Learning Flexibly Distributional Representation for Low-quality 3D Face Recognition

Zihui Zhang, Cuican Yu, Shuang Xu, Huibin Li *
Xi'an Jiaotong University
{zhangzihui247, ccy2017, shuangxu}@stu.xjtu.edu.cn, huibinli@xjtu.edu.cn

Abstract

Due to the superiority of using geometric information, 3D Face Recognition (FR) has achieved great successes. Existing methods focus on high-quality 3D FR which is unpractical in real scenarios. Low-quality 3D FR is a more realistic scenario but the low-quality data are born with heavy noises. Therefore, exploring noise-robust low-quality 3D FR methods becomes an urgent and challenging problem. To solve this issue, in this paper, we propose to learn flexibly distributional representation for low-quality 3D FR. Firstly, we introduce the distributional representation for low-quality 3D faces due to that it can weaken the impact of noises. Generally, the distributional representation of a given 3D face is restricted to a specific distribution such as Gaussian distribution. However, the specific distribution may be not up to describing the complex low-quality face. Therefore, we propose to transform this specific distribution to a flexible one via a Continuous Normalizing Flow (CNF), which can get rid of the form limitation. This kind of flexible distribution can approximate the latent distribution of the given noisy face more accurately, which further improves accuracy of low-quality 3D FR. Comprehensive experiments show that our proposed method improves both low-quality and cross-quality 3D FR performances on low-quality benchmarks. Furthermore, the improvements are more remarkable on low-quality 3D faces when the intensity of noise increases which indicate the robustness.

Introduction

Due to the rapid development of deep learning and big data techniques, 2D face recognition (FR) has achieved great successes (Deng et al. 2019; Liu et al. 2017; Wang et al. 2018; Schroff, Kalenichenko, and Philbin 2015) and has drawn extensive attention in both academic community and industrial community. However, existing 2D FR systems are still vulnerable to lighting, pose variations, makeup, presentation attacks and so on. 3D FR, which works by comparing facial geometry shapes, is inherently robust to 2D texture related disturbing factors such as different races, shading, face paint, video replay attack, print photo attack, etc.

With the public release of 3D face databases and benchmarks, e.g. FRGC (Phillips et al. 2005), Bosphorus (Savran et al. 2008), BU3D-FE (Yin et al. 2006), a large number of 3D FR methods have been proposed during the past two decades (Gilani, Mian, and Eastwood 2017; Blanz and Vetter 2003; Gilani and Mian 2018) and achieve high accuracies. It should be noted that almost all these methods focus on 3D FR with high-quality 3D face scans captured by high-precision 3D scanners. Although very high performances can be achieved, their 3D FR systems generally suffer from the problems of time-consuming data acquisition, difficulty of moving and high cost, and thus are still far from practical applications.

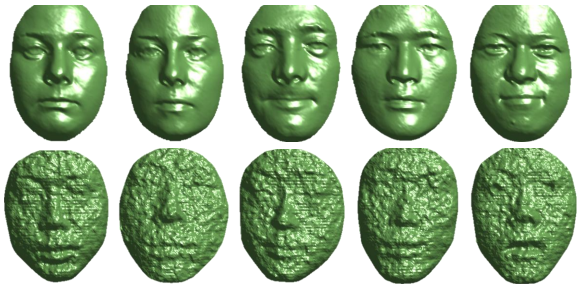

Figure 1: Illustration of high and low quality 3D face scans

Recent years, the popularization of consumer depth cameras such as Kinect and RealSense, paves the way for real applications of 3D FR systems. Comparing with high-precision 3D scanners, consumer depth cameras are much cheaper and smaller, and they have a very high frame rate which is very important for real-time data acquisition and algorithm processing. Unfortunately, the data captured by those consumer depth cameras are often noisy and thus low-quality as shown in Figure 1. Therefore, it is quite necessary to explore noise-robust algorithm for low-quality 3D FR.

There have been some prior studies on low-quality 3D FR. Early explorations use traditional methods (Berretti, Bimbo, and Pala 2012; Goswami et al. 2013; Min, Kose, and Dugelay 2014) such as ICP and PCA get promising results. However, the databases they used are unpractical because of few subjects or variations. Lock3DFace (Zhang et al. 2016) is the first large-scale benchmark which is appropriate for low-quality 3D FR. Later, Cui et al. gives a deep CNN based baseline, while the authors focus on RGB-D FR and use a private dataset to pre-train. In current state-of-the-art method

*Corresponding author.

(Mu et al. 2019), the authors designed sophisticated data augmentations and a novel lightweight deep architecture to achieve better performance in low-quality 3D FR. However, all these existing methods use deterministic representations to describe 3D faces.

For the first time, we propose to *learn flexibly distributional representation for low-quality 3D face recognition*. Noises on a 3D face usually obey a distribution and lead to disturbances on the corresponding feature, thus the feature with disturbances also obeys a latent distribution (i.e., the posterior of the given 3D face). From the view of maximum likelihood, the feature without disturbance (i.e., clean feature) can be inferred from this posterior. Hence, to avoid the influence of noises, we need to learn the posterior to achieve clean feature of the noisy 3D face. Whereas this posterior, namely true posterior, is intractable, thus it is generally approximated by a specific distribution (e.g. parametric Gaussian distribution). Since the form of the specific distribution limits the expressiveness, which causes an inaccurate estimation of the true posterior. For example, true posterior of a 3D face obeys a Laplacian distribution, but we search an approximation within a parametric family of Gaussian distributions. Therefore, we can not infer clean feature precisely from the inaccurate estimation, which is unfavourable to 3D FR. In order to precisely estimate the true posterior, we propose to get rid of the limitation of distribution form by exploiting the Continuous Normalizing Flow (CNF), that reversibly transforms the specific distribution into a flexible form. In summary, the flexibly distributional representation learning consists of two steps: (i) encode each noisy 3D face as a specific distribution, and (ii) transform it by a parameterized CNF to better approximate the true posterior.

The key insight in this paper is that we learn a form-free distribution to approximate the true posterior of a given low-quality 3D face, instead of using a specific distribution. As a result, this kind of form-free posterior leads to a more discriminative feature of the given 3D face, which can further improve the performance of 3D FR. Our contributions include the following three aspects:

- We employ latent distribution to represent a given low-quality 3D face and infer corresponding clean feature for 3D FR.
- We propose to learn a flexible distribution to approximate the true posterior of the noisy 3D face by CNF, rather than the widely used Gaussian distribution.
- Comprehensive experiments demonstrate that our approach can significantly improve accuracies for both low-quality and cross-quality 3D FR. The gaps between performances of our method and baseline in 3D FR with stronger noises indicates the robustness of our approach.

Related and Preliminary Works

Low-quality 3D Face Recognition

In the past two decades, with the releases of different 3D face databases such as FRGC v2.0, BU3D-FE, Bosphorus, 3D FR methods have received widely attention. The 3D FR approaches can be mainly divided into hand-crafted methods and deep learning based methods. Hand-crafted methods (Drira et al. 2013; Kakadiaris et al. 2007; Li et al. 2015; Gupta, Markey, and Bovik 2010; Mian, Bennamoun, and Owens 2008) design local or global shape descriptors to represent geometries of 3D faces. Deep learning based method (Kim et al. 2017; Gilani and Mian 2018; Soltanpour and Wu 2019; Cai et al. 2019; Lin et al. 2019) extract latent features to represent 3D faces. These approaches achieve great recognition results, but they only focus on high-quality 3D FR.

Low-quality 3D FR is a more practical scenario (Hu, Zhao, and Liu 2019), while the researches are limited. Hand-crafted (Berretti, Bimbo, and Pala 2012; Goswami et al. 2013; Min, Kose, and Dugelay 2014; Goswami, Vatsa, and Singh 2014) descriptors such as LGBP, HOG are used in early works while the databases they used contain few subjects and the total number of 3D face scans is quite limited. Zhang et al. firstly provide a public large low-quality 3D face database, and Cui et al. use a private dataset to train an Inception-V2 (Ioffe and Szegedy 2015) for RGB-D FR on Lock3DFace, Mu et al. designs a lightweight deep model and a series of data augmentations to achieve the state-of-the-art on low-quality 3D FR. However, due to the drive of real applications, low-quality 3D FR still needs to be largely explored.

Continuous Normalizing Flow

Normalizing Flow is a series of differentiable and invertible mappings $h_1, ..., h_n$ which can transfer a prior distribution into a more complex distribution. It is popularised in variational inference (Rezende and Mohamed 2015; van den Berg et al. 2018) and density estimation (Dinh, Krueger, and Bengio 2015). Given a random variable $\mathbf{z} \sim p(\mathbf{z})$, the log-density of $\mathbf{z}' = h_n \circ h_{n-1} \circ \cdots \circ h_1(\mathbf{z})$ is:

$$\log p(\mathbf{z}') = \log p(\mathbf{z}) - \sum_{k=1}^{n} \log \mid \det(\frac{\partial h_k}{\partial \mathbf{z}_{k-1}}) \mid \quad (1)$$

Instead of defining the transform as a discrete sequence, Continuous Normalizing Flow (CNF) (Chen et al. 2018) generalizes it to a continuous one. CNF defines a continuous mapping h as the solution of an ordinary differential equation (ODE) $\frac{\partial \mathbf{z}(t)}{\partial t} = h(\mathbf{z}(t), t)$ with the initial value $\mathbf{z}(t_0) = \mathbf{z}_0$, then the log-density of $\mathbf{z}(t_1)$ is

$$\log p(\mathbf{z}(t_1)) = \log p(\mathbf{z}(t_0)) - \int_{t_0}^{t_1} \mathrm{Tr}(\frac{\partial h}{\partial \mathbf{z}(t)}) dt \quad (2)$$

The outputs and gradients of CNF can be computed by a black box ODE solver (Chen et al. 2018; Grathwohl et al. 2019).

Data Uncertainty in 2D Face Recognition

Recently, several 2D FR methods (Shi and Jain 2019; Shi et al. 2020; Chang et al. 2020) introduce data uncertainty learning to deep model to improve the robustness of 2D face representation. Chang et al. adopts a Gaussian distribution to learn data uncertainty for 2D FR. In (Shi and Jain 2019), the authors represent each face image as a distribution, and

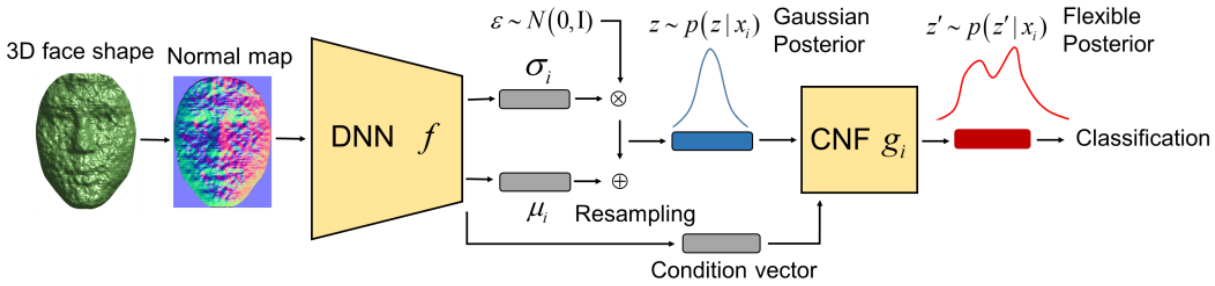

Figure 2: Overview of the proposed method. Normal map of a given 3D face is firstly encoded as a latent distribution (i.e. a Gaussian posterior $p(\mathbf{z}|\mathbf{x}_i)$) by a deep model. Then, a CNF is introduced to change the density of Gaussian posterior into a flexible posterior $p(\mathbf{z}'|\mathbf{x}_i)$. Once it is learned, the most likely sample in $p(\mathbf{z}'|\mathbf{x}_i)$ is selected for low-quality 3D FR

propose a metric between distributions to classify 2D face images. Shi et al. design a confidence-aware loss for 2D FR by considering data uncertainty. Experimental results show that these methods can reduce the influence of label noise, image blurring and other uncertainty factors, to a certain extent. However, as far as we know, there is no similar work on 3D FR so far.

Proposed Method

The aim of our method is to overcome the influence of noises and achieve robust representation of low-quality 3D faces. In Section , we embed low-quality 3D faces into latent distributions. In Section , we propose flexibly distributional representation of low-quality 3D faces by CNF. And finally, objective of our learning processes is draw. The overview of the proposed method is shown in Figure 2.

Distributional Representation for Low-quality 3D Faces

After encoded by deep networks, the noises on low-quality 3D faces cause noises on features of 3D faces. In particular, given an observed low-quality 3D face $\mathbf{x}_i$, we assume the noisy latent feature $\mathbf{z}$ obeys a posterior $q(\mathbf{z}|\mathbf{x}_i)$, and

$$\mathbf{z} = z^c + \mathbf{n} \tag{3}$$

where z^c is the clean feature of $\mathbf{x}_i$, $\mathbf{n}$ is the encoded noise. The samples $\{\mathbf{z}_j\}$ in $q(\mathbf{z}|\mathbf{x}_i)$ can be regarded as multiple observed results of z^c. From the view of maximum likelihood, z^c can be inferred as the most likely sample in $q(\mathbf{z}|\mathbf{x}_i)$. That is, we can choose the most likely sample in $q(\mathbf{z}|\mathbf{x}_i)$ as the clean feature z^c.

In general, the true posterior $q(\mathbf{z}|\mathbf{x}_i)$ is intractable, but can be approximated by some specific distribution $p(\mathbf{z}|\mathbf{x}_i)$, such as Gaussian distribution $\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2\mathbf{I})$ with learnable parameters mean $\boldsymbol{\mu}_i$ and variance $\boldsymbol{\sigma}_i^2$.

To make the training process differentiable and obtain the samples from $p(\mathbf{z}|\mathbf{x}_i)$, the re-parameterization trick (Kingma and Welling 2014) can be adopted.

$$\mathbf{z} = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}\boldsymbol{\sigma}_i, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{4}$$

The clean feature z^c of the given noisy 3D face can be approximated by the most likely sample (MLS) in the approximated distribution (i.e. $\boldsymbol{\mu}_i$ in Gaussian distribution). However, it is usually not realistic that regard the true posterior

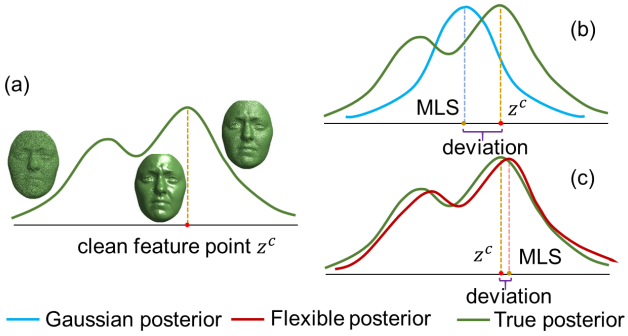

Figure 3: Comparisons of flexible posterior and Gaussian posterior. (a) Noisy feature obey a distribution $q(\mathbf{z}|\mathbf{x}_i)$ (i.e.true posterior). The clean feature can be treated as the most likely sample in $q(\mathbf{z}|\mathbf{x}_i)$. (b) Estimating the true posterior by a Gaussian distribution. (c) Estimating the true posterior by a flexible distribution. A specific distribution (Gaussian in this figure) can not approximate the true posterior accurately, which leads to a deviation between clean feature and MLS. This deviation can be reduced by a flexibly distributional representation

$q(\mathbf{z}|\mathbf{x}_i)$ as a Gaussian distribution. Actually, no matter which family of distributions we adopt to estimate the true posterior, it may not fall into the adopted family. The inaccurate estimation of the true posterior results in that the MLS in $p(\mathbf{z}|\mathbf{x}_i)$ away from the clean feature z^c. Thus, a flexible posterior is needed.

Flexibly Distributional Representation for Low-quality 3D Faces

As shown in Figure 3, there is a deviation between MLS of the Gaussian estimation and the clean feature of the true posterior. If the estimated posterior is closer to the true posterior, the deviation will be reduced. Thus precise estimation of the true posterior $q(\mathbf{z}|\mathbf{x}_i)$ is the key in learning distributional representation of 3D faces.

To push the MLS of estimated posterior towards the clean feature, we propose to model $q(\mathbf{z}|\mathbf{x}_i)$ via a Continuous Normalizing Flow (CNF) to flexibly estimate the true posterior. Figure 3 illustrates difference between Gaussian posterior

and our proposed flexible posterior. Unlike Gaussian distribution, which is restricted by form, our flexible posterior can approximate the true posterior more precisely.

The strategy of flexibly estimating the true posterior is described as follows. First, a latent variable is drawn from a parametric posterior distribution such as a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i)$. The sample of this Gaussian distribution is then transformed by a CNF g_i. After applying a continuous-time dynamics, the latent variable is given by:

$$\begin{aligned} \mathbf{z}' = g_i(\mathbf{z}(t_0)) = \mathbf{z}(t_0) + \int_{t_0}^{t_1} h_i(\mathbf{z}(t), t)dt, \\ \mathbf{z}(t_0) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) \end{aligned} \tag{5}$$

where $\mathbf{z} = \mathbf{z}(t_0)$ and $\mathbf{z}' = \mathbf{z}(t_1)$. Since each 3D face $\mathbf{x}_i$ has a corresponding true posterior, the transformation function CNF should vary according to $\mathbf{x}_i$. Therefore, we define the continuous-time dynamics conditioned on $\mathbf{x}_i$ as h_i and use another extractor $c(\cdot)$ to obtain the condition vector $c = c(\mathbf{x}_i)$.

$$h_i(\mathbf{z}(t), t) = h(\mathbf{z}(t), t, c), \quad g_i(\mathbf{z}(t_0)) = g(\mathbf{z}(t_0), c) \tag{6}$$

Hence, the log-density of the transferred latent variable is

$$\log p(\mathbf{z}'|\mathbf{x}_i) = \log p(\mathbf{z}(t_0)|\mathbf{x}_i) - \int_{t_0}^{t_1} \mathrm{Tr}(\frac{\partial h_i}{\partial \mathbf{z}(t)})dt \tag{7}$$

In this way, the Gaussian distribution is transformed into a form-free posterior. This transformation helps our network learn a more flexible posterior $p(\mathbf{z}'|\mathbf{x}_i)$.

In conclusion, we use $p(\mathbf{z}'|\mathbf{x}_i)$ as a better approximation to the true posterior $q(\mathbf{z}|\mathbf{x}_i)$, and the clean feature z^c of the given noisy 3D face can be approximated by the MLS in the flexibly distributional representation $p(\mathbf{z}'|\mathbf{x}_i)$.

Objective for Low-quality 3D FR

As described above, when we obtain the approximated posterior $p(\mathbf{z}'|\mathbf{x}_i)$ of a given 3D face $\mathbf{x}_i$, we can achieve the MLS to improve the recognition accuracy. Therefore, our goal is to determine parameters of the estimated posterior $p(\mathbf{z}'|\mathbf{x}_i)$, which can be solved by maximum posterior estimation

$$\arg\max_{f, g_i} p(\mathbf{z}'|\mathbf{x}_i) \tag{8}$$

where f and g_i represent the deep model and CNF respectively. And the maximum posterior estimation can be regarded as the following empirical risk minimization:

$$\min_{F} \sum_{\mathbf{x}_i \in \mathcal{X}} L(y_i, F(\mathbf{x}_i))) \tag{9}$$

where F denotes the entire recognition model, consisting of a feature extractor f, a CNF and a classifier cls. L is a classification loss and y_i is the label of $\mathbf{x}_i$.

In practice, in order to enforce the loss, i.e., Eq.(9), to be lower, deep networks tend to force the variance of the estimated posterior to be zero, that would cause collapse of the distribution and the posterior cannot be estimated. In contrast, a large variance will disrupt the identity information. Therefore we obey the widely used assumption in variational auto-encoder (Kingma and Welling 2014), and add the prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ to constrain the distribution learning.

$$\min_{F} \sum_{\mathbf{x}_i \in \mathcal{X}} L(y_i, F(\mathbf{x}_i)) + \lambda D_{KL}(p(\mathbf{z}'|\mathbf{x}_i)||\mathcal{N}(\mathbf{0}, \mathbf{I})) \tag{10}$$

where λ is a trade-off parameter. For the first term, samples of $p(\mathbf{z}'|\mathbf{x}_i)$ are used to minimize the classification loss. And the KL divergence in the second term acts as a regularization term, which can be computed by plugging Eq.(7) into Eq.(10) as follows:

$$\begin{aligned} &D_{KL}(p(\mathbf{z}'|\mathbf{x}_i)||\mathcal{N}(\mathbf{0}, \mathbf{I}))) \\ &= \mathbb{E}_{p(\mathbf{z}'|\mathbf{x}_i)}[\log p(\mathbf{z}'|\mathbf{x}_i) - \log \mathcal{N}(\mathbf{0}, \mathbf{I})] \\ &= -\frac{1}{2}(\log \boldsymbol{\sigma}_i^2 + 1) + \mathbb{E}_{p(\mathbf{z}'|\mathbf{x}_i)}[\frac{\mathbf{z}'^2}{2} - \int_{t_0}^{t_1} \mathrm{Tr}(\frac{\partial h_i}{\partial \mathbf{z}(t)})dt] \end{aligned} \tag{11}$$

Thus, the final loss function can be written as:

$$\begin{aligned} \min_{f, g_i} \sum_{\mathbf{x}_i \in \mathcal{X}} & L(y_i, cls(g_i(f(\mathbf{x}_i)), c)) - \lambda \frac{1}{2}(\log \boldsymbol{\sigma}_i^2 + 1) \\ & + \lambda \mathbb{E}_{p(\mathbf{z}'|\mathbf{x}_i)}[\frac{\mathbf{z}'^2}{2} - \int_{t_0}^{t_1} \mathrm{Tr}(\frac{\partial h_i}{\partial \mathbf{z}(t)})dt] \end{aligned} \tag{12}$$

In 3D face recognition task, we use the cross entropy as the classification loss. By minimizing the final loss function, flexible distribution of each noisy 3D face can be learned. In the training phase, samples from $p(\mathbf{z}'|\mathbf{x}_i)$ are used to compute the final loss. While in the testing phase, we select the MLS of $p(\mathbf{z}'|\mathbf{x}_i)$ as an approximation to the clean feature of the given noisy 3D face. Even though $p(\mathbf{z}'|\mathbf{x}_i)$ can not be represented explicitly, invertibility is equivalent to monotony for the continuous function $g_i(\cdot)$. So the MLS of $p(\mathbf{z}'|\mathbf{x}_i)$ can be obtained by sampling $\boldsymbol{\mu}_i$ from Gaussian distribution and transform it through the CNF.

Owing to the re-parameterization trick, the flexibly distributional representation can also be viewed as an effective data augmentation that acts on feature space rather than data space. Hand-crafted data augmentations (Mu et al. 2019) only contain limited variations (e.g., pose, scale) and the model trained by these data can only deal with corresponding variations. Our approach adaptively learn data augmentation from input data, which guarantees better generalization.

Experiments

Implementation Details

As shown in Figure 2, the normal maps calculated from the low-quality 3D faces are fed into the deep model. For fair comparison, the backbone of our deep model is the same as Led3D (Mu et al. 2019), i.e., a five-layer CNN architecture with an attention block and four shortcut blocks. Our deep model is optimized by the Adam with batch size of 300 for all 100 epochs. The learning rate is initially set to 1e-4 and reduced by a factor of 10 per 3000 iterations. We extract a vector c of 10-dimension from the last feature map of Led3D network by a fully connected layer as the condition vector of CNF. The CNF module consists of 3 fully connected layers with softplus activations. λ in Eq.(12) is chosen as 5e-3.

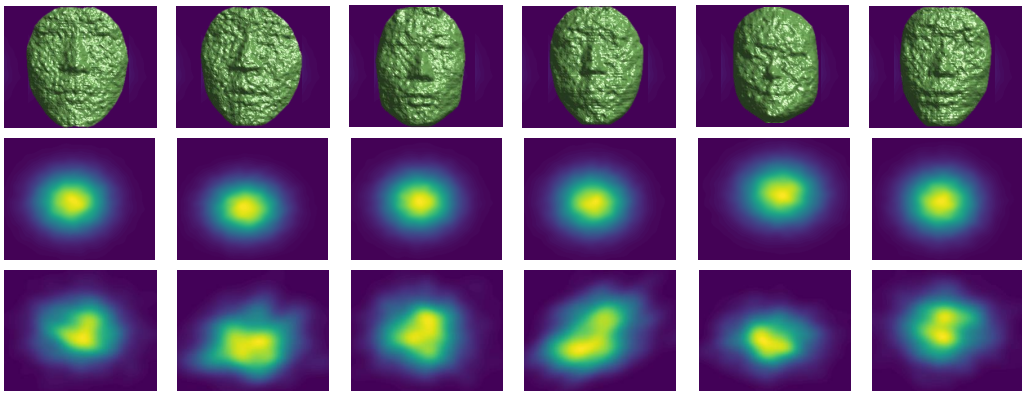

Figure 4: Illustration of the flexibly distributional representations. The first row shows six noisy 3D faces from different subjects, the second row displays their corresponding Gaussian posterior densities and the third row displays flexible posterior densities. The brighter the color, the higher the probability density

3D Face Databases

To show the effectiveness of proposed method, two low-quality databases, i.e., Lock3DFace (Zhang et al. 2016) and IIIT-D (Goswami et al. 2013) and four high-quality databases, namely FRGC v2 (Phillips et al. 2005), Bosphorus (Savran et al. 2008), BU3D-FE (Yin et al. 2006) and BU4D-FE (Yin et al. 2008) are used in our experiments.

Lock3DFace. It is the largest and most comprehensive low-quality 3D face database. The 3D faces are collected by Kinect V2 and include 5,671 videos of 509 subjects. Each subject has neutral expression (NU) and four variations: facial expression (FE), occlusion (OC), pose (PS) and time (TM). It is the most suitable database for low-quality 3D FR and we mainly evaluate our method on it.

IIIT-D. It has 4603 depth maps of 106 subjects, which were captured by Kinect V1 with moderate pose and expression variations.

FRGC v2. FRGC v2 consists of 4,007 3D face scans of 466 subjects, with expression variations.

Bosphorus. It includes 4,666 3D faces of 105 subjects with expression, occlusion and pose variations.

BU3D-FE. It contains 2,500 scans of 100 subjects. And each subject has one neutral scan and six expression variations of four intensity levels.

BU4D-FE. It contains 101 subjects and each one has 6 videos. Each video in BU4D-FE has more than 100 frames.

3D FR on Real Low-quality Data

Protocol I and Results on Lock3DFace. For fair comparison, we conduct the same data augmentations (i.e. 12 variations of pose generating, shape jittering, and shape scaling) as Mu et al. and demonstrated experiments under two experimental protocols. The first protocol is the same as Mu et al.. That is, For each subject, six frames are sampled from the first neural expression depth videos with an equal interval, and are used for data augmentation. The training set contains totally 39,702 (3054×12+3054) frames. The other videos of four types (FE, OC, PS and TM) are used as the test subsets. In the test phase, all frames of the remaining four subsets are fed into the deep model, and the entire video labels are predicted by simple voting. Our model and all the baselines are pre-trained on the FRGC v2 and Bosphorus databases, and then fine-tuned on the training set. Cui et al. pre-train a Inception-V2 on their private data which cause we cannot compare with them fairly, so we report the results of Inception-V2 trained under our setting.

	Test subset				
	FE	OC	PS	TM	AVG
VGG16 (2015)	79.63	36.95	21.7	12.84	42.8
ResNet34 (2016)	62.83	20.32	22.56	2.07	32.23
Inception-V2 (2015)	80.48	32.17	33.23	12.54	44.77
MobileNet-V2 (2018)	85.38	32.77	28.3	10.6	44.92
Led3D (2019)	86.94	48.01	37.67	26.12	54.28
Ours	**92.38**	**49.30**	**43.34**	**31.80**	**58.68**

Table 1: Comparison of the accuracies on Lock3DFace

Table 1 shows the experimental results of Protocol I. From this table, we can see that our proposed flexibly distributional representation can significantly outperform all existing state-of-the-art methods over all the four test subsets. In particularly, compared with Led3D (Mu et al. 2019), we use the same backbone network and data augmentations, but achieve more than 4% improvements on average. These results demonstrate that our proposed representation is more suitable for low-quality 3D FR, the insight analyze presented at Section 4.6.

Protocol II and Results on Lock3DFace. The second protocol divides training and test set according to subjects. We respectively select all the frames of the first 100, 200, 300, and 400 subjects as the training set and the remaining 409, 309, 209, and 109 subjects are used for test. In the test phase, the gallery set is composed of the first frame of the first neutral expression video of each subject, and all the frames and videos of the four subsets (FE, OC, PS and TM) in the test set are used as the probes. We denote the four divisions as A, B, C and D. For fair comparison, we also select six frames of the first neutral expression videos in the

Division	Evaluation type	Method	Test subset				
			FE	OC	PS	TM	AVG
A	Video	Inception-V2 (2015)	89.95	29.71	24.48	10.00	41.38
		Led3D (2019)	89.13	48.89	34.14	12.87	52.02
		Ours(Gaussian posterior)	91.89	52.93	35.74	22.57	55.48
		Ours	**94.11**	**58.19**	**37.34**	**23.03**	**58.72**
	Frame	Inception-V2 (2015)	77.15	27.46	23.22	8.04	38.56
		Led3D (2019)	86.05	45.58	31.84	11.73	48.66
		Ours(Gaussian posterior)	91.54	51.19	33.13	19.46	53.84
		Ours	**92.72**	**54.85**	**34.18**	**20.17**	**55.52**
B	Video	Inception-V2 (2015)	81.26	33.96	28.04	8.99	42.49
		Led3D (2019)	88.14	51.81	36.95	11.73	50.49
		Ours(Gaussian posterior)	93.11	53.72	28.69	25.76	54.12
		Ours	**94.47**	**57.77**	**36.30**	**27.90**	**57.65**
	Frame	Inception-V2 (2015)	77.67	33.96	26.27	7.52	39.95
		Led3D (2019)	85.30	47.70	32.80	11.07	47.74
		Ours(Gaussian posterior)	92.07	52.79	27.63	23.40	52.79
		Ours	**93.16**	**56.71**	**32.98**	**23.13**	**55.13**
C	Video	Inception-V2 (2015)	89.57	49.76	38.27	39.21	58.94
		Led3D (2019)	94.32	59.56	38.03	42.10	63.32
		Ours(Gaussian posterior)	95.55	67.94	38.99	51.98	67.61
		Ours	**96.17**	**70.57**	**41.87**	**58.16**	**70.56**
	Frame	Inception-V2 (2015)	87.89	47.87	36.52	34.58	56.60
		Led3D (2019)	93.52	56.77	36.34	37.94	61.10
		Ours(Gaussian posterior)	94.91	66.07	37.23	47.53	65.91
		Ours	**95.70**	**67.56**	**40.56**	**51.76**	**68.13**
D	Video	Inception-V2 (2015)	89.16	53.75	43.95	42.00	65.43
		Led3D (2019)	94.30	64.22	40.82	59.00	70.28
		Ours(Gaussian posterior)	96.47	69.72	40.83	55.00	72.04
		Ours	**96.48**	**73.39**	**45.87**	**60.00**	**74.48**
	Frame	Inception-V2 (2015)	88.81	52.84	42.43	39.14	63.49
		Led3D (2019)	93.22	63.44	39.03	53.91	68.65
		Ours(Gaussian posterior)	95.30	67.64	39.44	50.74	70.26
		Ours	**95.52**	**69.08**	**43.64**	**54.36**	**71.89**

Table 2: Comparison results on different divisions of Lock3DFace

training set and take the same data augmentations with Mu et al.. All the baseline models and ours are pre-trained on the FRGC v2 and Bosphorus databases with additive Gaussian noise of zero mean and variance of 16.

Table 2 reports that our method achieves the state-of-the-art performances on both video based and frame based FR. Another important conclusion is that the flexible posterior estimated by CNF outperforms the Gaussian posterior in almost all subsets, which indicates that the flexible posterior can better approximate the true posterior than the simple Gaussian posterior. We also visualize the learned Gaussian posterior and the flexible posterior of given noisy 3D faces in Figure 4. Form the figure, we can find that without the form limitation, CNF helps us learn a more flexible density.

Protocol and Results on IIIT-D. IIIT-D database contains 106 subjects, and we choose all 3D faces of the first 53 subjects for model training and faces of the remaining 53 subjects for model evaluation. Similar to the Protocol II on Lock3DFace, all the baseline models and ours are pre-trained on the noisy FRGC v2 and Bosphorus databases. From table 3, we can see that our proposed method can largely improve the recognition accuracies, which indicates the good generalization.

Method	Inception-V2 (2015)	Led3D(2019)	Ours
Accuracy	65.58	74.27	**80.28**

Table 3: Recognition results on IIIT-D

3D FR on Synthetic Low-quality Data

Due to the real low-quality 3D face databases are rare, we also add random noises to high-quality data to simulate low-quality 3D faces. Specifically, we add Gaussian noise with a zero mean and variance of 16 to the FRGC v2 and Bosphorus then combine them with the original two datasets as the training data. All the noises are added on z-value only. To evaluate the robustness of the proposed method, for test data, we add Gaussian additive noises with different variances (i.e., 4, 8, 16, 32, 64) to synthesize noisy BU3D-FE and BU4D-FE databases. In each synthetic BU3D-FE database, for each subject, a face with neutral expression is selected as the gallery and the rest faces are probes. In each synthetic BU4D-FE database, for each subject, the first frame of the first video is used as the gallery sample and all the rest frames are used as probes.

σ^2	BU-3DFE Led3D (2019)	BU-3DFE Ours	BU-4DFE Led3D (2019)	BU-4DFE Ours
4	95.08	**96.88**	90.07	**94.87**
8	92.87	**95.42**	84.83	**90.85**
16	87.54	**91.72**	72.77	**81.09**
32	74.66	**82.33**	50.51	**61.32**
64	55.37	**60.37**	28.73	**36.80**

Table 4: Recognition results on synthesis noisy data

σ^2	BU-3DFE Led3D (2019)	BU-3DFE Ours	BU-4DFE Led3D (2019)	BU-4DFE Ours
4	92.58	**96.21**	89.07	**93.99**
8	92.04	**96.00**	86.89	**92.01**
16	88.20	**93.92**	80.79	**86.96**
32	81.00	**88.13**	63.77	**75.00**
64	63.45	**71.25**	33.26	**46.50**

Table 5: Recognition results for cross-quality 3D FR

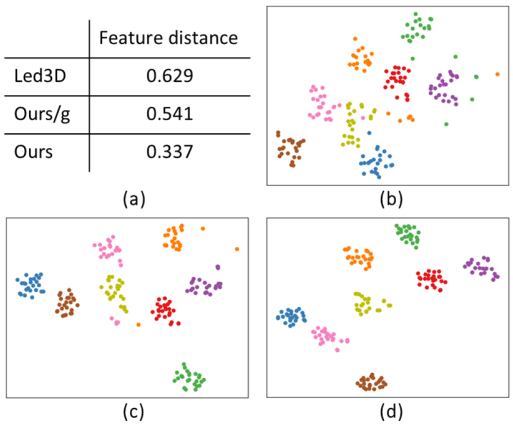

Figure 5: (a) Averaged feature distances between high-low quality pairs, ours/g mean Gaussian posterior representation. (b-d) Visualization of laten feature clustering results of Led3D, Gaussian distributional representation, and flexibly distributional representation

Table 4 reports the recognition accuracies of Led3D (Mu et al. 2019) and our method on the synthetic low-quality BU3D-FE and BU4D-FE databases. Whether the intensity of noise on the test set is strong or weak than the noise on training set, our model consistently performs better than baseline. Especially faced with stronger noise, the gaps between the performances of our model and baseline become widen. This phenomenon exactly illustrates that our model is expert in learning robust representations of noisy 3D faces, and it is important for low-quality 3D FR.

Cross-quality 3D FR

In practical applications, a commonly used 3D FR setting is that high-quality data as galleries and low-quality data are used as probes. The reason is that high-precision acquisition equipments are not convenient for real-time 3D FR, while consumer 3D depth cameras are efficient but acquired data are usually noisy.

To simulate this situation, we use the same training set described in Section 4.4. We evaluate on the BU3D-FE and BU4D-FE databases. The way we generate low-quality 3D face probes is the same as we did in Section 4.4. And the neutral scan of the original BU3D-FE and the first frame of the first video of the original BU4D-FE are used as the high-quality galleries, respectively.

Table 5 shows the accuracies for cross-quality 3D FR. Similar to the synthetic low-quality data experiment, our model still significantly exceeds the baseline method in all cases. All the results further prove that our method can achieve more robust representations of noisy 3D faces.

Understanding of the Flexibly Distributional Representation

In this section, we will analyze the effect of distributional representation and flexibly distributional representation insightfully.

To verify our proposed method can reduce influence of noises, we compare the distances between high-low quality faces in feature space. Firstly we add noises on high-quality 3D faces of BU3D-FE database to obtain pairs of high-low quality 3D faces. Then, We extract features of high-low quality data pairs by using our model and Led3D (Mu et al. 2019) respectively, and MLS is chosen as the feature of a given 3D face in our model specially. From Figure 5 (a), we can find that whether Gaussian or flexible posterior is selected, the average euclidean distance between features of high-low quality faces is smaller than that of Led3D (Mu et al. 2019). Furthermore, MLS of flexible posterior is closer to feature of high-quality faces, compared with Gaussian posterior. This phenomenon indicate that our representation plays a role of feature denoising, which partly verifies our claim in Figure 3.

For 3D FR task, we need features of 3D faces to be compact and discriminative. Features extracted by Led3D, Gaussian posterior distribution, and flexible posterior distribution are clustered respectively. As shown in Figure 5, among them, our proposed flexible posterior is favor of learning more compact features, which indicates that our method can achieve more discriminative features for low-quality 3D FR.

Conclusion

In this paper, we propose a novel flexibly distributional representation for low-quality 3D FR. Different from the normally used distribution such as Gaussian distribution, we breakthrough the limitation of distribution form. The specific posterior can be transformed into a flexible one via CNF, which benefits a better estimation of the true posterior. Based on more robust representation, the proposed method achieves the state-of-the-art results on several low-quality and cross-quality 3D FR tasks.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61976173.

References

Berretti, S.; Bimbo, A. D.; and Pala, P. 2012. Superfaces: A Super-Resolution Model for 3D Faces. In Fusiello, A.; Murino, V.; and Cucchiara, R., eds., *Computer Vision - ECCV. Workshops and Demonstrations*.

Blanz, V.; and Vetter, T. 2003. Face Recognition Based on Fitting a 3D Morphable Model. *IEEE Trans. Pattern Anal. Mach. Intell* .

Cai, Y.; Lei, Y.; Yang, M.; You, Z.; and Shan, S. 2019. A fast and robust 3D face recognition approach based on deeply learned face representation. *Neurocomputing* .

Chang, J.; Lan, Z.; Cheng, C.; and Wei, Y. 2020. Data Uncertainty Learning in Face Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*.

Chen, T. Q.; Rubanova, Y.; Bettencourt, J.; and Duvenaud, D. 2018. Neural Ordinary Differential Equations. In *Annual Conference on Neural Information Processing Systems, NeurIPS*.

Cui, J.; Zhang, H.; Han, H.; Shan, S.; and Chen, X. 2018. Improving 2D Face Recognition via Discriminative Face Depth Estimation. In *International Conference on Biometrics, ICB*.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

Dinh, L.; Krueger, D.; and Bengio, Y. 2015. NICE: Nonlinear Independent Components Estimation. In *3rd International Conference on Learning Representations, ICLR*.

Drira, H.; Amor, B. B.; Srivastava, A.; Daoudi, M.; and Slama, R. 2013. 3D Face Recognition under Expressions, Occlusions, and Pose Variations. *IEEE Trans. Pattern Anal. Mach. Intell* .

Gilani, S. Z.; and Mian, A. 2018. Learning From Millions of 3D Scans for Large-Scale 3D Face Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

Gilani, S. Z.; Mian, A. S.; and Eastwood, P. R. 2017. Deep, dense and accurate 3D face correspondence for generating population specific deformable models. *Pattern Recognit* .

Goswami, G.; Bharadwaj, S.; Vatsa, M.; and Singh, R. 2013. On RGB-D face recognition using Kinect. In *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems, BTAS*.

Goswami, G.; Vatsa, M.; and Singh, R. 2014. RGB-D Face Recognition With Texture and Attribute Features. *IEEE Trans. Inf. Forensics Secur* .

Grathwohl, W.; Chen, R. T. Q.; Bettencourt, J.; Sutskever, I.; and Duvenaud, D. 2019. FFJORD: Free-Form Continuous Dynamics for Scalable Reversible Generative Models. In *7th International Conference on Learning Representations, ICLR*.

Gupta, S.; Markey, M. K.; and Bovik, A. C. 2010. Anthropometric 3D Face Recognition. *Int. J. Comput. Vis* .

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

Howard, A.; Zhmoginov, A.; Chen, L.-C.; Sandler, M.; and Zhu, M. 2018. Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

Hu, Z.; Zhao, Q.; and Liu, F. 2019. Revisiting Depth-Based Face Recognition From a Quality Perspective. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops*.

Ioffe, S.; and Szegedy, C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning, ICML*.

Kakadiaris, I. A.; Passalis, G.; Toderici, G.; Murtuza, M. N.; Lu, Y.; Karampatziakis, N.; and Theoharis, T. 2007. Three-Dimensional Face Recognition in the Presence of Facial Expressions: An Annotated Deformable Model Approach. *IEEE Trans. Pattern Anal. Mach. Intell* .

Kim, D.; Hernandez, M.; Choi, J.; and Medioni, G. G. 2017. Deep 3D face identification. In *IEEE International Joint Conference on Biometrics, IJCB*.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR*.

Li, H.; Huang, D.; Morvan, J.; Wang, Y.; and Chen, L. 2015. Towards 3D Face Recognition in the Real: A Registration-Free Approach Using Fine-Grained Matching of 3D Keypoint Descriptors. *International Journal of Computer Vision* .

Lin, S.; Liu, F.; Liu, Y.; and Shen, L. 2019. Local Feature Tensor Based Deep Learning for 3D Face Recognition. In *14th IEEE International Conference on Automatic Face & Gesture Recognition, FG*.

Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; and Song, L. 2017. SphereFace: Deep Hypersphere Embedding for Face Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

Mian, A. S.; Bennamoun, M.; and Owens, R. A. 2008. Keypoint Detection and Local Feature Matching for Textured 3D Face Recognition. *Int. J. Comput. Vis.* .

Min, R.; Kose, N.; and Dugelay, J. 2014. KinectFaceDB: A Kinect Database for Face Recognition. *IEEE Trans. Syst. Man Cybern. Syst* .

Mu, G.; Huang, D.; Hu, G.; Sun, J.; and Wang, Y. 2019. Led3D: A Lightweight and Efficient Deep Approach to Recognizing Low-Quality 3D Faces. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

Phillips, P. J.; Flynn, P. J.; Scruggs, W. T.; Bowyer, K. W.; Chang, J.; Hoffman, K.; Marques, J.; Min, J.; and Worek, W. J. 2005. Overview of the Face Recognition Grand Challenge. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*.

Rezende, D. J.; and Mohamed, S. 2015. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on Machine Learning, ICML*.

Savran, A.; Alyüz, N.; Dibeklioglu, H.; Çeliktutan, O.; Gökberk, B.; Sankur, B.; and Akarun, L. 2008. Bosphorus Database for 3D Face Analysis. In *Biometrics and Identity Management, First European Workshop, BIOID*.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

Shi, Y.; and Jain, A. K. 2019. Probabilistic Face Embeddings. In *IEEE/CVF International Conference on Computer Vision, ICCV*.

Shi, Y.; Yu, X.; Sohn, K.; Chandraker, M.; and Jain, A. K. 2020. Towards Universal Representation Learning for Deep Face Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR*.

Soltanpour, S.; and Wu, Q. J. 2019. Weighted Extreme Sparse Classifier and Local Derivative Pattern for 3D Face Recognition. *IEEE Trans. Image Process* .

van den Berg, R.; Hasenclever, L.; Tomczak, J. M.; and Welling, M. 2018. Sylvester Normalizing Flows for Variational Inference. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI*.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018. CosFace: Large Margin Cosine Loss for Deep Face Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

Yin, L.; Chen, X.; Sun, Y.; Worm, T.; and Reale, M. 2008. A high-resolution 3D dynamic facial expression database. In *8th IEEE International Conference on Automatic Face and Gesture Recognition FG*.

Yin, L.; Wei, X.; Sun, Y.; Wang, J.; and Rosato, M. J. 2006. A 3D Facial Expression Database For Facial Behavior Research. In *Seventh IEEE International Conference on Automatic Face and Gesture Recognition FGR*.

Zhang, J.; Huang, D.; Wang, Y.; and Sun, J. 2016. Lock3DFace: A large-scale database of low-cost Kinect 3D faces. In *International Conference on Biometrics, ICB*.