SynthDistill: Face Recognition with Knowledge Distillation from Synthetic Data

Hatef Otroshi Shahreza[1,2], Anjith George[1], Sébastien Marcel[1,3]
[1]Idiap Research Institute, Martigny, Switzerland
[2]École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[3]Université de Lausanne (UNIL), Lausanne, Switzerland
{hatef.otroshi,anjith.george,sebastien.marcel}@idiap.ch

Abstract

State-of-the-art face recognition networks are often computationally expensive and cannot be used for mobile applications. Training lightweight face recognition models also requires large identity-labeled datasets. Meanwhile, there are privacy and ethical concerns with collecting and using large face recognition datasets. While generating synthetic datasets for training face recognition models is an alternative option, it is challenging to generate synthetic data with sufficient intra-class variations. In addition, there is still a considerable gap between the performance of models trained on real and synthetic data. In this paper, we propose a new framework (named SynthDistill) to train lightweight face recognition models by distilling the knowledge of a pretrained teacher face recognition model using synthetic data. We use a pretrained face generator network to generate synthetic face images and use the synthesized images to learn a lightweight student network. We use synthetic face images without identity labels, mitigating the problems in the intra-class variation generation of synthetic datasets. Instead, we propose a novel dynamic sampling strategy from the intermediate latent space of the face generator network to include new variations of the challenging images while further exploring new face images in the training batch. The results on five different face recognition datasets demonstrate the superiority of our lightweight model compared to models trained on previous synthetic datasets, achieving a verification accuracy of 99.52% on the LFW dataset with a lightweight network. The results also show that our proposed framework significantly reduces the gap between training with real and synthetic data. The source code for replicating the experiments is publicly released.

1. Introduction

Recent advancements in face recognition systems have been driven by deep neural networks trained on large-scale datasets, leading to remarkable progress in accuracy [16, 27]. However, the state-of-the-art face recognition networks are often computationally heavy and the deployment of these networks on edge devices poses practical challenges. Nevertheless, it is possible to develop efficient networks from these large models that achieve comparable accuracy with significantly reduced computational load, making them suitable for edge device deployment.

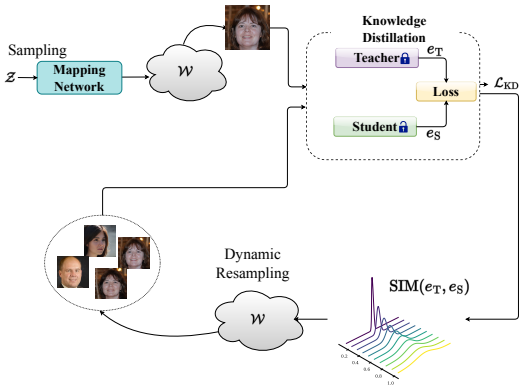

Figure 1. Schematic showing the proposed approach (SynthDistill). Latent space of StyleGAN is first sampled from $\mathcal{Z}$ space, and then dynamically re-sampled from $\mathcal{W}$ space based on teacher-student agreement. This dynamic re-sampling leads to the generation of challenging samples that facilitate efficient learning.

One strategy is training lightweight and efficient networks on the large-scale face recognition datasets [1, 6, 9, 18, 30]. However, training an efficient face recognition model using large-scale face recognition datasets requires access to such a dataset. Nonetheless, large-scale face recognition datasets, such as VGGFace2 [11], MS-Celeb [19], WebFace [56], etc., were collected by crawling images from the Internet, thus raising legal, ethical, and privacy concerns [10]. To address such concerns, recently several works proposed generating synthetic face datasets and use the synthetic face images for training face recogni-

tion models [3, 7, 28]. However, generating synthetic face datasets with sufficient inter-class and intra-class variations is still a challenging problem. Our experimental results also show that there is still a large gap in the recognition performance when training a lightweight face recognition model on real data and existing synthetic face datasets.

Another strategy to train a lightweight face recognition model is to transfer the knowledge of a model trained on a large dataset to a lightweight network through knowledge distillation [21]. Notwithstanding, the knowledge distillation from a teacher model often requires access to the original or another large-scale real dataset. Meanwhile, access to a real dataset for knowledge distillation may not always be feasible due to the size of the datasets. Even if there is access to real large-scale dataset, there remain ethical and legal concerns of using large-scale face recognition datasets crawled from internet. In this work, we propose a new framework to distill the knowledge of a pretrained teacher using synthetic face images without identity labels, and thus mitigating the need for real identity-labeled data during the distillation phase. We propose dynamic sampling from the intermediate latent space of a StyleGAN to generate new images and enhance training.

In contrast to previous approaches that rely on static generation of synthetic face datasets [3, 7, 28] and then using the generated dataset for training the FR model, we combine these two steps with an online-generation of synthetic images and training the lightweight network in the image generation loop within a knowledge distillation based framework. This avoids the requirements of hard identity labels for the generated images, and further assists the generation network to produce challenging samples though a feedback mechanism while exploring more image variations, thus enabling the training of more robust models. In addition, compared to previous works for the training of face recognition models on synthetic datasets, our proposed knowledge distillation framework does not require identity labels in the training, simplifying the process of generating synthetic face images. We should also note that previous synthetic datasets still used a face recognition model in the dataset generation pipeline.

In our case, we also employ a pre-trained face recognition model in our pipeline, but with the role as a teacher. However, instead of generating a static synthetic dataset with identity labels, we dynamically create synthetic face images during the knowledge distillation process. This novel approach allows us to frame our knowledge distillation as a label-free training paradigm, utilizing synthetic data to effectively train lightweight face recognition models.

It is noteworthy that we do not need access to the complete whitebox knowledge of the teacher network in our proposed knowledge distillation approach, and thus our method can also be used in case of a blackbox access to the teacher model that can used to generate the embeddings, given the embeddings are available. We adapt the TinyNet [20] architecture and train lightweight face recognition models (called *TinyFaR*) in our knowledge distillation approach. We provide an extensive experimental evaluation on five different face recognition benchmarking datasets, including LFW [23], CA-LFW [54], CP-LFW [53], CFP-FP [40] and AgeDB-30 [36]. Our experimental results demonstrate the effectiveness of our approach in achieving efficient face recognition systems with reduced computational requirements, while avoiding the use of real data for knowledge distillation. This opens new possibilities for developing privacy-aware and resource-efficient face recognition models suitable for edge devices. Fig. 1 illustrates the general block diagram of our proposed knowledge distillation framework with dynamic sampling.

The main contributions of this work are listed below:

- We propose a novel framework to train a lightweight face recognition model using knowledge distillation. The proposed knowledge distillation framework is based on synthetic face images and does not require real training data. In addition, we do not need identity-labeled training data in our knowledge distillation framework, mitigating problems in generating synthetic face recognition datasets.
- Our proposed knowledge distillation framework is based on a dynamic sampling of difficult samples during training to enhance the training. Dynamic sampling helps the student network to simultaneously learn on new images (i.e., increase generalization), while focusing on difficult samples. Therefore, the training images are synthesized online and during the distillation process.
- We provide extensive experimental results on different face recognition datasets, showing superior recognition accuracy for lightweight face recognition models trained in our framework compared to training lightweight face recognition from scratch using other synthetic datasets.

The remainder of the paper is organized as follows. In Section 2 we review the related works in the literature. We describe our proposed framework for knowledge distillation with synthetic data using dynamic latent sampling in Section 3. We report our experimental results in Section 4 and also discuss our results in Section 5. Finally, the paper is concluded in Section 6.

2. Related works

In this section, we discuss the relevant literature on synthetic datasets, light-weight face recognition networks, and

knowledge distillation in the context of face recognition.

2.1. Synthetic Datasets

Several works have explored the generation of synthetic datasets for training face recognition. It is worth noting that many large-scale datasets are typically collected through web-crawling without explicit informed consent. By leveraging synthetic datasets, it becomes possible to mitigate concerns regarding the privacy of individuals while also potentially addressing issues such as bias [24, 41]. These synthetic datasets are often generated using variations of StyleGAN, 3D models, and diffusion models.

Several prior works, including FaceID-GAN [42], identity-preserving face images [4] [51], have employed synthesis techniques to generate facial images. Notably, FF-GAN [51] (e.g., 3DMM [5]) and DiscoFaceGAN [17] leverages 3D priors. In [37], authors proposed an approach called SynFace which incorporates the use of identity mixup (IM) and domain mixup (DM) techniques to address the performance gap. They use a small portion of labeled real data in the training process to reduce the domain gap between real and synthetic data to improve the performance. Additionally, the controllable face synthesis model provides a convenient means to manipulate various aspects of synthetic face generation, such as pose, expression, illumination, the number of identities, and samples per identity. Boutros et al. [7], presented a method to generate synthetic data using a class conditional generative adversarial network. The authors trained the StyleGAN2-ADA model [25] on the CASIA-WebFace [49] datasets, using identities as class labels. They have conducted experiments using the generated SFace dataset to show its utility in training face recognition models. Bae et al. [3], introduced a large-scale synthetic dataset for face recognition named DigiFace-1M. This dataset was created by utilizing a computer graphics pipeline to render digital faces. Each identity within the dataset is generated by incorporating randomized variations in facial geometry, texture, and hairstyle. The rendered faces exhibit diverse attributes such as different poses, expressions, hair color, hair thickness, and density, as well as accessories. Through the implementation of aggressive data augmentation techniques, they reduced the domain gap between the generated images and real face images leading to gains in face recognition performance. In [28], authors proposed a Dual Condition Face Generator (DCFace) utilizing a diffusion model. This approach incorporates a novel Patch-wise style extractor and Time-step dependent ID loss, enabling DCFace to consistently generate face images depicting the same individual in different styles, while maintaining precise control over the process.

Despite the advantages of synthetic data in terms of privacy and consent, the performance of face recognition models trained on these datasets falls short when compared to models trained on real data. This severely limits real-world usage of models trained on synthetic datasets. To address these challenges, we propose a novel strategy for training face recognition models using synthetic data within an kowledge distillation framework. Our method generates data online dynamically and eliminates the need for real data during the distillation phase.

2.2. Efficient Face Recognition

As edge computing gained prevalence, there is an increased focus on developing lightweight face recognition models without compromising accuracy. In the initial phase of efficient model development, Wu et al. introduced LightCNN, a lightweight architecture [47]. MobileNets [22, 39] employed depth-wise separable convolutions to improve the performance. Building upon the MobileNet architecture, MobileFaceNets were designed for real-time face verification tasks [15]. The concept of MixConv, which incorporates multiple kernel sizes in a single convolution, was used to develop MixFaceNet networks for lightweight face recognition [44, 6]. Inspired by ShuffleNetV2 [34], ShuffleFaceNet models were proposed for face recognition, with parameter counts ranging from 0.5M to 4.5M and verification accuracies exceeding 99.20% on the LFW dataset [35]. Neural architecture search (NAS) was utilized in [9] to automatically design an efficient network called PocketNet for face recognition. The PocketNet architecture was learned using the differential architecture search (DARTS) algorithm on the CASIA-WebFace dataset, and knowledge distillation (KD) was employed during training. Yan et al. [48] employed knowledge distillation (KD) and variable group convolutions to address computational intensity imbalances in face recognition networks. Alansari et al. proposed GhostFaceNets, which exploit redundancy in convolutional layers to create compact networks [1]. These modules generate a fixed percentage of convolutional feature maps using computationally inexpensive depth-wise convolutions. Recently, George et al. introduced EdgeFace, a combination of CNN-Transformer architecture that achieved strong verification performance with minimal FLOP and parameter complexity [18].

2.3. Knowledge Distillation

The concept of Knowledge Distillation was first introduced by Hinton et al. [21]. The primary goal of knowledge distillation is to transfer the knowledge from a pre-trained, complex "teacher" model to a simpler, more efficient "student" model. The methods for distillation in classification tasks can primarily be learned through the utilization of soft labels from a teacher and ground truth [21]. Another approach involves feature-based learning, where the student aims to match the intermediate layers of the teacher [38]. Additionally, contrastive-based methods have also been em-

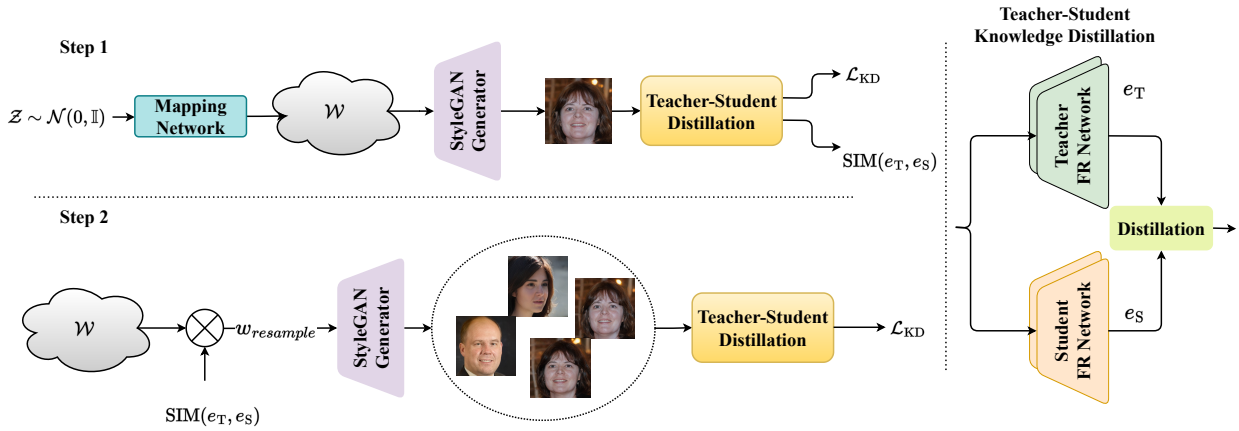

Figure 2. Schematic showing the proposed approach (SynthDistill). In step 1, $\mathcal{Z}$ space of the StyleGAN is sampled to generate face images. In step 2, the $\mathcal{W}$ space is re-sampled based on the teacher-student agreement to generate more challenging samples. The student model is is updated based on the distillation loss $\mathcal{L}_{\text{KD}}$, all the other network blocks remains frozen.

ployed [45] for distilling the knowledge of a teacher to a student.

Over the years, several methods have been proposed in the literature [38, 31, 26, 14, 32, 55, 52, 12] to enhance the efficiency of distillation. However, most of these methods rely on the availability of original or similar training datasets, which can be limited due to security and privacy concerns. Consequently, traditional data-dependent distillation methods become impractical. To address this challenge, researchers have introduced Data-free knowledge distillation (DFKD), without relying on the original or real training data. DFKD aims to develop a distillation strategy using a synthesis-based approach. These approaches utilize either whitebox teacher model [33, 13, 50] or data augmentation techniques [2] to generate synthetic samples. These synthetic samples act as substitute training datasets for distillation. By training on such synthetic data, the student model can effectively learn from the teacher model without needing access to real training data making it privacy friendly. Along the same lines, Boutros et al. [8] proposed an unsupervised face recognition model based on unlabeled synthetic data. They used contrastive learning to maximize the similarity between two augmented images (using geometric and color transformations) of the same synthetic image. However, since the data augmentation cannot provide enough inter-class variations, it affects the performance of trained face recognition model when evaluating on benchmark datasets.

3. Proposed Framework

In this section, we describe our proposed framework for training a lightweight face recognition model using synthetic data using knowledge distillation. We describe the architecture of lightweight face recognition model in Section 3.1 and explain our knowledge distillation framework using synthetic data in Section 3.2.

3.1. Lightweight Network Architecture

As discussed in Section 2, lightweight face recognition models in the literature usually adapt lightweight neural network models for face recognition tasks. However, our knowledge distillation framework can be applied to any lightweight model with only the condition that the output of the lightweight network should have the same dimensions as the embedding of the teacher model. To eliminate this condition so that the proposed framework can be used for any lightweight network with different output sizes, we use a fully connected layer at the output of the lightweight network to have output with the same size as the teacher model.

In this paper, we use TinyNet [20] as the backbone for the lightweight FR model. The TinyNet is an optimized version of EfficientNet [43], which uses a structure that simultaneously enlarges the resolution, depth, and width in a Rubik's cube for neural networks and find networks with high efficiency by changing these three dimensions. However, authors in [20] show that the resolution and depth are more important than width for small networks, and propose smaller models derived from the EfficientNet-B0 as different variations of TinyNet, which are efficient and achieve high accuracy in recognition tasks. The feature layer of TinyNet has 1280 dimensions and the embedding of our teacher network has 512 dimensions. Therefore, we add a fully connected layer to generate 512-length feature at the output of TinyNet and call our lightweight face recognition network based on TinyNet *TinyFaR*. We should note that to our knowledge, TinyNet lightweight network structure has

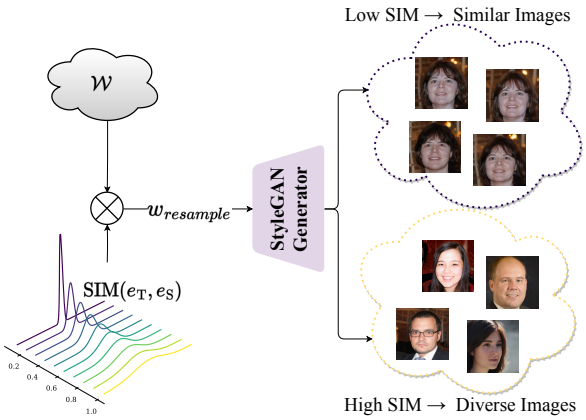

Figure 3. Schematic showing the re-sampling strategy in the proposed approach. When teacher-student agreement is high, the re-sampling method generates diverse images. Conversely, when the similarity is low, i.e, when the given sample is challenging, re-sampling generates similar (challenging) samples facilitating the learning.

not been used before for face recognition in the literature.

3.2. Knowledge Distillation with Synthetic Data

Let F_T and F_S denote the teacher[1] and student (lightweight) face recognition models, respectively. In this paper, we consider StyleGAN [25] as a pretrained face generator model, which consists of a mapping network M and a generator network G. The mapping network takes a a noise $\boldsymbol{z} \in \mathcal{Z} \sim N(0, \mathbb{I})$ from input latent space $\mathcal{Z}$ with Gaussian distribution and generates an intermediate latent code $\boldsymbol{w} \in \mathcal{W}$. Then, the intermediate latent code $\boldsymbol{w}$ is used by the generator network to generate a face image $I = G(\boldsymbol{w})$. In our knowledge distillation framework, we first generate a batch of synthetic face images and extract the teacher's embeddings $e_T = F_T(I)$. Then, we train the student network by minimizing the mean squared error (MSE) of the teacher and student's embeddings as follows:

$$\mathcal{L}_{KD} = \|\boldsymbol{e}_T - F_S(I)\|_2^2 . \quad (1)$$

Minimizing the MSE of embeddings helps the student network to extract embeddings similar to the teacher's embeddings from a given face image.

After updating the weights of student network with our knowledge distillation loss $\mathcal{L}_{KD}$ (as in Eq. 1), we sample around the intermediate latent codes based on the similarity of embeddings extracted by the student $\boldsymbol{e}_S$ and teacher $\boldsymbol{e}_T$ networks in our batch. To this end, we use the cosine

[1]Note that the teacher model can be blackbox and we do not use teacher's gradients in our method.

Algorithm 1 Our proposed knowledge distillation approach

Require: n_{epoch}: number of epochs, $n_{iteration}$: number of iterations in each epoch, α: learning rate, c: re-sampling coefficient.

procedure TRAINING
  Initialize weights θ_S of the student network
  for epoch $= 1, ..., n_{epoch}$ **do**
    for itr $= 1, ..., n_{iteration}$ **do**
      Step 1: Train with random samples
        Sample a batch of random noise vectors:
          $\boldsymbol{z} \in \mathcal{Z} \sim \mathcal{N}(0, \mathbb{I})$
        Generate synthetic face images:
          $\boldsymbol{w} = M(\boldsymbol{z})$
          $\boldsymbol{I} = G(\boldsymbol{w})$
        Extract teacher's embeddings e_T:
          $e_T = F_T(I)$
        Calculate loss $\mathcal{L}_{KD}$ and optimize θ_S:
          $g_{\theta_S} \leftarrow \nabla_{\theta_S} \mathcal{L}_{KD}$
          $\theta_S \leftarrow \theta_S - \alpha \cdot \text{Adam}(\theta_S, g_{\theta_S})$
      Step 2: Train with dynamic re-sampling
        Calculate similarity of e_T and e_S:
          $\boldsymbol{s}_{sim} = \text{SIM}(e_T, e_S)$
        Re-sample based on similarity:
          $\boldsymbol{w}_{resample} = \boldsymbol{w} + c \times \boldsymbol{s}_{sim} \times \boldsymbol{n}$,
            where $\boldsymbol{n} \sim \mathcal{N}(0, \mathbb{I})$
        Generate synthetic face images:
          $\boldsymbol{I} = G(\boldsymbol{w}_{resample})$
        Extract teacher's embeddings e_T:
          $e_T = F_T(I)$
        Calculate loss $\mathcal{L}_{KD}$ and optimize θ_S:
          $g_{\theta_S} \leftarrow \nabla_{\theta_S} \mathcal{L}_{KD}$
          $\theta_S \leftarrow \theta_S - \alpha \cdot \text{Adam}(\theta_S, g_{\theta_S})$
    end for
  end for
end procedure

similarity and normalize it in (0,1) interval as follows:

$$\text{SIM}(\boldsymbol{e}_T, \boldsymbol{e}_S) = 0.5 \times (1 + \frac{\boldsymbol{e}_S \cdot \boldsymbol{e}_T}{\|\boldsymbol{e}_S\|_2 \cdot \|\boldsymbol{e}_T\|_2}). \quad (2)$$

Having normalized similarity score $\boldsymbol{s}_{sim} = \text{SIM}(e_T, e_S)$, we re-sample around each latent code:

$$\boldsymbol{w}_{resample} = \boldsymbol{w} + c \times \boldsymbol{s}_{sim} \times \boldsymbol{n}, \quad (3)$$

where $\boldsymbol{n} \sim \mathcal{N}(0, \mathbb{I})$ is a random noise with Gaussian distribution and c is a constant coefficient. As a matter of fact, in our re-sampling based on similarity score $\boldsymbol{s}_{sim}$ as in Eq. 3, we sample with higher standard deviation values around the latent codes which achieved higher similarity in our initial sampling, and thus letting more variation in re-sampling. While, for the lower similarity between embeddings extracted by the student and teacher networks, the

standard deviation values for re-sampling are smaller so that during re-sampling we can sample around the same latent codes. Therefore, our dynamic re-sampling approach helps us further sample difficult images while exploring the latent space. Fig. 3 illustrates our re-sampling strategy. After re-sampling new latent codes, we generate synthetic face images and optimize our student network with our knowledge distillation loss $\mathcal{L}_{\text{KD}}$ (as in Eq. 1). Our knowledge distillation framework using synthetic data (named *SynthDistill*) is depicted in Fig. 2 and summarized in Algorithm 1.

4. Experiments

In this section, we report our experiments and discuss our results. First, in Section 4.1 we describe our evaluation datasets, and in Section 4.2 we explain our training details. In Section 4.3, we compare our method with previous methods based on synthetic data for face recognition in the literature. Then, we report different ablation studies and discuss effect of each part in our proposed framework in Section 4.4.

4.1. Datasets

We evaluate our trained student models using five different benchmarking datasets. The datasets chosen for evaluation comprised Labeled Faces in the Wild (LFW) [23], Cross-age LFW (CA-LFW) [54], CrossPose LFW (CP-LFW) [53], Celebrities in Frontal-Profile in the Wild (CFP-FP) [40], AgeDB-30 [36]. To maintain consistency with previous work, we present recognition accuracy values on these datasets.

Table 1. Complexity of different network structures

Role in our KD	Network	M FLOPS	M Params
Teacher	IResNet100	24,179.2	65.2
Student	TinyFaR-A	254.3	5.6
	TinyFaR-B	151.3	3.1
	TinyFaR-C	76.8	1.8

4.2. Training Details

For the teacher network, we use the pretrained ArcFace model[2] with IResnet100 backbone from Insightface [16] trained on the MS-Celeb dataset [19]. The embedding of our teacher network has 512 dimensions, but the feature layer of TinyNet has 1280 dimensions. Therefore, as discussed in Section 3 we use a fully connected layer at the output of our TinyNet model so that it can generate embeddings with the same dimension as the teacher's embeddings and call it *TinyFaR*. In our experiments, we use different variations of TinyNet [20] and build corresponding version of TinyFaR with 512-length feature as our student (lightweight) network. Table 1 compares IResnet100 with different variations of TinyFaR in terms of computation complexity and number of parameters. We use StyleGAN2-ADA model [25] to generate synthetic face images with 256×256 resolution and crop and resize images to have 112×112 face images for our knowledge distillation. We train our student networks with 17 epochs, where in each epoch we sampled one million images in step 1 of our algorithm 1 and re-sampled the same number of images with the re-sampling coefficient of $c = 1$. We trained our student networks using Adam optimizer [29] on a system equipped with a single NVIDIA GeForce RTX$^{\text{TM}}$ 3090. For training face recognition from scratch in our experiments, we used CosFace [46] loss function. The source codes of our experiments are publicly available[3].

[2]The performance of our teacher network on our benchmarking datasets in terms of recognition accuracy is as follows: LFW (99.77 ± 0.28), CA-LFW (96.10 ± 1.10), CP-LFW (92.88 ± 1.52), CFP-FP (96.27 ± 1.10), and AgeDB-30 (98.25 ± 0.71).

Table 2. Synthetic and real face datasets

Dataset	#Images	#Subjects	Data	Method
WebFace-4M [56]	4,235,242	205,990	Real	Web-crawled
SFace [7] (IJCB 2022)	1,885,877	10,572	Synthetic	StyleGAN model
DigiFace [3] (WACV 2023)	1,219,995	109,999	Synthetic	Rendering
DCFace [28] (CVPR 2023)	1,300,000	60,000	Synthetic	Diffusion model

4.3. Comparison

We compare the performance of our proposed knowledge distillation framework with training the same network using synthetic datasets in the literature, including DigiFace [3], SFace [7], and DCFace [28]. In addition, we also consider training with real data using WebFace-4M [56] as our baseline. Table 2 compares these datasets in terms of the number of images and samples and their generation method. All these datasets are generated to have inter-class and intra-class variation, and thus have identity labels. Therefore, these datasets can be used for training lightweight face recognition from scratch using the classification training. In contrast, our proposed framework based on dynamic sampling approach does not provide idenity labels and can be used within a knowledge distillation training. Table 3 reports the recognition performance of different variations of TinyFaR when training with datasets. As the results in this table show, our knowledge distillation approach with synthetic data (and no identity labels) far outperforms training from scratch using synthetic data and has comparable performance with training using real data.

4.4. Ablation studies

Effect of dynamic sampling: To evaluate the effect of dynamic sampling in our proposed framework, we compare the performance network trained with knowledge distillation using our dynamic sampling (sampling + re-sampling)

[3]Source code: https://gitlab.idiap.ch/bob/bob.paper.ijcb2023_synthdistill

Table 3. Comparison of our knowledge distillation approach with training from scratch using other synthetic datasets

Network	Training	Dataset	LFW	CA-LFW	CP-LFW	CFP-FP	AgeDB-30
TinyFaR-A	Classification	WebFace-4M (real)	99.58 ± 0.37	95.02 ± 1.00	91.82 ± 1.29	95.09 ± 1.15	94.62 ± 1.21
		DCFace (synthetic)	97.35 ± 0.66	90.08 ± 1.27	79.63 ± 2.08	82.01 ± 1.62	85.12 ± 2.05
		SFace (synthetic)	90.48 ± 1.54	75.48 ± 2.27	71.40 ± 1.89	72.07 ± 2.38	68.65 ± 2.53
		DigiFace (synthetic)	89.12 ± 1.30	71.65 ± 2.14	69.63 ± 1.70	76.24 ± 1.34	68.60 ± 1.23
	Knowledge Distillation	**SynthDistill (synthetic) [ours]**	**99.52 ± 0.31**	**94.57 ± 1.01**	**87.00 ± 1.64**	**90.89 ± 1.54**	**94.93 ± 1.35**
TinyFaR-B	Classification	WebFace-4M (real)	99.55 ± 0.40	94.73 ± 0.88	90.95 ± 1.43	94.00 ± 1.23	93.72 ± 1.37
		DCFace (synthetic)	97.40 ± 0.75	89.62 ± 1.37	78.93 ± 1.74	82.47 ± 1.74	85.03 ± 1.97
		SFace (synthetic)	91.10 ± 1.22	76.15 ± 1.46	72.02 ± 1.34	71.13 ± 2.43	68.73 ± 1.68
		DigiFace (synthetic)	88.03 ± 1.05	70.27 ± 2.17	68.22 ± 1.74	75.29 ± 2.14	66.38 ± 1.82
	Knowledge Distillation	**SynthDistill (synthetic) [ours]**	**99.20 ± 0.41**	**93.78 ± 0.78**	**84.93 ± 2.10**	**88.19 ± 1.34**	**93.02 ± 1.30**
TinyFaR-C	Classification	WebFace-4M (real)	99.37 ± 0.26	93.08 ± 1.11	88.98 ± 1.12	92.30 ± 1.74	91.18 ± 1.80
		DCFace (synthetic)	96.78 ± 0.73	88.48 ± 1.02	77.22 ± 1.80	80.59 ± 1.80	83.65 ± 2.14
		SFace (synthetic)	91.12 ± 1.01	76.70 ± 1.25	71.27 ± 1.98	72.24 ± 1.53	71.13 ± 1.35
		DigiFace (synthetic)	87.47 ± 0.87	69.18 ± 1.94	68.05 ± 1.94	74.16 ± 2.68	67.23 ± 1.85
	Knowledge Distillation	**SynthDistill (synthetic) [ours]**	**98.58 ± 0.44**	**91.80 ± 1.04**	**82.00 ± 2.14**	**84.54 ± 1.57**	**88.98 ± 1.49**

Table 4. Ablation study on the effect of dynamic sampling

Sampling	# Samples/epoch	LFW	CA-LFW	CP-LFW	CFP-FP	AgeDB-30
static	1 M	98.87 ± 0.39	92.93 ± 0.94	83.52 ± 1.63	87.60 ± 1.44	91.25 ± 2.18
static	2 M	98.95 ± 0.47	93.67 ± 0.78	84.75 ± 1.94	88.51 ± 1.63	92.83 ± 1.76
dynamic (re-sampling in $\mathcal{Z}$)	1M + 1M	99.28 ± 0.30	93.88 ± 1.09	84.45 ± 1.95	87.59 ± 1.20	92.45 ± 1.69
dynamic (re-sampling in $\mathcal{W}$)	1M + 1M	99.52 ± 0.31	94.57 ± 1.01	87.00 ± 1.64	90.89 ± 1.54	94.93 ± 1.35

using static sampling (with no re-sampling). Table 4 compares the performance of TinyFaR-A trained with knowledge distillation using our dynamic sampling (sampling + re-sampling in $\mathcal{W}$ space) with one million samples plus one million re-sampling (1M+1M) in each epoch as well as static sampling with one million and two million samples in each epoch. As the results in this table show knowledge distillation using our dynamic sampling with one million iterations in each epoch outperforms the same number of iterations or sample total samples with static sampling. This table also compares our dynamic re-sampling in $\mathcal{W}$ space to dynamic re-sampling in $\mathcal{Z}$ space. As the results show dynamic re-sampling in both spaces achieves better performance than static sampling. In addition, comparing dynamic re-sampling space, the results show that dynamic re-sampling in $\mathcal{W}$ leads to superior performance.

Table 5. Ablation study on the effect of number of sampling

# Itr	LFW	CA-LFW	CP-LFW	CFP-FP	AgeDB-30
0.5 M	99.43 ± 0.37	93.90 ± 1.11	86.13 ± 1.81	89.46 ± 1.48	93.53 ± 1.36
1 M	99.52 ± 0.31	94.57 ± 1.01	87.00 ± 1.64	90.89 ± 1.54	94.93 ± 1.35
2 M	99.48 ± 0.39	95.07 ± 0.97	87.78 ± 1.64	91.31 ± 1.99	95.25 ± 1.19

Effect of number of sampled images: To evaluate the effect of the number of sample images in our dynamic sampling, we train TinyFaR-A with different numbers of iterations (sampling and re-sampling) per epoch in our knowledge distillation approach. Table 6 reports the performance of the trained model with different numbers of iterations. As the results in this table show, higher iterations help our knowledge distillation with the cost of more training computation. However, to reduce computations in our experiments we use one million iterations (1M sampling + 1M re-sampling) in our experiments.

Table 6. Ablation study on the effect of re-sampling coefficient

coef. (c)	LFW	CA-LFW	CP-LFW	CFP-FP	AgeDB-30
0.8	99.45 ± 0.32	94.58 ± 0.95	86.10 ± 2.23	90.23 ± 1.68	94.82 ± 1.15
0.9	99.40 ± 0.41	94.90 ± 1.09	87.23 ± 2.02	90.36 ± 1.45	94.72 ± 1.07
1	99.52 ± 0.31	94.57 ± 1.01	87.00 ± 1.64	90.89 ± 1.54	94.93 ± 1.35
1.1	99.47 ± 0.44	94.95 ± 0.84	87.53 ± 1.78	90.81 ± 1.61	95.13 ± 1.08
1.2	99.53 ± 0.32	94.95 ± 0.90	87.47 ± 1.27	90.94 ± 1.63	94.52 ± 1.47
1.3	99.52 ± 0.31	94.50 ± 0.97	87.58 ± 1.84	91.17 ± 1.50	95.05 ± 1.28
1.4	99.48 ± 0.32	94.77 ± 0.97	87.40 ± 1.74	90.56 ± 1.49	94.78 ± 1.29
1.5	99.47 ± 0.32	94.58 ± 1.00	88.17 ± 1.64	90.84 ± 1.24	94.80 ± 1.07

Effect of re-sampling coefficient: As another ablation study, we evaluate the effect of re-sampling coefficient c in our dynamic sampling. Table 6 reports the performance of TinyFaR-A trained with our knowledge distillation using different re-sampling coefficient values. As the results in this table show, with a higher re-sampling coefficient our dynamic re-sampling can generate more diverse images and achieve higher recognition performance. However, a very high re-sampling coefficient can also cause $\boldsymbol{w}_{\text{resample}}$ to be out of the distribution of $\mathcal{W}$, and thus drop the performance.

5. Discussions

The results in Table 3 show that our proposed knowledge distillation framework outperforms training using synthetic datasets in the literature and achieves comparable performance with training using real face images. Comparing the performance of networks trained with previous synthetic datasets to networks trained with real data, we observe a considerable gap in the performance of trained face recognition models with synthetic and real data. Meanwhile, our proposed knowledge distillation method still achieves lower but is very close to the performance of training with real data.

Unlike previous synthetic face datasets, our method does not require identity labels, and thus does not have many issues in generating synthetic datasets with inter-class and intra-class variations. Instead, our knowledge distillation approach with dynamic sampling leverages the most capacity of StyleGAN to generate training samples, which helps to achieve comparable performance to training with real data. Our proposed framework avoids the requirements of hard identity labels for the generated images, which further assists the generation network to produce challenging samples though a feedback mechanism during our knowledge distillation, thus enabling the training of much robust models. We should also note that, for generation of synthetic face datsets in the literature, a pretrained face recognition model (which has been trained on a large-scale real face recognition dataset) is used in the process of generation of synthetic dataset. Therefore, training with synthetic face datasets in the literature indirectly benefits from the information and knowledge of the pretrained face recognition model (trained on real images) used for generating the synthetic dataset. In our proposed framework, we also use the pretrained face recognition model, but instead of following common two-step approach (generation of dataset and training with new dataset), we use the pretrained face recognition model as a teacher in our knowledge distillation approach and generate synthetic face images used in our training with no identity label.

Our ablation studies show the effect of each part in our knowledge distillation framework. In particular, the results demonstrate that our dynamic sampling improves our knowledge distillation compared to static sampling. In addition, using our dynamic sampling and with more number of iterations or higher re-sampling coefficient can improve the knowledge distillation, as it helps our student to learn embeddings of more face images from the teacher.

6. Conclusions

In this paper, we proposed a data-free framework (named *SynthDistill*) to train lightweight face recognition models based on knowledge distillation using synthetic data. We combined the two steps of data generation and training the lightweight network and have an online-generation and training in the loop using a distillation framework. We dynamically generated synthetic face images during training and distilled the knowledge of a pretrained and blackbox face recognition model. Our dynamic sampling helps our student network to further see difficult samples while exploring new samples, leading to more robust training. Our knowledge distillation framework does not require identity-labeled training data, and thus mitigates challenges in generating intra-class variations in synthesized datasets. We adapted the TinyNet architecture to use in our knowledge distillation framework and trained lightweight face recognition models (called *TinyFaR*). We reported extensive experimental evaluation on five different face recognition benchmarking datasets, including LFW, CA-LFW, CP-LFW, CFP-FP, and AgeDB-30. The experimental results demonstrate the superiority of our proposed knowledge distillation approach compared to training previous synthetic datasets.

Our experimental results also showed that while there is a considerable gap between training with synthetic datasets and real data, our knowledge distillation framework based on synthetic data achieves comparable performance with training with real data and significantly reduces the gap between models trained on synthetic data and models trained on real data. Achieving such an improvement in training using synthetic data within our proposed framework shows more potential in training with synthetic data and motivates further research on training with synthetic data. Furthermore, our results for lightweight student networks pave the way for developing privacy-aware and resource-efficient face recognition models.

Acknowledgments

This research is based upon work supported by the H2020 TReSPAsS-ETN Marie Skłodowska-Curie early training network (grant agreement 860813).

This research is also based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2022-21102100007]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] M. Alansari, O. A. Hay, S. Javed, A. Shoufan, Y. Zweiri, and N. Werghi. Ghostfacenets: Lightweight face recognition model from cheap operations. *IEEE Access*, 2023.

[2] Y. M. Asano and A. Saeed. The augmented image prior: Distilling 1000 classes by extrapolating from a single image. *arXiv preprint arXiv:2112.00725*, 2021.

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

[4] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua. Towards open-set identity preserving face synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6713–6722, 2018.

[5] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[6] F. Boutros, N. Damer, M. Fang, F. Kirchbuchner, and A. Kuijper. Mixfacenets: Extremely efficient face recognition networks. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021.

[7] F. Boutros, M. Huber, P. Siebke, T. Rieber, and N. Damer. Sface: Privacy-friendly and accurate face recognition using synthetic data. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11. IEEE, 2022.

[8] F. Boutros, M. Klemt, M. Fang, A. Kuijper, and N. Damer. Unsupervised face recognition using unlabeled synthetic data. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2023.

[9] F. Boutros, P. Siebke, M. Klemt, N. Damer, F. Kirchbuchner, and A. Kuijper. Pocketnet: Extreme lightweight face recognition network using neural architecture search and multi-step knowledge distillation. *IEEE Access*, 10:46823–46833, 2022.

[10] F. Boutros, V. Struc, J. Fierrez, and N. Damer. Synthetic data for face recognition: Current state and future prospects. *Image and Vision Computing*, page 104688, 2023.

[11] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[12] D. Chen, J.-P. Mei, C. Wang, Y. Feng, and C. Chen. Online knowledge distillation with diverse peers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 3430–3437, 2020.

[13] H. Chen, Y. Wang, C. Xu, Z. Yang, C. Liu, B. Shi, C. Xu, C. Xu, and Q. Tian. Data-free learning of student networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3514–3522, 2019.

[14] P. Chen, S. Liu, H. Zhao, and J. Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5008–5017, 2021.

[15] S. Chen, Y. Liu, X. Gao, and Z. Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13*, pages 428–438. Springer, 2018.

[16] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019.

[17] Y. Deng, J. Yang, D. Chen, F. Wen, and X. Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5154–5163, 2020.

[18] A. George, C. Ecabert, H. O. Shahreza, K. Kotwal, and S. Marcel. Edgeface: Efficient face recognition model for edge devices. *arXiv preprint arXiv:2307.01838*, 2023.

[19] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016.

[20] K. Han, Y. Wang, Q. Zhang, W. Zhang, C. Xu, and T. Zhang. Model rubik's cube: Twisting resolution, depth and width for tinynets. *Advances in Neural Information Processing Systems*, 33:19353–19364, 2020.

[21] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[22] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

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

[24] A. Jain, N. Memon, and J. Togelius. Zero-shot racially balanced dataset generation using an existing biased stylegan2. *arXiv preprint arXiv:2305.07710*, 2023.

[25] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.

[26] J. Kim, S. Park, and N. Kwak. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems*, 31, 2018.

[27] M. Kim, A. K. Jain, and X. Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18750–18759, 2022.

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

[29] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, California., USA, May 2015.

[30] J. N. Kolf, F. Boutros, J. Elliesen, M. Theuerkauf, N. Damer, M. Alansari, O. A. Hay, S. Alansari, S. Javed, N. Werghi,

et al. Efar 2023: Efficient face recognition competition. *arXiv preprint arXiv:2308.04168*, 2023.

[31] N. Komodakis and S. Zagoruyko. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017.

[32] Z. Li, X. Li, L. Yang, B. Zhao, R. Song, L. Luo, J. Li, and J. Yang. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1504–1512, 2023.

[33] R. G. Lopes, S. Fenu, and T. Starner. Data-free knowledge distillation for deep neural networks. *arXiv preprint arXiv:1710.07535*, 2017.

[34] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.

[35] Y. Martindez-Diaz, L. S. Luevano, H. Mendez-Vazquez, M. Nicolas-Diaz, L. Chang, and M. Gonzalez-Mendoza. Shufflefacenet: A lightweight face architecture for efficient and highly-accurate face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

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

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

[38] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[39] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

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

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

[42] Y. Shen, P. Luo, J. Yan, X. Wang, and X. Tang. Faceidgan: Learning a symmetry three-player gan for identity-preserving face synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 821–830, 2018.

[43] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[44] M. Tan and Q. V. Le. Mixconv: Mixed depthwise convolutional kernels. *arXiv preprint arXiv:1907.09595*, 2019.

[45] Y. Tian, D. Krishnan, and P. Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

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

[47] X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.

[48] M. Yan, M. Zhao, Z. Xu, Q. Zhang, G. Wang, and Z. Su. Vargfacenet: An efficient variable group convolutional neural network for lightweight face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

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

[50] H. Yin, P. Molchanov, J. M. Alvarez, Z. Li, A. Mallya, D. Hoiem, N. K. Jha, and J. Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8715–8724, 2020.

[51] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Towards large-pose face frontalization in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3990–3999, 2017.

[52] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3713–3722, 2019.

[53] T. Zheng and W. Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5(7), 2018.

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

[55] X. Zhu, S. Gong, et al. Knowledge distillation by on-the-fly native ensemble. *Advances in neural information processing systems*, 31, 2018.

[56] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021.