the variances of the IDs should be taken into consideration. Since angular margin-based losses are used for training FR, FR features are considered to be on the spherical feature space [30, 31]. In total, we use spherical interpolation with the proportion of two IDs' element-wise variance. A feature of a fictional ID F_c which is the interpolation between the feature

F_a and F_b can be formulated as follows,

$$\begin{aligned} F_c = & \frac{\sin(V_b \Omega_{ab}/(V_a + V_b))}{\sin(\Omega_{ab})} \odot F_a \\ & + \frac{\sin(V_a \Omega_{ab}/(V_a + V_b))}{\sin(\Omega_{ab})} \odot F_b \end{aligned} \tag{3.4}$$

where V_a and V_b are the element-wise variance of each feature F_a, F_b. Note that $\odot$ is Hadamard product and Ω_{ab} is the angle subtended by the arc whose first and last points are F_a and F_b.

3.4.2 Diversity-Aware Non-Maximum ID Suppression

According to our method for generating fictional IDs, the number of fictional IDs that can be generated is ${}_nC_2$ where n is the number of IDs in the training set. For instance, if the number of IDs in the training set for the diffusion model is 10k, the number of fictional IDs that can be generated is approximately 50M. Considering that the number of IDs of WebFace260M [2] is 4M, which is the largest public FR dataset in existence, a process of selecting core IDs from among the numerous fictional IDs is needed.

Similar to the concept of the global sparsity in Face-NMS [24], which can be seen as intra-class sparsity, we assume that improving the inter-class sparsity of FR feature space can also improve FR performance. We propose Non-Maximum Suppression ID Selection (ID-NMS), an approach that applies NMS to IDs, selecting IDs that improve inter-class sparsity to achieve comprehensive coverage of the FR feature space. Moreover, we consider the intra-class diversities of IDs during the selection process, as intra-class diversity is a crucial factor for an effective FR dataset.

As defined in Algorithm 1. ID-NMS sequentially finds the furthest ID from the selected core IDs and samples one ID among k-nearest neighbors of the furthest ID based on their intra-class diversities. We employ softmax of predicted intra-class diversities of the k-nearest neighbors with temperature τ to form a categorical distribution

as

$$p_{\text{div}}(x_i) = \frac{e^{d_i/\tau}}{\Sigma_{i=1}^{k} e^{d_i/\tau}}, \tag{3.5}$$

where $k = 10$ is the k from k-nearest neighbors, and d_i is the predicted intra-class diversity of x_i. And then, ID-NMS truncates IDs if their similarities to already selected IDs are greater than the specified threshold. The threshold is set based on cosface [31] margin value, 0.35. The predicted intra-class diversities are obtained through training a simple linear regressor that infers intra-class diversities from ID features.

Consequently, as shown in Fig. 3.3, the distribution of similarities with nearest IDs selected by ID-NMS indicates an average of 0.3864, which is lower overall compared to that selected randomly with an average of 0.5226. This implies that ID-NMS increases inter-class sparsity while random selection exhibits ID redundancy, as evidenced by the considerable number of pairs whose similarities exceed 0.45. Furthermore, we provide measurements of metrics that demonstrate that the IDs selected by ID-NMS exhibit higher levels of ID uniqueness and intra-class diversities in Sec. 4.4.

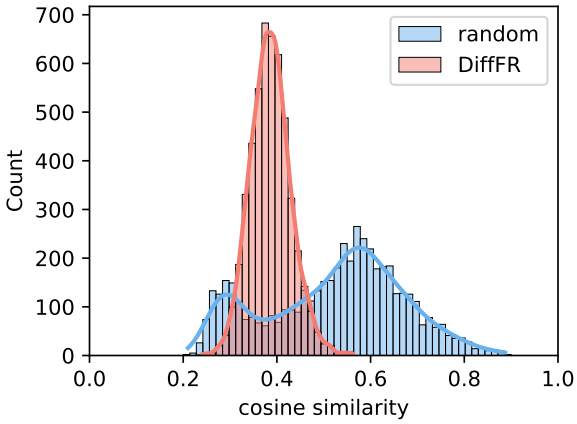

Figure 3.3 Distribution of similarities with nearest IDs. The distribution of DiffFR is lower overall than that of randomly selected. This indicates that DiffFR preferentially selects IDs whose inter-class similarities are small.

Algorithm 1 ID-NMS sampling

Require: set of N fictional IDs $\mathcal{X}$, set of similarities $\mathcal{S}$, set of predicted diversities $\mathcal{D}$, similarity threshold t

Result: set of selected core IDs $\mathcal{C}$

$\mathcal{X} = \{x_1, \cdots, x_N\}$
$\mathcal{S} = \{s_{1,2}, s_{1,3}, \cdots, s_{N-1,N}\}$
$\mathcal{D} = \{d_1, \cdots, d_N\}$
$\mathcal{C} = \{\}$
x_{curr} = random($\mathcal{X}$)
while $\mathcal{X} \neq \emptyset$ **do**
    $\mathcal{C} \leftarrow \mathcal{C} + \{x_{\text{curr}}\}$
    $\mathcal{X} \leftarrow \mathcal{X} - \{x_{\text{curr}}\}$
    for x_i in $\mathcal{X}$ **do**
        if $s_{\text{curr},i} \geq t$ **then** ▷ Discard x_i if too similar
            $\mathcal{X} \leftarrow \mathcal{X} - \{x_i\}$
            $\mathcal{S} \leftarrow \mathcal{S} - \{s_{\text{curr},i}\}$
        end if
    end for
    $j \leftarrow \arg\min_j s_{\text{curr},j}$ ▷ Find the furthest ID
    $\mathcal{M} \leftarrow \{m | x_m \in k\text{-NN}(x_j, \mathcal{X})\}$
    $x_l \sim p_{\text{div}}, \;\; p_{\text{div}} = \text{softmax}(\mathcal{D}_{\mathcal{M}}, \tau)$ ▷ Eq. (3.5)
    $x_{\text{curr}} \leftarrow x_l$
end while
return $\mathcal{C}$

Chapter 4

Experiments

In this section, we evaluate the performance of different generative methods for synthesizing facial images to augment FR datasets in terms of depth and width. In addition, we evaluate and compare all the synthesized datasets based on three crucial virtues of synthesized FR datasets, which include the number of unique IDs, ID preservation, and intra-class diversity. We also show qualitative results of generating images of both real and fictional IDs, and the results with and without variance-based interpolating.

4.1 Experimental Settings

4.1.1 Settings for the Diffusion Model

The ID-preserving diffusion model is trained on CelebA [32] consisting of 10,177 IDs and 202,599 face images. Same as introduced in DDPM [17], we use the UNet [33] model architecture and cosine scheduling [18] for the noise level. For detailed architectural settings, we follow the settings in [19]. The FR model for conditioning the diffusion model is a pre-trained model from Insightface [34] which is a modified

ResNet100 [30] trained on Glint360k [35]. All features for conditioning are normalized to locate them on the spherical feature space.

4.1.2 Settings for the FR

For the base training FR datasets, CelebA [32] and CASIA-WebFace [36] are used. We augment the base dataset with synthesized data and train the FR model on the combined set, following the domain mixup strategy employed in [1]. For training the FR, we employ a modified ResNet50 modified in [30] with CosFace [31] loss function. Optimizer SGD [37, 38, 39] is applied with a momentum of 0.9 and a weight decay of 5e-4. One NVIDIA Tesla A100 GPUs is used in the training with 1,024 batch size. The learning rate is initially set to 0.1 and decreased by 10 at 10, 16, 21, and 25 epochs and training terminates at 30 epochs. During training, we only use flip data augmentation. For evaluation, we use widely used verification sets such as LFW [40], CFP-FP, CFP-FF [41], AgeDB-30[42], CPLFW [43] and CALFW [44].

4.2 Fictional ID Augmentation

In this experiment, we compare the FR performance improvement when adding fictional IDs using DiffFR and other synthesizing methods. To demonstrate the effectiveness of the core ID selection, we conduct an additional comparison between our DiffFR with and without ID-NMS. We select face synthesis methods for comparison that are able to generate face images of fictional IDs, including SynFace [1], DiscoFaceGAN [12], DiffAE [20], DCFace [6]. Since DiffuseIT [21] does not have the ability to generate fictional IDs, we utilize IDs from FFHQ [45] for augmenting additional IDs.

Base Dataset	Method	LFW	CFP-FP	AgeDB	CFP-FF	CALFW	CPLFW	Avg.
CelebA	Base set	97.88	80.29	86.25	97.90	90.78	78.52	88.60
	SynFace	97.85	81.37	85.83	97.84	89.98	78.45	88.56
	DiscoFaceGAN	98.33	84.59	86.70	98.40	91.08	81.52	90.10
	DiffAE	92.98	61.34	64.40	78.60	72.77	65.13	72.54
	DiffuseIT	95.26	63.56	72.43	80.04	76.95	67.52	75.96
	DCFace	98.53	86.21	90.17	98.54	91.57	81.42	91.07
	DiffFR (wo/ ID-NMS)	98.83	86.03	89.57	98.53	91.63	81.50	91.02
	DiffFR	**99.05**	**86.96**	**91.40**	**98.81**	**92.20**	**83.10**	**91.92**
CASIA	Base set	99.17	93.31	92.62	99.13	92.88	87.10	94.03
	SynFace	99.20	93.54	92.40	99.17	92.72	87.53	94.09
	DiscoFaceGAN	99.32	93.37	93.13	99.13	92.85	87.33	94.19
	DiffAE	99.17	93.99	92.72	99.31	92.80	87.05	94.17
	DiffuseIT	99.25	93.74	92.97	99.37	93.17	87.68	94.36
	DCFace	**99.38**	94.16	93.15	**99.50**	93.07	88.02	94.54
	DiffFR (wo/ ID-NMS)	99.30	94.37	94.87	99.40	93.53	88.93	95.07
	DiffFR	**99.38**	**94.57**	**95.00**	99.46	**93.78**	**89.57**	**95.29**

Table 4.1 Verification accuracy(%) comparison on benchmark sets of DiffFR, various synthesizing methods, and DiffFR without ID-NMS when enhancing FR dataset **width**.

4.2.1 Results

Tab. 4.1 illustrates the results on the verification sets when the FR datasets are supplemented in terms of width. The results indicate that supplying additional IDs to the datasets improves FR performances, especially when DiffFR is used for synthesis, compared to the performance achieved by previous approaches. When the base dataset is CelebA, even though DCFace achieves the best average accuracy among previous methods on verification sets of 90.07%, DiffFR outperforms them with an accuracy of 91.92%. We also observe a similar trend when CASIA is used for the base dataset. Despite the promising results achieved by DCFace on the verification sets with an accuracy of 94.54%, DiffFR continues to exhibit outstanding accuracy with a score of 95.29%. In addition, as can be seen in Fig. 4.1 (a), we show that the FR accuracy improves as the number of fictional IDs increases for all methods tested, which aligns with the observations reported in SynFace [1].

4.2.2 Effectiveness of ID-NMS

The comparison of DiffFR with and without ID-NMS reveals that the inclusion of ID-NMS leads to improved overall FR accuracy. When using CASIA as the base set, the incorporation of ID-NMS brings an improvement in the best average validation accuracy, with a score of 95.29% compared to 95.07% without ID-NMS. As will be demonstrated through measurements of metrics in Sec. 4.4, we assert that these improvements in FR performance are attributable to the improvement of ID uniqueness and intra-class diversity in the datasets achieved by ID-NMS.

4.3 Supplying Images to Tail IDs

This experiment aims to compare DiffFR with other generative models for supplying images to tail IDs in the base FR dataset. We select generative models that can generate

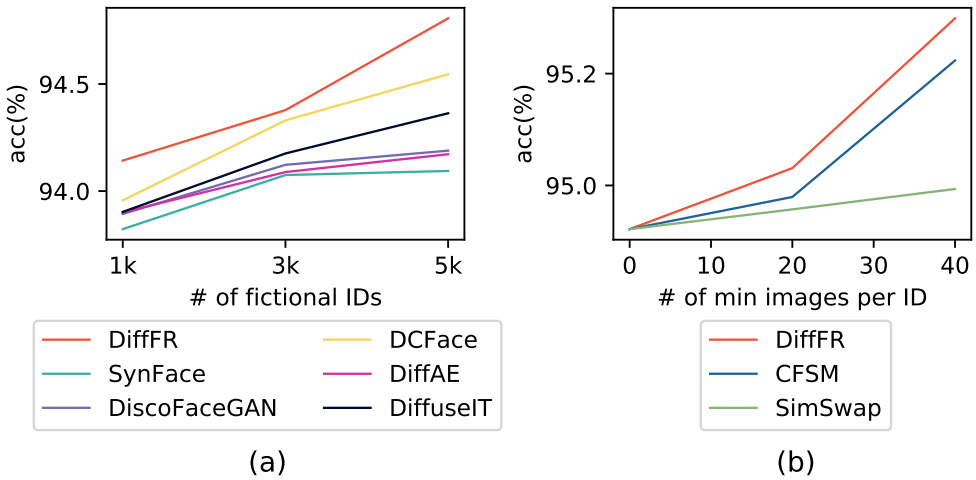

Figure 4.1 Verification performance(%) trends as the dataset width and depth increase. (a) shows accuracies according to the number of fictional IDs added(width), and (b) shows accuracies according to the number of images supplied to tail classes(depth). The base dataset is CASIA.

Base Dataset	Method	LFW	CFP-FP	AgeDB	CFP-FF	CALFW	CPLFW	Avg.
CelebA	Base set	97.88	80.29	86.25	97.90	90.78	78.52	88.60
	SimSwap	98.47	84.89	91.82	98.99	91.70	80.67	91.09
	CFSM	97.48	79.01	87.12	97.96	89.62	77.77	88.16
	DiffFR	**98.93**	**86.97**	**92.05**	**98.79**	**92.50**	**82.80**	**92.01**
CASIA +CelebA	Base set	**99.40**	94.40	94.27	**99.51**	93.30	88.65	94.92
	SimSwap	99.23	94.36	94.97	99.27	93.47	88.67	94.99
	CFSM	99.30	94.40	95.05	99.46	93.87	89.27	95.22
	DiffFR	99.25	**94.41**	**95.08**	99.43	**93.92**	**89.70**	**95.30**

Table 4.2 Verification accuracy (%) comparison on benchmark sets of DiffFR and GAN-based synthesizing methods when enhancing FR dataset **depth**. We supplied synthesized images to make every ID have at least 40 images.

diverse images within the given existing IDs. We add synthesized images to make every ID have at least 40 images since SynFace [1] shows that FR accuracy saturates when the depth of a dataset is greater than 30.

4.3.1 Results

Tab. 4.2 displays benchmark results for FR datasets enhanced in terms of depth. DiffFR shows the highest accuracy among all the other methods on the average of the benchmark sets, whether CelebA or the composite of CASIA and CelebA (CASIA+CelebA) is used as the base dataset. In case of LFW and CFP-FF, which are highly saturated benchmark sets, all the methods exhibit a slight degradation of FR performance, when CASIA+CelebA is used for the base dataset. However, the overall results show that enhancing FR datasets in terms of depth leads to improved FR performance. Additionally, Fig. 4.1 (b) illustrates that the FR accuracy for all the methods improves as

Method	$U_{\text{class}} \uparrow$	$C_{\text{intra}} \uparrow$	$S_{\text{intra}} \downarrow$
SimSwap	0.9570	0.8276	0.6734
CFSM	0.9465	0.9258	0.7520
SynFace	0.0000	0.4465	0.7637
DiscoFaceGAN	0.3374	0.9741	0.8385
DiffAE	0.9794	0.9922	0.8809
DiffuseIT	0.8922	**0.9984**	0.8118
DCFace	0.9924	0.6816	0.7048
DiffFR (wo/ var)	0.9364	0.6466	**0.6543**
DiffFR (wo/ ID-NMS)	0.7322	0.7325	0.6783
DiffFR	**0.9986**	**0.9517**	**0.6615**

Table 4.3 Comparison of dataset evaluation metrics based on three crucial trait: the number of unique IDs (U_{class}), ID preservation (C_{intra}), and intra-class diversity (S_{intra}). A lower value for S_{intra} indicates higher degree of intra-class diversity. "wo/ var" refers to that interpolated without variance.

the number of supplied images increases, consistent with the observations reported in SynFace [1].

By leveraging the knowledge of the FR, DiffFR can stably generate diverse images of a real ID even when the number of images for training is extremely limited, or the images are of low quality. In the first three rows of the Fig. 4.2, each ID has only a single training image, and in the remaining rows, only two or three low-quality images were used for training each ID. Despite these limitations, DiffFR generates diverse samples while stably preserving the target ID.

Figure 4.2 Samples of IDs with few training images. The first column indicates the number of images for the ID utilized for training. Even with a small number of low-quality images, DiffFR can generate diverse images.

4.4 Dataset Evaluation

In this section, we evaluate datasets based on three crucial traits: the number of unique IDs, ID preservation, and intra-class diversity following DCFace [6]. Among three metrics proposed in DCFace [6], we utilize two of the class-dependent metrics proposed in DCFace [6], namely uniqueness (U_{class}) and intra-class consistency (C_{intra}). However, the dataset synthesized by DiffFR does not have the real images that correspond to the synthesized images, making it impossible to measure the metric for intra-class diversity proposed in DCFace [6]. Moreover, the metric measures the distance of the style distribution between the real datasets used for diversification and the synthesized datasets, which implies that it does not provide a direct measurement of intra-class diversity. We introduce intra-class sparsity (S_{intra}) to measure how diverse images are generated within the same ID by modifying the global sparsity measure proposed in face-nms [24]. To make the measure less susceptible to ID information, we define S_{intra} by replacing FR features of global sparsity in [24] with Inception features [46]. In order to attain statistically reliable results, at least 5k IDs and 20 images per ID are involved when calculating global sparsity. To ensure fairness, another pre-trained FR model from Insightface [34] with a modified ResNet50 backbone [30] trained on Glint360k [35], is employed for measuring the metrics. For uniqueness and intra-class consistency, r is set to 0.45, aligning with the threshold of the FR model for determining matches or non-matches.

4.4.1 Results

Tab. 4.3 summarizes the evaluation result measuring U_{class}, C_{intra}, and S_{intra}. Among all methods, DiffFR demonstrates the highest ID uniqueness. However, when ID-NMS is not applied to DiffFR, the ID uniqueness is degraded, as ID-NMS plays a crucial role in removing redundant IDs. The ID uniqueness of SynFace [1] collapses because

they mix IDs to diversify images within IDs, resulting in the breakdown of the ID boundaries. Regarding ID preservation, it has been reported by DCFace [6] that there exists a tradeoff between ID preservation and intra-class diversity, and achieving the best performance in FR requires a balance between these two factors. This is because higher intra-class diversity suggests that the images associated with a particular ID are less similar to each other, thus implying a higher degree of diversity. Consistent with the aforementioned consensus, DiffuseIT [21] and DiffAE [20] exhibit high levels of ID preservation but show lower intra-class diversities. On the other hand, DiffFR shows high intra-class diversity while simultaneously maintaining high levels of ID preservation. Without variance-based interpolation, DiffFR displays the highest intra-class diversity, but ID preservation considerably decreases from 0.9517 to 0.6466. In sum, the dataset generated by DiffFR exhibits the highest ID uniqueness and achieves a well-balanced combination of intra-class diversity and ID preservation, resulting in the best subsequent FR performance.

4.5 Analysis

4.5.1 Effect of Variance-Based Interpolation

Fig. 4.3 illustrates the qualitative distinction between samples of fictional IDs generated with and without variance-based interpolation. When the fictional ID is generated without considering variances, the fictional ID may be too similar to a real ID with higher variance than the other real ID. On the other hand, variance-based interpolation enables the fictional ID to be distinct from the real IDs which are the interpolation endpoints. This observation aligns with the findings in Tab. 4.3, indicating that when fictional IDs are interpolated without variance, their uniqueness is degraded compared to when variance-based interpolation is used. Thus, employing variance-based interpolation contributes to minimizing redundancy in the generation of a fictional ID.

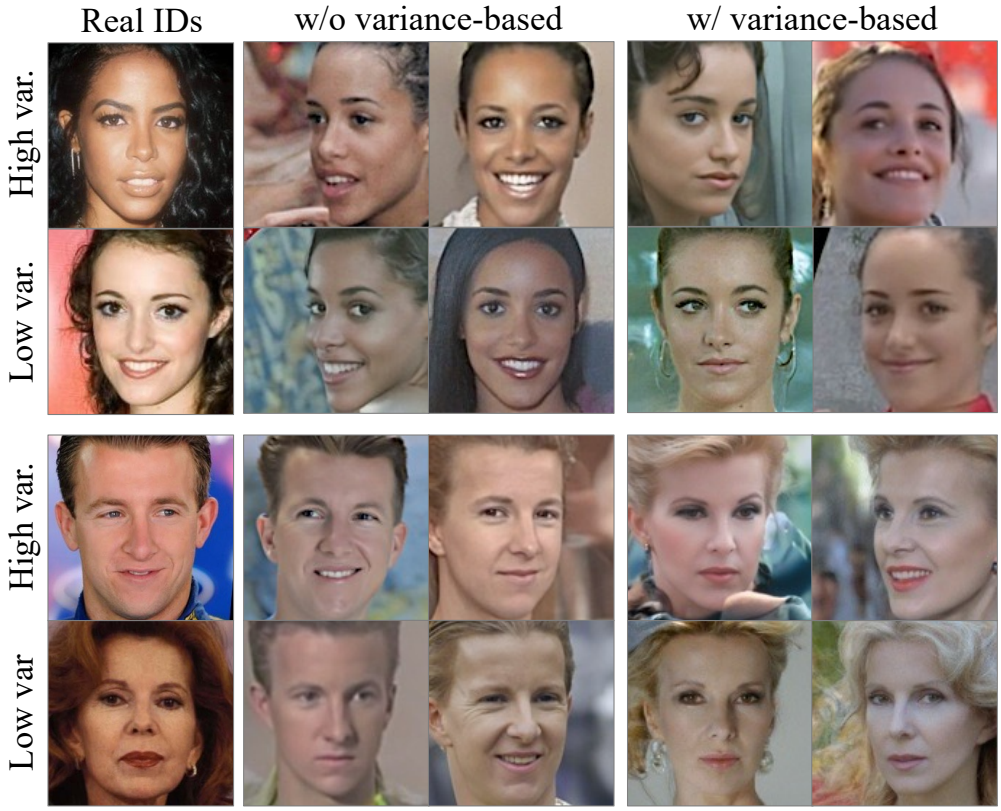

Figure 4.3 Comparison of samples without and with variance-based interpolation. The real IDs in the first column are the endpoints of interpolations. For each pair of real IDs, the upper ID has a greater average variance than the below one. Variance-based interpolation helps prevent the fictional IDs from being too similar to the real IDs with high variance.

4.5.2 Comparison of Real and Synthetic Datasets

The performance of FR models is observed to deteriorate when trained on synthetic datasets, in comparison to real datasets, due to the existing domain gap between them [1, 8, 6]. Here, we conduct a comparative analysis of the FR performance between real and synthetic datasets, covering both real IDs and fictional IDs. We maintain the number of IDs at 10,177, which is that of CelebA while manipulating the number of images per ID and the authenticity of the identifiers by varying them between real and fictional. The gap between synthetic and real datasets in Tab. 4.4 is calculated as $(\mathrm{REAL} - \mathrm{SYN})/\mathrm{SYN}$, which represents the improvement required for the synthetic dataset to match the performance of the real dataset.

Real/Synth	Authenticity	# of imgs	LFW	CFP-FP	AgeDB	CFP-FF	CALFW	CPLFW	Avg.	Gap to Real
Synth.	Real ID	20	96.20	70.87	77.68	94.74	85.83	71.63	82.83	6.96
		40	**97.58**	**72.04**	**79.83**	**96.00**	**87.20**	72.73	**84.23**	**5.18**
	Fict. ID	20	95.18	69.71	74.63	93.56	67.97	83.93	80.83	9.61
		40	96.75	70.34	78.20	93.94	69.23	**84.18**	82.11	7.90
Real	-	20	97.88	80.29	86.25	97.90	90.78	78.52	88.60	0.00

Table 4.4 Verification accuracy (%) comparison on benchmark sets of real and synthetic datasets generated by DiffFR. "# of imgs" is the average number of images per ID. The number of IDs is fixed with 10,177 which is that of the original dataset.

As can be seen in Tab. 4.4, a noticeable performance gap continues to exist between real and synthetic datasets. In addition to the observation, if the number of images is the same, the datasets with real IDs perform better. Despite the observed degradation, the results show that synthetic datasets with fictional IDs can serve as a viable alternative to real datasets, mitigating privacy concerns, with a marginal gap of 7.90%. On top of that, synthetic datasets with fictional IDs demonstrate robustness to pose variations, achieving an accuracy of 84.18%, compared to the accuracy of 78.52% for the real dataset. In line with the consistent observations, an increase in the number of images per ID positively impacts the overall performance.

Chapter 5

Conclusion

In this work, we propose DiffFR, a method for generating diverse face images within core fictional IDs using an ID-preserving diffusion model to enhance the coverage of the FR feature space thoroughly. DiffFR designates fictional IDs that increase the inter-class sparsity of the FR feature space, leading to the improvement of FR. The FR trained on the synthesized dataset generated by DiffFR shows better performance on FR benchmark datasets than other synthesizing approaches. Moreover, due to its ability to exploit the knowledge of a pre-trained FR, DiffFR can generate high-quality samples of an ID whose number of training images is extremely small. The experiments conducted show that the synthesized dataset generated by DiffFR has a high degree of intra-class diversity, and inter-class sparsity.

Appendix A

Comparison of Real and Synthetic Datasets

The performance of FR models is observed to deteriorate when trained on synthetic datasets, in comparison to real datasets, due to the existing domain gap between them [1, 8]. In this section, we present a qualitative comparison of real and synthetic datasets.

A.1 Qualitative Results

Fig. A.1 provides a comparison of real and synthetic images. Within each row of the figure, the images share identical IDs. Synthetic images demonstrate comparable levels of photo-realism and competitive diversity in comparison to real images.

Figure A.1 Comparison of real and synthetic images. Images in each row has the same ID. In comparison to real images, synthetic images exhibit comparable photo-realism and competitive diversity.

Appendix B

Qualitative Results of ID-NMS

This section presents a qualitative comparison of fictional IDs that are discarded and selected according to ID-NMS.

Figure B.1 Samples of **discarded** fictional IDs. Samples are generated based on the interpolations between features of the real IDs on their left and right sides. According to the ID-NMS, IDs are discarded when it is too similar to existing IDs or already selected IDs.

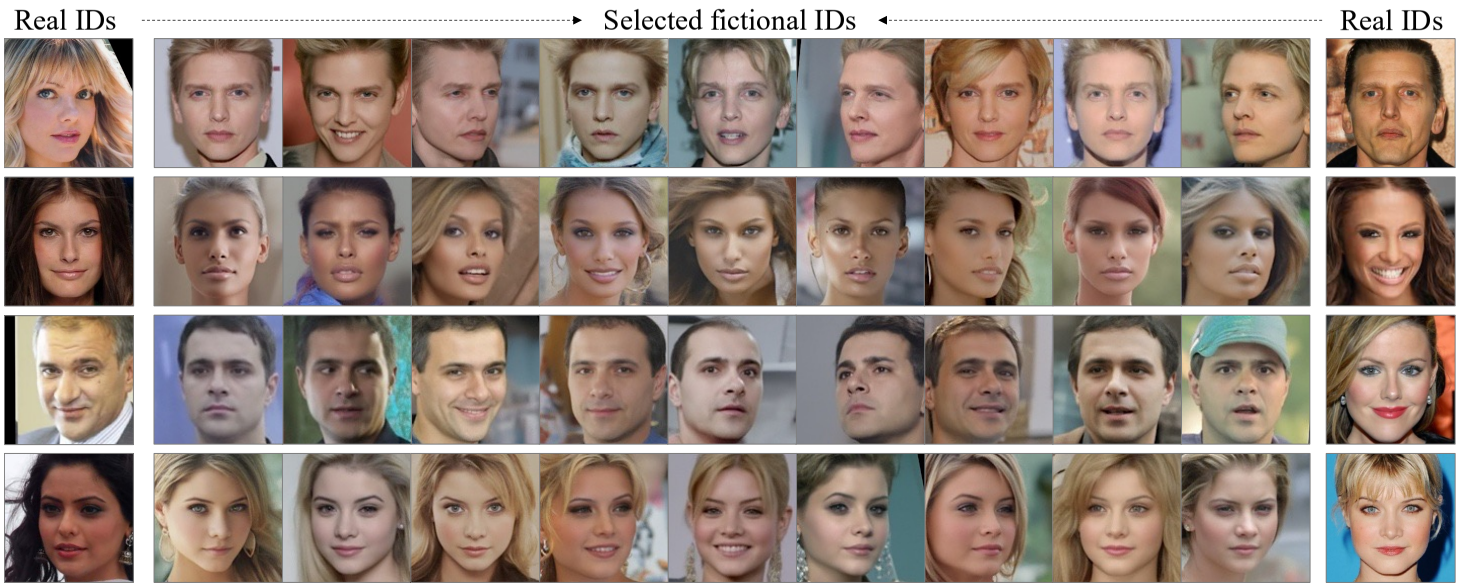

Figure B.2 Samples of **selected** fictional IDs. Samples are generated based on the interpolations between features of the real IDs on their left and right sides.

B.1 Discarded and Selected Fictional IDs

According to ID-NMS, it discards IDs when they are too similar to existing IDs or already selected IDs. Fig. B.1 depicts samples of discarded fictional IDs. Samples are generated based on the interpolations between features of the real IDs on their left and right sides. As can be seen in Fig. B.1, discarded fictional IDs resemble the interpolation endpoints, as observed empirically. In contrast, the selected fictional IDs depicted in Fig. B.2 exhibit a lesser degree of resemblance to the interpolation endpoints.

Bibliography

[1] H. Qiu, B. Yu, D. Gong, Z. Li, W. Liu, and D. Tao, "Synface: Face recognition with synthetic data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10880–10890, 2021.

[2] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. 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*, pp. 10492–10502, 2021.

[3] Y. Guo, L. Zhang, Y. Hu, X. He, and J. 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*, pp. 87–102, Springer, 2016.

[4] A. Nech and I. Kemelmacher-Shlizerman, "Level playing field for million scale face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7044–7053, 2017.

[5] J. Cao, Y. Li, and Z. Zhang, "Celeb-500k: A large training dataset for face recognition," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 2406–2410, IEEE, 2018.

[6] M. Kim, F. Liu, A. Jain, and X. Liu, "Dcface: Synthetic face generation with dual condition diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12715–12725, 2023.

[7] F. Wang, L. Chen, C. Li, S. Huang, Y. Chen, C. Qian, and C. Change Loy, "The devil of face recognition is in the noise," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 765–780, 2018.

[8] G. Bae, M. de La Gorce, T. Baltrušaitis, C. Hewitt, D. Chen, J. Valentin, R. Cipolla, and J. Shen, "Digiface-1m: 1 million digital face images for face recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3526–3535, 2023.

[9] L. Yang, X. Zhan, D. Chen, J. Yan, C. C. Loy, and D. Lin, "Learning to cluster faces on an affinity graph," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2298–2306, 2019.

[10] F. Liu, M. Kim, A. Jain, and X. Liu, "Controllable and guided face synthesis for unconstrained face recognition," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, pp. 701–719, Springer, 2022.

[11] A. Sevastopolsky, Y. Malkov, N. Durasov, L. Verdoliva, and M. Nießner, "How to boost face recognition with stylegan?," *arXiv preprint arXiv:2210.10090*, 2022.

[12] Y. Deng, J. Yang, D. Chen, F. Wen, and X. 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*, pp. 5154–5163, 2020.

[13] Y. Shen, P. Luo, J. Yan, X. Wang, and X. 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*, pp. 821–830, 2018.

[14] Y. Shen, B. Zhou, P. Luo, and X. Tang, "Facefeat-gan: a two-stage approach for identity-preserving face synthesis," *arXiv preprint arXiv:1812.01288*, 2018.

[15] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[16] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[18] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning*, pp. 8162–8171, PMLR, 2021.

[19] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.

[20] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn, "Diffusion autoencoders: Toward a meaningful and decodable representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10619–10629, 2022.

[21] G. Kwon and J. C. Ye, "Diffusion-based image translation using disentangled style and content representation," *arXiv preprint arXiv:2209.15264*, 2022.

[22] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1701–1708, 2014.

[23] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

[24] Y. Chen, J. Huang, J. Zhu, Z. Zhu, T. Yang, G. Huang, and D. Du, "Face-nms: A core-set selection approach for efficient face recognition," *arXiv preprint arXiv:2109.04698*, 2021.

[25] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter, "Analyzing and reducing the damage of dataset bias to face recognition with synthetic data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

[26] D. S. Trigueros, L. Meng, and M. Hartnett, "Generating photo-realistic training data to improve face recognition accuracy," *Neural Networks*, vol. 134, pp. 86–94, 2021.

[27] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4507–4515, 2017.

[28] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*, pp. 2256–2265, PMLR, 2015.

[29] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[30] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

[31] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5265–5274, 2018.

[32] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[33] O. Ronneberger, P. Fischer, and T. 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*, pp. 234–241, Springer, 2015.

[34] J. Guo, J. Deng, X. An, J. Yu, and B. Gecer, "insightface." `https://github.com/deepinsight/insightface`, 2021.

[35] X. An, X. Zhu, Y. Xiao, L. Wu, M. Zhang, Y. Gao, B. Qin, D. Zhang, and F. Ying, "Partial fc: Training 10 million identities on a single machine," in *Arxiv 2010.05222*, 2020.

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

[37] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *International conference on machine learning*, pp. 1139–1147, PMLR, 2013.

[38] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural networks*, vol. 12, no. 1, pp. 145–151, 1999.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[40] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[41] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *2016 IEEE winter conference on applications of computer vision (WACV)*, pp. 1–9, IEEE, 2016.

[42] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "Agedb: the first manually collected, in-the-wild age database," in *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 51–59, 2017.

[43] T. Zheng and W. Deng, "Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments," *Beijing University of Posts and Telecommunications, Tech. Rep*, vol. 5, no. 7, 2018.

[44] T. Zheng, W. Deng, and J. Hu, "Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments," *arXiv preprint arXiv:1708.08197*, 2017.

[45] V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1867–1874, 2014.

[46] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

초록

심층 얼굴 인식(Deep Face Recognition)에서 학습 데이터셋의 크기와 다양성은 성능 향상에 중요한 요소이다. 하지만, 수집된 데이터셋(crawled dataset)은 레이블 오류, 긴꼬리 문제(long-tailed problem)와 개인정보 문제로 인해 사용에 한계가 있다. 이러한 문제를 해결하기 위해 기존 연구에서는 실제 데이터셋을 보완하기 위한 얼굴 데이터셋 합성 방법들이 제안되었다. 그러나 기존 접근법들은 명시적인 특징 제어를 필요로 하거나 생성된 이미지의 다양성이 부족하여 성공적인 얼굴 인식 성능을 달성하지 못하였다. 본 논문에서는 인물 보존 확산 모델(ID-preserving diffusion model)을 활용하여 핵심 가상 인물(ID)을 생성하고, 이를 기반으로 얼굴 인식 성능을 향상시키는 DiffFR 방법을 제안한다. DiffFR는 인물 정보를 압축하여 추출한 인물 특징 벡터(ID feature)를 조건부 확산 모델(conditional diffusion model)에 입력하여 학습하여 기존 인물 뿐만 아니라 가상 인물의 얼굴 이미지도 생성이 가능하도록 한다. 또, 생성 가능한 무수히 많은 가상 인물들 중, 클래스 간 희소성(inter-class sparsity)을 향상 시켜 얼굴 인식 특징 공간의 빈 공간을 채워주는 핵심 인물들을 선정하여 생성한다. 결과적으로, DiffFR가 다른 얼굴 데이터 합성 방법을 통한 데이터 증강(data augmentation)에 비해 우수한 성능을 보인다는 것을 실험적으로 보인다. 또, 이는 클래스 내 다양성(intra-class diversity)과 클래스간 희소성에 기인한 것임을 실험적으로 보인다.

주요어: 얼굴 인식, 이미지 생성, 확산 모델

학번: 2017-21172