Enhancing Face Recognition with Self-Supervised 3D Reconstruction

Mingjie He[1,2], Jie Zhang[1,2], Shiguang Shan[1,2,3], Xilin Chen[1,2]
[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
[3]Peng Cheng Laboratory, Shenzhen, 518055, China
{hemingjie, zhangjie, sgshan, xlchen}@ict.ac.cn

Abstract

Attributed to both the development of deep networks and abundant data, automatic face recognition (FR) has quickly reached human-level capacity in the past few years. However, the FR problem is not perfectly solved in case of uncontrolled illumination and pose. In this paper, we propose to enhance face recognition with a bypass of self-supervised 3D reconstruction, which enforces the neural backbone to focus on the identity-related depth and albedo information while neglects the identity-irrelevant pose and illumination information. Specifically, inspired by the physical model of image formation, we improve the backbone FR network by introducing a 3D face reconstruction loss with two auxiliary networks. The first one estimates the pose and illumination from the input face image while the second one decodes the canonical depth and albedo from the intermediate feature of the FR backbone network. The whole network is trained in end-to-end manner with both classic face identification loss and the loss of 3D face reconstruction with the physical parameters. In this way, the self-supervised reconstruction acts as a regularization that enables the recognition network to understand faces in 3D view, and the learnt features are forced to encode more information of canonical facial depth and albedo, which is more intrinsic and beneficial to face recognition. Extensive experimental results on various face recognition benchmarks show that, without any cost of extra annotations and computations, our method outperforms state-of-the-art ones. Moreover, the learnt representations can also well generalize to other face-related downstream tasks such as the facial attribute recognition with limited labeled data.

1. Introduction

With abundant data [2,8] and the development of margin-based loss functions [5, 17, 30, 31], great progresses have been achieved on face recognition. The accuracy on the famous LFW benchmark [10] is almost perfect, e.g., 99.83% by ArcFace [5]. Even though, existing methods degenerate severely under large poses, various illuminations and partial occlusions. Face recognition under unconstrained scenarios still remains a challenging task. Unfortunately, conventional methods with fixed-margins do not fit well with these factors as they do not consider the difficultness of each sample, and may result in convergence issue. To alleviate this problem, AdaptiveFace [16], AdaCos [39], MagFace [20] and CurricularFace [12] propose to dynamically tune the margin during the training process. MagFace [20] proposes a quality assessment method to evaluate the face quality and then adaptively adjust its margins. To some extent, these adaptive margin methods have alleviated the convergence issue. However, the performance still degenerates when testing under large poses, various illuminations and occlu-

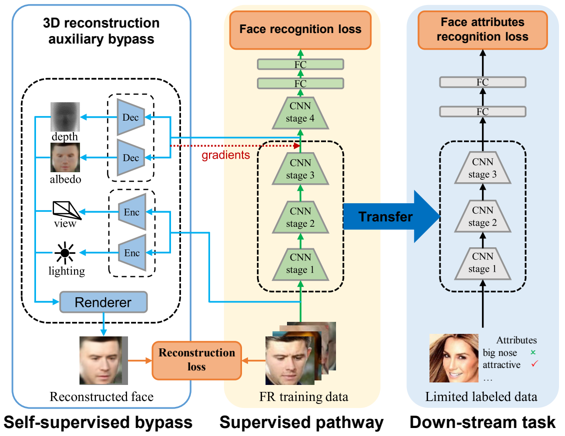

Figure 1. Illustration of the proposed 3D bypass enhanced representation learning (3D-BERL). Inspired by the physical model of image formation, 3D-BERL incorporates a self-supervised 3D reconstruction bypass to improve the face recognition. Besides, the shallow stages of the backbone can also serve as a good foundation model for downstream tasks such as facial attribute recognition with only limited labeled data.

sions as they do not explicitly consider kicking out identity-irrelevant features during training process.

An ideal solution is to annotate all the image factors of each face, and then disentangle these identity-irrelevant factors from the face embedding via multi-task supervised-learning. However, this kind of annotation is labor-intensive and is almost infeasible, especially when current face recognition training sets commonly contain millions of faces. To be free from this, we resort to another technology roadmap, namely self-supervise learning. Apart from supervised learning methods, it does not require manual annotations. Instead, the image representations are learnt via pretext tasks [7, 22] which commonly apply a transformation to the image and then enforce the neural network to predict the properties of the transformation. The self-supervised learning has shown the potential to become an alternative approach to learn feature representation. It even outperforms supervised learning methods on image classification task. However, the work in [27] has shown that the commonly used pretext tasks such as jigsaw puzzle [22] and rotation prediction [7] do not work well on face-related task. The possible reason is that it is hard to prevent existing pretext tasks collapse to trivial solutions for face images which have unified structure and similar textures.

In this paper, we propose a novel 3D Bypass Enhanced Representation Learning (3D-BERL) method to improve face recognition under unconstrained scenarios. As shown in Fig 1, our 3D-BERL incorporates an auxiliary bypass of self-supervised 3D face reconstruction into traditional 2D face recognition pathway. Inspired by the physical model of image formation, we carefully design two auxiliary networks in the auxiliary bypass. The first one estimates the identity-irrelevant viewpoint (pose) and illumination parameter, and the second one decodes the canonical depth and albedo from the intermediate block of existing face recognition backbone (the stage 3 in ResNet). The pose, illumination, depth and albedo are learnt by self-supervised 3D reconstruction, and the reconstruction process is based on the physical model of image formation. Among these four factors, the learning of depth and albedo will enforce the shallow part of FR backbone (the stage 1, 2, 3 of ResNet [9]) to focus on the identity-relevant depth and albedo. Besides, as shown in our visualization in Fig 3, the image formation model has regularized both the depth and albedo to with a canonical view, meaning that the effect of pose and illumination is eliminated. Moreover, even if the face image is occluded, the prediction of depth is still good enough for face recognition. Then, the succeeding layers in the FR backbone can extract features robust to pose, illumination and partial occlusion. Here, the self-supervised reconstruction acts as a regularization that enables the recognition network to understand faces in 3D view. The whole network is trained in end-to-end manner with both classic face identification loss and the loss of 3D face reconstruction with the physical parameters. Extensive results on various face recognition benchmarks show that our method outperforms state-of-the-art methods without any cost of extra annotations.

Moreover, as seen in Fig 1, the deeper part of the backbone (stage 4) acts as a face recognition specified head that focuses on learning the face embedding for recognizing identities. In contrast, the shallow stages of the backbone jointly supervised by both the FR task and the self-supervised 3D auxiliary task can provide a more foundational face representation. Here, we transfer these shallow stages to the facial attribute recognition task. Experiments show that it significantly outperforms pervious method especially when only limited labeled data are available.

Briefly, the main contributions of this paper are summarized as follows:

- We propose a novel 3D bypass enhanced representation learning (3D-BERL) method which improves face recognition by incorporating a self-supervised 3D reconstruction bypass into traditional 2D face recognition pathway.

- The self-supervised reconstruction in the proposed auxiliary bypass acts as a regularization that enables the recognition network to understand faces in 3D view and generate more robust face embedding under unconstrained scenarios.

- The proposed method outperforms state-of-the-art methods on various face recognition benchmarks, and the shallow stages of the backbone trained by our method can also serve as a good foundation model for downstream tasks such as facial attribute recognition with only limited labeled data.

2. Related Works

Learning from Labeled Data. The recent advance of face recognition comes from large-scale labeled training data and the rapid evolution of loss functions. Thanks to the efforts by previous researchers, various annotated face datasets are available. For instance, MS1MV2 [5] contains 5.8 million labeled faces and Glint360K [1] contains 17 million labeled faces.

Given these labeled datasets, existing face recognition methods mostly utilize a softmax-based loss function to train a deep neural network. To improve the performance of softmax loss, SphereFace [17], AM-Softmax [30], CosFace [31] and ArcFace [5] incorporate additional margins into conventional softmax loss function, leading to improved recognition accuracy. However, these methods do not consider the diversity in the difficulty of each sample and only use fixed margins, which may lead to convergence issues. More recently, AdaptiveFace [16], AdaCos [39], MagFace [20] and CurricularFace [12] propose to dynamically tune the margin during the training process. MagFace [20] proposes a quality assessment method to adaptively decide the angular margins for low-quality samples, leading to an improved within-class feature distribution. CurricularFace [12] takes advantage of curriculum learning and adaptively adjusts the importance of easy and hard samples during different training stages. These adaptive margin methods relieve the convergence issues of together training hard and easy samples, and achieve better performance than ArcFace [5]. However, the performance still degenerates when testing under large poses, various illuminations and occlusions as they do not explicitly consider kicking out identity-irrelevant features during training process. Another trend resorts to 3D face reconstruction for tackling large pose face recognition [6, 23, 40], which has shown promising results under unconstrained scenarios.

Learning from Unlabeled Data. The self-supervised learning has drawn increasing attention in recent years as it can learn feature representation without requiring manual annotations. A popular self-supervised learning pipeline is to design annotation-free pretext tasks for neural networks to solve. The pretext tasks commonly apply a transformation to the image and the network is trained via predicting the properties of the transformation. For instance, the jigsaw puzzle [22] shuffles the patches of an image and then trains the neural network to identify the correct location of each patch. Some other famous pretext tasks are image colorizing [38] and rotation prediction [7], which learn the feature representation via predicting the color of each image pixel or predicting the rotated angle of the input image.

To the best of our knowledge, most of the existing pretext tasks are based on pixel-level transformation designed for general images. As shown in [27], although some of them even outperforms supervised learning methods on image classification task, they does not work well on face-related task. It is necessary to design new pretext task for face images. In [27], the face parsing and the facial component prediction are employed as pretext tasks. These two face perception tasks can learn semantic-aware fine-grained feature representations and the pretrained model achieves an improved performance on downstream facial attribute recognition task.

Considering that the albedo and depth are another two identity-relevant characters which inherited in face images, it is reasonable to design a pretext task based on the albedo and depth. Besides, when the face image is occluded, the depth can provide more robust 3D cues than the 2D face texture. Inspired by the self-supervised 3D reconstruction [35], we propose to incorporate a self-supervised bypass of 3D reconstruction into traditional 2D face recognition pathway and the face embedding are forced to focus on identity-relevant depth and albedo information, which is beneficial to face recognition.

3. Method

In this section, we first briefly introduce the preliminary knowledge on image formation model. Then, we present an overview of the proposed method and the details of each key components.

3.1. Preliminary on Imaging Model

Given a single view face image, we predict the face's depth and albedo via neural network and reconstruct the face with imaging model. In this work, we employ the imaging model in [35] to design our 3D reconstruction bypass. It assumes that an image $\mathbf{I} \in \mathbb{R}^{3\times W\times H}$ with width W and height H can be reconstructed with four factors, including the canonical *depth map* $\mathbf{d} \in \mathbb{R}^{1\times W\times H}$, the canonical *albedo map* $\mathbf{a} \in \mathbb{R}^{3\times W\times H}$, the *viewpoint matrix* $\omega \in \mathbb{R}^{2\times 3}$ and the *illumination parameter vector* $\mathbf{l} \in \mathbb{R}^4$.

The *depth map* $\mathbf{d}$ contains the depth value d_{uv} of each pixel (u, v). It is assumed that the distance between the face and the camera is 1 meter and the camera's field of view (FOV) θ_{FOV} is 10°. The *albedo map* $\mathbf{a}$ contains the 3 channels RGB albedos of each pixel. The *viewpoint matrix* ω consists of the rotation angle vector $\mathbf{r} \in \mathbb{R}^3$ and the translation vector $\mathbf{t} \in \mathbb{R}^3$, in which each element is the value along x, y and z axes respectively. The *illumination parameter vector* $\mathbf{l} = [k_a, k_d, l_{dx}, l_{dy}]$, where k_a and k_d are the magnitude of the ambient and diffuse terms of lighting, and $[l_{dx}, l_{dy}]$ presents the lighting direction $\mathbf{l}_d$.

To reconstruct an canonical image $\mathbf{I}_c$, the *normal map* $\mathbf{n}$ is firstly derived from the canonical *depth map* $\mathbf{d}$, in which each element n_{uv} is a vector normal to the underlying 3D surface of pixel (u, v). Then, the coefficients of directional illumination, namely the *shading map* $\mathbf{s}$, is calculated as $s_{uv} = max\{0, \langle \mathbf{l}_d, n_{uv} \rangle\}$. Finally, the canonical face $\mathbf{I}_c$ is generated via the illumination model as follows:

$$\mathbf{I}_c = \mathbf{a} \circ (k_a + k_d \mathbf{s}), \tag{1}$$

where $\circ$ denotes the element-wise multiplication. Note that the canonical face $\mathbf{I}_c$ is obtained with *viewpoint matrix* $\omega = 0$, we need to further warp $\mathbf{I}_c$ to obtain the actual image $\widehat{\mathbf{I}}$ with original view point $\omega = [\mathbf{r}; \mathbf{t}]$. The warping function $\Pi(\mathbf{I}_c, \mathbf{d}, \omega, K)$ which maps the pixel (u, v) in $\mathbf{I}_c$ to the pixel (u', v') in $\widehat{\mathbf{I}}$ is given in Eq. 2:

$$p' \propto K(d_{uv} \cdot \mathbf{r} K^{-1} p + \mathbf{t}) \ , \tag{2}$$

where $p' = (u', v', 1)$, $p = (u, v, 1)$ and K is the camera intrinsic matrix:

$$K = \begin{bmatrix} \frac{W-1}{2tan\frac{\theta_{FOV}}{2}} & 0 & \frac{W-1}{2} \\ 0 & \frac{W-1}{2tan\frac{\theta_{FOV}}{2}} & \frac{H-1}{2} \\ 0 & 0 & 1 \end{bmatrix}. \tag{3}$$

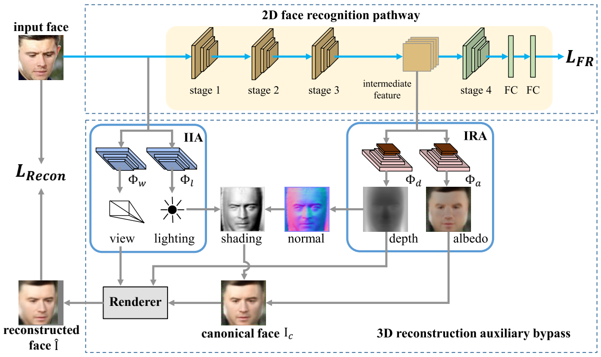

Figure 2. Overview of the 3D reconstruction auxiliary bypass. The identity irrelevant auxiliary (IIA) network estimates the viewpoint and illumination of the input face image. Besides, the identity relevant auxiliary (IRA) network extracts the canonical albedo map and the canonical depth map from the intermediate feature of FR backbone. The whole network is trained in end-to-end manner with both classic face identification loss and the loss of 3D face reconstruction with the physical parameters. In this way, the learnt features are forced to encode more information of canonical facial depth and albedo, which is more intrinsic and beneficial to face recognition.

3.2. Overview

Based on the aforementioned image formation model, to reconstruct a face image, we design several sub-networks to explicitly extract the canonical depth map, canonical albedo map, view matrix and illumination parameter. As shown in Fig 2, we put the these sub-networks into two groups. The first group, namely the identity irrelevant auxiliary (IIA) network, extracts identity irrelevant information, including the viewpoint matrix and the illumination parameter by taking a face image as input. The second group, namely the identity relevant auxiliary (IRA) network, extracts identity relevant information, including the canonical albedo map and the canonical depth map, which is beneficial to the succeeding face recognition task learning.

We hope that the FR backbone mainly encodes the information of the canonical depth and albedo, and neglects the pose and illumination variances. With this in mind, we treat the FR backbone as a unified encoder for these identity relevant information, and embed the identity relevant auxiliary (IRA) network to decoder such identity relevant information from the intermediate feature of FR backbone. As shown in Fig 2, the output feature of the 3-th stage in the ResNet backbone is fed into the identity relevant auxiliary (IRA) network. Meanwhile, the identity irrelevant auxiliary (IIA) network estimates the viewpoint and illumination from the input face image.

Both the IRA network and IIA network are trained by self-supervised 3D reconstruction, and the reconstruction process is based on the aforementioned model of image formation and a differentiable renderer [13]. Among these two auxiliary networks, the learning of IRA network will enforce the shallow part of FR backbone (the stage 1, 2, 3) to focus on the identity-relevant depth and albedo. Moreover, as the depth and albedo learnt by IRA network are with a canonical view, the pose and illumination are mostly kicked out from the intermediate feature of FR backbone. Therefore, with the supervision from the margin-based loss function, the succeeding layers (stage 4) can get ride of the pose and illumination, leading to a more robust face embedding. It is worth mentioning that both the IRA and IIA networks will be detached at inference time, thus the face embedding is improved without any cost of inference speed.

3.3. Identity Irrelevant Auxiliary (IIA) Network

The identity irrelevant auxiliary (IIA) network aims to estimate the identity irrelevant information. It consists of two sub networks Φ_w and Φ_l, which take the image $\mathbf{I}$ as input and estimate the *viewpoint matrix* $\omega = \Phi_w(\mathbf{I})$ and the *illumination parameter vector* $\mathbf{l} = \Phi_l(\mathbf{I})$, respectively.

These two sub networks have identical backbone structure with 10 layers (Conv-ReLU × 5). Besides, each of them has a output head (Conv-Tanh) to finally generate the *viewpoint matrix* $\omega \in \mathbb{R}^{2\times 3}$ or the *illumination parameter vector* $\mathbf{l} \in \mathbb{R}^4$.

3.4. Identity Relevant Auxiliary (IRA) Network

The identity relevant auxiliary (IRA) network aims to extract the canonical albedo map and the canonical depth map which are relevant to face identity. It consists of two sub network Φ_d and Φ_a, which are both with the encoder-decoder structure and extract the canonical *depth map* $\mathbf{d}$ and the canonical *albedo map* $\mathbf{a}$. As we treat the shallow stages of the backbone (stage 1, 2, 3) as a unified encoder for the identity-relevant encoder, the Φ_d and Φ_a takes the 14×14 intermediate features $\mathbf{f}_{s3}$ from the stage 3 as their inputs.

To further enlarge the reception field, we still design a tiny encoder structure with 7 layers (Conv-GroupNorm-LeakyReLU, Conv-LeakyReLU, Conv-ReLU) in the Φ_d and Φ_a. For the decoder in Φ_d and Φ_a, we employ the decoder structure designed in [35].

After obtaining all the outputs of the IIA network and the IRA network, image $\widehat{\mathbf{I}}$ is reconstructed by Eq. 4:

$$\widehat{\mathbf{I}} = \Pi(\Phi_a(\mathbf{f}_{s3}) \circ (k_a + k_d\mathbf{s}), \Phi_d(\mathbf{f}_{s3}), \Phi_w(\mathbf{I}), K). \quad (4)$$

The reconstruction process of Eq. 4 is briefly shown in the Fig 2. At first, the canonical *depth map* $\mathbf{d}$ learnt by the IRA network will be used to calculate the *normal map* $\mathbf{n}$. Then, with the lighting parameters estimated by the IIA network, the *shading map* $\mathbf{s}$ is derived. Afterwards, we obtain the reconstructed canonical face $\mathbf{I}_c = \Phi_a(\mathbf{f}_{s3}) \circ (k_a + k_d\mathbf{s})$ by the lighting model in Eq 1. Finally, using the warping function Π, the renderer module warps the canonical face $\mathbf{I}_c$ to the view of $\Phi_w(\mathbf{I})$, and produces the final reconstructed face $\widehat{\mathbf{I}}$.

3.5. Joint Loss Function

In our 3D-BERL framework, the FR backbone is jointly trained with the conventional face recognition loss and the reconstruction loss of our 3D auxiliary bypass. Here, we employ CurricularFace [12] as the FR loss L_{FR}. Given a training batch of N face samples, the L_{FR} is formulated as:

$$L_{FR} = -\frac{1}{N}\sum_{i=1}^{N}\frac{e^{s\,cos(\theta_{y_i}+m)}}{e^{s\,cos(\theta_{y_i}+m)}+\sum_{j\neq y_i}e^{s\,G(t,cos(\theta_j))}}, \tag{5}$$

where if $cos(\theta_{y_i}+m)-cos(\theta_j)\geq 0$, $G(t,cos(\theta_j)) = cos(\theta_j)$, otherwise $G(t,cos(\theta_j)) = cos(\theta_j)(t+cos(\theta_j))$. The m and s are the margin and scale parameters. The t is an adaptively updated hyper-parameter that modulates the negative cosine similarities.

The reconstruction loss L_{Recon} is defined in Eq 6. It minimize the L_1 distance between $\widehat{\mathbf{I}}_i$ and $\mathbf{I}_i$.

$$L_{Recon} = \frac{1}{N}\sum_{i=1}^{N}|\widehat{\mathbf{I}}_i - \mathbf{I}_i|. \tag{6}$$

The joint loss function is formulated in Eq. 7, where the γ_1 and γ_2 are used to balance the two losses. The gradients of L_{FR} loss will back propagate through the whole backbone and the gradients of L_{Recon} will update both the two auxiliary networks and the shallow part of the FR backbone (stage 1, 2, 3).

$$L_{total} = \gamma_1 L_{FR} + \gamma_2 L_{Recon}. \tag{7}$$

As seen in Fig 2, the deeper part of the backbone (stage 4) acts as a face recognition specified head that focuses on learning the face embedding for recognizing identities. By encoding more identity-relevant canonical depth and albedo into the first three stages, experiments in section 4 shows that this face recognition head has extracted more discriminative embeddings for face identification. Furthermore, the shallow stages of the backbone jointly trained by both the FR loss and the self-supervised 3D reconstruction loss can provide more foundational face representations for various face analysis tasks. The experiments in section 4 demonstrate that those face representations can achieve superior performance after transferring to the downstream facial attribute recognition task.

4. Experiments

4.1. Dataset

We employ MS1MV2 [5] as our training set for fair comparisons with other methods. As a refined version of the MS-Celeb-1M [8] dataset, MS1MV2 contains 5.8 million face images from 85K identities. To evaluate the effectiveness of the proposed method, we extensively test our 3D-BERL on three popular benchmarks, including IJB-B [34], IJB-C [19] and MegaFace [14]. Moreover, we transfer the learnt face representations to other face analysis tasks like facial attribute recognition and conducts experiments on CelebA [18] and LFWA [18] to further verify the effectiveness of our method. Here, we briefly introduce the aforementioned public datasets.

IJB-B and IJB-C. The IJB-B [34] dataset consists of 55K video frames and 21.8K images. As a further extension of IJB-B, the IJB-C [19] is a larger dataset with 117.5K video frames and 31.3K images. IJB-C also contains more natural occlusions and increased diversity of geographic origin. In this work, we conduct comparisons using the standard protocol of IJB-B and IJB-C. Besides, to further verify the performance of recognizing occluded faces images, we also employed an occlusion specified protocol of IJB-C, in which each testing sample contains at least one occluded facial region.

MegaFace. The probe set of the MegaFace [14] dataset contains 106.8K face images from 530 subjects. Since it has a huge gallery set with more than 1 million face images from 690K individuals, the MegaFace dataset can evaluate whether the face recognition model can handle million-scale distractors. The standard MegaFace challenge 1 (MF1) protocol is employed in our experiments.

CelebA. The CelebA [18] is a large-scale facial attribute dataset, which contains more than 202K face images with annotations of 40 attributes.

LFWA. The LFWA [18] is another widely-used facial attribute dataset. It contains 13.2K face images and the 40 facial attributes are annotated in an identical way with the CelebA dataset.

4.2. Implementation Details

In the experiments, the face images are aligned to the size of 112×112 with five detected facial landmarks, *i.e.*, eyes, mouth corners and nose. We adopt the ResNet-100 in [5] as our backbone which is trained with the CurricularFace loss [12]. The backbone is trained by SGD optimizer with momentum of 0.9. The two auxiliary networks is trained by Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The loss weights of the supervised face recognition loss and the self-supervised 3D reconstruction loss are identically set to 1. The batch size is set to 200 and the weight decay is set to 5e-4. Our 3D-BERL framework is implemented in Pytorch [24] and all the models are trained on four NVIDIA TITAN XP GPUs. To speed up the training, the size of the reconstruction image $\widehat{\mathbf{I}}_i$ and other auxiliary variables ($\mathbf{d}$, $\mathbf{a}$, $\mathbf{n}$, $\mathbf{s}$) is set to 64×64. The $\mathbf{I}_i$ in Eq 6 is also the 64×64 resized version of the input image. The overall training procedure contains three stages. Firstly, we train the FR backbone only using the FR loss function. Then, we train the proposed IIA and IRA networks with the FR backbone frozen. Finally, in the third stage, both the FR backbone

Table 1. Ablation studies: after transferring from each pretext tasks, the recognition accuracy (%) with different proportions of labeled LFWA training samples. **Base**: the baseline. **3D**: the self-supervised 3D reconstruction. **FR**: the supervised face recognition.

			Proportion of training samples				
Base	3D	FR	5%	10%	20%	50%	100%
✓			77.82	80.24	78.82	81.47	86.43
✓	✓		79.96	83.15	84.16	85.45	86.85
✓		✓	83.39	84.62	85.11	85.74	86.34
✓	✓	✓	84.08	85.33	85.96	86.13	86.39

and the two auxiliary networks are jointly optimized with the loss defined in Eq 7.

For the downstream facial attribute recognition task, we discard the layers in the stage 4 of ResNet-100 and add two FC layers as an attribute classification head on the top of the stage 3 which is previously regularized by the self-supervised 3D reconstruction. To evaluate the performance of the facial attribute recognition with limited training data, we choose a small proportion of the training set for training. The batch size is set to 40 and the whole network is finetuned for 60 epochs.

4.3. Ablation Study

In our framework, both the supervised face recognition task and the self-supervised 3D reconstruction task can be treated as pretext tasks for the downstream facial attribute recognition. To investigate the effectiveness of these two pretext tasks, we perform an ablation study on the LFWA dataset by evaluating three variants of the proposed method. The first one is the fully self-supervised setting that only the 3D reconstruction task is used for feature learning. Secondly, only the face recognition task is employed. Finally, these two tasks are combined together to learn face representations. These three variants are all pre-trained on MS1MV2 and then finetuned with different proportions of labeled LFWA training data. Moreover, training from scratch on the same LFWA data is conducted as the baseline.

The results are shown in Table 1. As seen, when only a small proportion of labeled training data is available, initialization with the self-supervised 3D reconstruction task and the supervised face recognition task both outperforms the baseline. Specifically, when using 10% of the training data, the 3D task and the FR task achieve improvements up to 2.91% and 4.38%, respectively. These results illustrate that both the two tasks are beneficial to the downstream facial attribute recognition task. When the two kinds of tasks are jointly used for pre-training, the attribute recognition accuracy can be further improved, which demonstrates that the self-supervised 3D reconstruction task and the supervised face recognition task benefit from each other to achieve better face representations for downstream tasks.

Table 2. Performance on IJB-B dataset.

	IJB-B (TAR@FAR)		
Methods	1e-6	1e-5	1e-4
SphereFace [17]	39.40	73.58	89.19
CosFace [31]	40.41	89.25	94.01
ArcFace [5]	-	-	94.20
NPCFace [37]	-	85.59	92.02
MagFace [20]	40.91	89.88	94.33
CurricularFace [12]	42.26	89.02	94.83
3D-BERL	45.77	90.60	94.98

4.4. Results on IJB-B and IJB-C

For fair comparisons with state-of-the-art methods, we follow the same testing protocol in [5, 12] to conduct evaluations on two template-based face recognition benchmarks, *i.e.*, IJB-B and IJB-C. The average feature of all images in one template is utilized as the final feature embedding of the template. The true accept rate (TAR) and the false accept rate (FAR) are employed as the evaluation metrics for the 1:1 verification task. Here, we mainly report the TARs when FAR is 1e-6, 1e-5 and 1e-4, respectively.

We Firstly conduct a comparison with state-of-the-art methods on IJB-B dataset. As shown in Table 2, 3D-BERL outperforms the strong baseline CurricularFace [12] with a large improvement up to 3.51% in terms of TAR when FAR=1e-6. Attributed to the enhanced face representations by the 3D reconstruction, 3D-BERL also shows superior performance than other FR methods including SphereFace [17], CosFace [31], ArcFace [5] and NPCFace [37]. Moreover, the proposed method also surpasses the more recent method MagFace [20] which proposes a quality assessment method to improve the margin-based loss function. An improvement up to 4.86% is witnessed, which demonstrates the effectiveness of integrating self-supervised 3D face reconstruction into face recognition learning.

Since only the TAR at FAR=1e-4 on IJB-C dataset is reported in [12], to get a detailed comparison with CurricularFace, we re-implement it and get a result which is on par with the original paper. Table 3 summarizes the result on IJB-C dataset. Similar conclusions can be obtained that our 3D-BERL outperforms state-of-the-art methods. Compared to this strong baseline [12], the proposed method achieves improvements of 0.99% and 0.45% on the TAR at FAR=1e-6 and 1e-5, respectively. It is worth mentioning that 3D-BERL uses exactly the same FR backbone as CurricularFace, thus their comparison also takes the role of the ablation study on FR task. The abovementioned results have demonstrated that it is the proposed 3D reconstruction bypass that helps to improve the discriminant ability of the backbone.

We then evaluate 3D-BERL on the more challenging

Table 3. Performance on IJB-C dataset.

	IJB-C (TAR@FAR)		
Methods	1e-6	1e-5	1e-4
CenterFace [33]	-	78.10	85.30
SphereFace [17]	68.86	83.33	91.77
AdaCos [39]	83.28	88.03	92.40
DDL [11]	-	88.40	93.10
PFE [26]	-	89.64	93.25
CosFace [31]	87.96	92.68	95.56
ArcFace [5]	-	-	95.60
DUL [4]	-	90.23	94.21
NPCFace [37]	-	88.08	92.90
DiscFace [15]	-	92.42	94.82
CircleLoss [29]	-	89.10	93.25
MagFace [20]	89.26	93.67	95.81
CurricularFace [12]	87.46	93.85	96.20
3D-BERL	88.45	94.30	96.20

Table 4. Performance on IJB-C occlusion subset.

	IJB-C occlusion (TAR@FAR)			
Methods	1e-6	1e-5	1e-4	1e-3
InterpretFR [36]	-	-	-	89.80
CurricularFace [12]	70.23	89.23	93.41	95.33
3D-BERL	75.09	90.24	93.41	95.44

IJB-C occlusion subset, in which each face contains at least one occluded facial region. As illustrated in Table 4, 3D-BERL significantly outperforms the CurricularFace [12]. The TAR at FAR=1e-6 is improved by 4.86%, demonstrating that the feature learnt by 3D-BERL is more robust to occlusion. The reason is that the depth branch regularizes face recognition features to encode more information of 3D shape which is more robust to occlusion. To verify this, we visualize the depth extracted by 3D-BERL in Fig 3. The results of two occluded faces of IJB-C dataset are shown in the last two rows of Fig 3. As seen, although the occlusion has damaged the facial texture, 3D-BERL still obtained reasonable depth map. In this way, 3D-BERL improves the face recognition performance under occluded scenarios.

4.5. Results on MegaFace

Table 5 summarizes the rank-1 accuracies of state-of-the-art methods on large-scale MegaFace [14] benchmark. Compared with the recent strong competitors, 3D-BERL achieves a state-of-the-art result of 98.63%, demonstrating the superiority of the proposed method again.

Table 5. Rank-1 identification accuracy (%) on MegaFace Challenge 1.

Methods	MF1 Rank1
AdaptiveFace [16]	95.02
AdaCos [39]	97.41
CosFace [31]	97.91
MV-Arc-Softmax [32]	97.74
ArcFace [5]	98.35
CircleLoss [29]	98.50
3D-BERL	98.63

4.6. Results on CelebA and LFWA

Table 6. Recognition accuracy (%) on CelebA dataset.

	Proportion of training samples				
Methods	0.2%	0.5%	1%	2%	100%
SlimCNN [25]	79.90	80.20	80.96	82.32	91.24
DeepCluster [3]	83.21	86.13	87.46	88.86	91.68
JigsawPuzzle [22]	82.88	84.71	86.25	87.77	91.57
Rot [7]	83.25	86.51	87.67	88.82	91.69
FixMatch [28]	80.22	84.19	85.77	86.14	89.78
VAT [21]	81.44	84.02	86.30	87.28	91.44
Unsup3D [35]	83.36	84.61	86.04	86.88	90.31
SSPL [27]	86.67	88.05	88.84	89.58	91.77
Baseline	80.53	82.25	83.06	83.89	89.48
3D-BERL	87.34	88.02	88.95	89.42	90.23

We believe the shallow stages jointly trained by both the FR loss and the self-supervised 3D reconstruction loss can provide robust face representations to various face analysis tasks. In this section, we conduct experiments on a challenging task of facial attribute prediction with limited labeled data. We utilize CelebA and LFWA datasets for experiments and take five different proportions of training data to finetune the downstream facial attribute recognition task. As seen in Table 6, 3D-BERL achieves comparable or even superior performance than state-of-the-art methods on CelebA dataset. Comparing to the self-supervised methods [3, 7, 22] designed for general objects, the proposed method achieves significant improvements, especially when less training data (such as 0.2%) is used. This is because the proposed auxiliary bypass exploits the 3D information of face structure, which is more favorable for backbone to capture facial attributes. Moreover, the proposed method also achieves higher accuracy than the semi-supervised methods FixMatch [28] and VAT [21].

Recently, attributed to the pretext task of face paring, SSPL [27] outperforms all previous methods for facial attribute prediction with limited training data. Differently, 3D-BERL focuses on learning 3D face structure and achieves an improvement of 0.67% when 0.2% of training data is used. It is worth noting that SSPL uses larger image size of 224×224, while 3D-BERL only requires 112×112 images. Even with a lower resolution, our method still achieves comparable results with SSPL.

Here, we also compare 3D-BERL with Unsup3D [35], which is most related to our work. We concatenate the out-

Table 7. Recognition accuracy (%) on LFWA dataset.

	Proportion of training samples				
Methods	5%	10%	20%	50%	100%
SlimCNN [25]	70.90	71.49	72.12	73.45	76.02
DeepCluster [3]	74.21	77.42	80.77	84.27	85.90
JigsawPuzzle [22]	73.90	77.01	79.56	83.29	84.86
Rot [7]	74.40	76.67	81.52	84.90	85.72
FixMatch [28]	71.42	72.78	75.10	80.87	83.84
VAT [21]	72.19	74.42	76.26	80.55	84.68
SSPL [27]	78.68	81.65	83.45	85.43	86.53
Baseline	77.82	80.24	78.82	81.47	86.43
3D-BERL	84.08	85.33	85.96	86.13	86.39

put of the albedo network and the depth network in Unsup3D and finetune it on the downstream facial attribute recognition task. As shown in Table 6, benefited from the initialization with the self-supervised 3D reconstruction, Unsup3D achieves better results than the baseline. Moreover, our 3D-BERL significantly outperforms Unsup3D with an improvement up to 3.98% when 0.2% of training data is used. The possible reason is that our self-supervised 3D reconstruction bypass enforces the features learnt by face recognition task to focus on the pose and illumination irrelevant features while face recognition pathway assists the features learnt by 3D reconstruction bypass to be more specific to face perceptron tasks. In other words, they benefit from each other.

Tabel 7 summarizes the results on LFWA dataset. As seen, 3D-BERL significantly outperforms previous methods by a large margin when only 5%, 10% or 20% of training data is used. When only 5% training data is used, 3D-BERL outperforms SSPL with an improvement up to 5.4%, which further demonstrates the effectiveness of 3D-BERL for learning robust face representations.

4.7. Visualization Results

In Fig 3, we shows the visualization results of our 3D reconstruction auxiliary bypass. As seen, both the albedo and depth are recovered in the natural canonical view and with high fidelity, meaning that the FR backbone has encoded sufficient information of canonical facial depth and albedo, which is more intrinsic and beneficial to face recognition. As shown in the last two rows of Fig 3, even in the presence of uncontrolled pose and occlusion, the proposed auxiliary bypass still decodes high fidelity canonical depth from the intermediate features of FR backbone, and this has explained why our method significantly improves the occluded face recognition in Table 4.

4.8. Limitations

While 3D-BERL has achieved promising results even with partial occlusion and medium pose, the 3D reconstruction fails in case of extreme poses, partially due to the severe self-occlusion not considered in the imaging model. This may be improved by imposing more complex imaging model and complicated reconstruction networks.

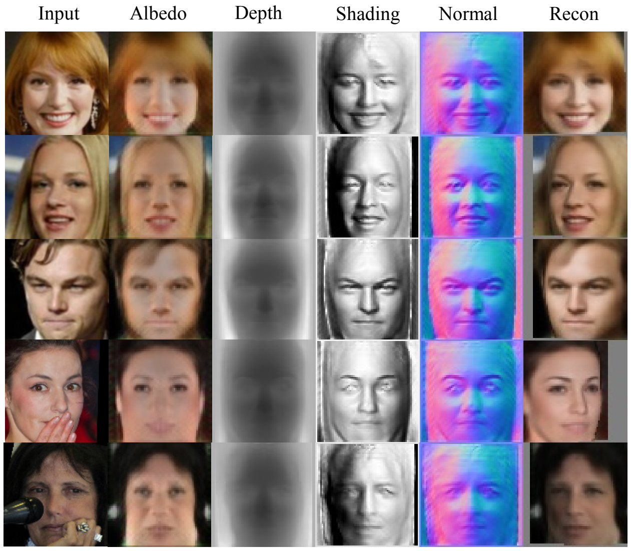

Figure 3. The visualization results of the 3D reconstruction auxiliary bypass.

5. Conclusion

To learn more robust face embedding under unconstrained scenarios, we propose to incorporate a self-supervised 3D reconstruction bypass into traditional 2D face recognition pathway. Inspired by the physical model of image formation, the 3D reconstruction bypass consists of two auxiliary networks: one for pose and lighting and the other for canonical depth and albedo. Then, we reconstruct the face image via image formation model and the reconstruction loss is exploited to enforce the FR backbone to encode more information of canonical facial depth and albedo. In other words, it enables the FR backbone to understand faces in 3D view which is intrinsic and beneficial to face recognition. Extensive experimental results show that the proposed method not only significantly improves the face recognition accuracy, but also provides a good foundation model for downstream tasks such as facial attribute recognition task with only limited labeled data.

Acknowledgments

This work is partially supported by National Key R&D Program of China (No. 2017YFA0700800), National Natural Science Foundation of China (Nos. 62176251 and Nos. 61976219).

References

[1] Xiang An, Xuhan Zhu, Yang Xiao, Lan Wu, Ming Zhang, Yuan Gao, Bin Qin, Debing Zhang, and Ying Fu. Partial fc: Training 10 million identities on a single machine. *arXiv preprint arXiv:2010.05222*, 2020. 2

[2] Jiajiong Cao, Yingming Li, and Zhongfei Zhang. Celeb-500k: A large training dataset for face recognition. In *IEEE International Conference on Image Processing (ICIP)*, pages 2406–2410, 2018. 1

[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision (ECCV)*, pages 139–156, 2018. 7, 8

[4] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5710–5719, 2020. 7

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 1, 2, 3, 5, 6, 7

[6] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos P Zafeiriou. Fast-ganfit: Generative adversarial network for high fidelity 3d face reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2021. 3

[7] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018. 2, 3, 7, 8

[8] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large scale face recognition. In *European Conference on Computer Vision (ECCV)*, pages 87–102, 2016. 1, 5

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2

[10] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 1

[11] Yuge Huang, Pengcheng Shen, Ying Tai, Shaoxin Li, Xiaoming Liu, Jilin Li, Feiyue Huang, and Rongrong Ji. Improving face recognition from hard samples via distribution distillation loss. In *European conference on computer vision (ECCV)*, pages 138–154, 2020. 7

[12] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5901–5910, 2020. 1, 3, 5, 6, 7

[13] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3907–3916, 2018. 4

[14] Ira Kemelmacher-Shlizerman, Steven M. Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4873–4882, 2016. 5, 7

[15] Insoo Kim, Seungju Han, Seong-Jin Park, Ji won Baek, Jinwoo Shin, Jae-Joon Han, and Changkyu Choi. Discface: Minimum discrepancy learning for deep face recognition. In *Asian Conference on Computer Vision (ACCV)*, pages 358–374, 2020. 7

[16] Hao Liu, Xiangyu Zhu, Zhen Lei, and Stan Z. Li. Adaptiveface: Adaptive margin and sampling for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11947–11956, 2019. 1, 3, 7

[17] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 212–220, 2017. 1, 2, 6, 7

[18] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision (ICCV)*, December 2015. 5

[19] Brianna Maze, Jocelyn Adams, James A. Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K. Jain, W. Tyler Niggel, Janet Anderson, Jordan Cheney, and Patrick Grother. Iarpa janus benchmark-c: Face dataset and protocol. In *International Conference on Biometrics (ICB)*, pages 158–165, 2018. 5

[20] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14225–14234, 2021. 1, 3, 6, 7

[21] Takeru Miyato, Shin-Ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 41(8):1979–1993, 2019. 7, 8

[22] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision (ECCV)*, pages 69–84, 2016. 2, 3, 7, 8

[23] Konstantinos Papadopoulos, Anis Kacem, Abdelrahman Shabayek, and Djamila Aouada. Face-gcn: A graph convolutional network for 3d dynamic face identification/recognition. *arXiv preprint arXiv:2104.09145*, 2021. 3

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. volume 32, pages 8026–8037, 2019. 5

[25] Ankit Kumar Sharma and Hassan Foroosh. Slim-cnn: A light-weight cnn for face attribute prediction. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 329–335, 2020. 7, 8

[26] Yichun Shi and Anil Jain. Probabilistic face embeddings. In *IEEE International Conference on Computer Vision (ICCV)*, pages 6902–6911, 2019. 7

[27] Ying Shu, Yan Yan, Si Chen, Jing-Hao Xue, Chunhua Shen, and Hanzi Wang. Learning spatial-semantic relationship for facial attribute recognition with limited labeled data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11916–11925, 2021. 2, 3, 7, 8

[28] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 7, 8

[29] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6398–6407, 2020. 7

[30] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters (SPL)*, 25(7):926–930, 2018. 1, 2

[31] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5265–5274, 2018. 1, 2, 6, 7

[32] Xiaobo Wang, Shifeng Zhang, Shuo Wang, Tianyu Fu, Hailin Shi, and Tao Mei. Mis-classified vector guided softmax loss for face recognition. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 12241–12248, 2020. 7

[33] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision (ECCV)*, pages 499–515, 2016. 7

[34] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K. Jain, James A. Duncan, Kristen Allen, Jordan Cheney, and Patrick Grother. Iarpa janus benchmark-b face dataset. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 592–600, 2017. 5

[35] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–10, 2021. 3, 4, 7

[36] Bangjie Yin, Luan Tran, Haoxiang Li, Xiaohui Shen, and Xiaoming Liu. Towards interpretable face recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 9348–9357, 2019. 7

[37] Dan Zeng, Hailin Shi, Hang Du, Jun Wang, Zhen Lei, and Tao Mei. Npcface: Negative-positive collaborative training for large-scale face recognition. *arXiv preprint arXiv:2007.10172*, 2020. 6, 7

[38] Richard Yi Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *European Conference on Computer Vision (ECCV)*, pages 649–666, 2016. 3

[39] Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10823–10832, 2019. 1, 3, 7

[40] Jian Zhao, Lin Xiong, Jianshu Li, Junliang Xing, Shuicheng Yan, and Jiashi Feng. 3d-aided dual-agent gans for unconstrained face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 41(10):2380–2394, 2019. 3