Grouped Knowledge Distillation for Deep Face Recognition

Weisong Zhao [1,3*], Xiangyu Zhu [2,4*], Kaiwen Guo [2], Xiao-Yu Zhang [1,3†], Zhen Lei [2,4,5]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2] CBSR&NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[4] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[5] Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences, Hong Kong, China
{zhaoweisong, zhangxiaoyu}@iie.ac.cn, guokaiwen1991@gmail.com, {xiangyu.zhu, zlei}@nlpr.ia.ac.cn

Abstract

Compared with the feature-based distillation methods, logits distillation can liberalize the requirements of consistent feature dimension between teacher and student networks, while the performance is deemed inferior in face recognition. One major challenge is that the light-weight student network has difficulty fitting the target logits due to its low model capacity, which is attributed to the significant number of identities in face recognition. Therefore, we seek to probe the target logits to extract the primary knowledge related to face identity, and discard the others, to make the distillation more achievable for the student network. Specifically, there is a tail group with near-zero values in the prediction, containing minor knowledge for distillation. To provide a clear perspective of its impact, we first partition the logits into two groups, i.e., Primary Group and Secondary Group, according to the cumulative probability of the softened prediction. Then, we reorganize the Knowledge Distillation (KD) loss of grouped logits into three parts, i.e., Primary-KD, Secondary-KD, and Binary-KD. Primary-KD refers to distilling the primary knowledge from the teacher, Secondary-KD aims to refine minor knowledge but increases the difficulty of distillation, and Binary-KD ensures the consistency of knowledge distribution between teacher and student. We experimentally found that (1) Primary-KD and Binary-KD are indispensable for KD, and (2) Secondary-KD is the culprit restricting KD at the bottleneck. Therefore, we propose a Grouped Knowledge Distillation (GKD) that retains the Primary-KD and Binary-KD but omits Secondary-KD in the ultimate KD loss calculation. Extensive experimental results on popular face recognition benchmarks demonstrate the superiority of proposed GKD over state-of-the-art methods.

Introduction

Face recognition has achieved great success in various application domains (Li and Jain 2011; Lei, Pietikäinen, and Li 2014). However, a large number of light-weight yet discriminative face recognition models are required due to the development of mobile and edge devices. An intuitive solution is to optimize the neural network architectures for mobile devices, e.g., MobileFaceNet (Chen et al. 2018) and MobileNetV3 (Howard et al. 2019). However, discriminative networks always benefit from a large model capacity, which will introduce more computational and storage costs. Given a heavy teacher network, Knowledge Distillation (KD) aims to improve the accuracy of light-weight networks, where the knowledge from a heavy teacher network is transferred to the light-weight student network.

The idea of KD (Hinton, Vinyals, and Dean 2015) was first introduced to transfer knowledge by reducing the Kullback-Leibler (KL) divergence between the prediction probabilities of the teacher and the student networks. In the past decade, the research attention has been drawn to conducting instance-wise constraints on the activation of the hidden layers, e.g., FitNet (Romero et al. 2015) distills knowledge from deep features of intermediate layers. How-

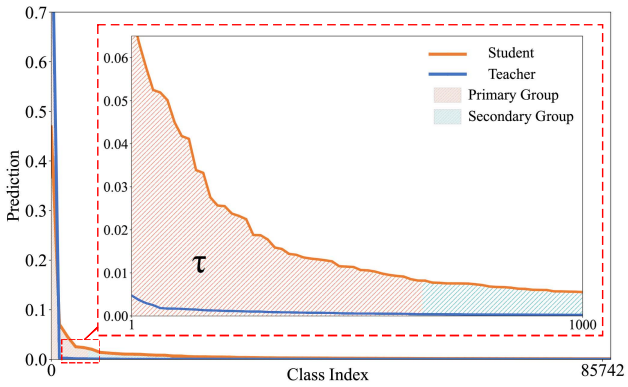

Figure 1: The distribution of the predictions (softened probabilities) of student and teacher networks trained from scratch. The analysis is conducted on MS1MV2 (Deng et al. 2019a) dataset. The predictions are visualized based on a randomly selected sample. There is a long tail group that has several near-zero values, containing minor knowledge for distillation. We partition the original logits into two groups, i.e., primary group (pink) and secondary group (blue), via a cumulative probability threshold τ of ranked student prediction. Best viewed in color.

*These authors contributed equally.
†Corresponding Author.

ever, such feature-based methods require the teacher and student networks to share the same representation space, which is unrealistic for student networks with low model capacities (Huang et al. 2022). Additionally, extra computational and storage usage (e.g., additional network modules and identical feature dimension requirements) are introduced for distilling deep features.

Unlike feature-based distillation, logits distillation does not require the student to mimic the teacher's representation space, but rather to preserve the high semantic consistency with the teacher, which can liberalize the requirements of consistent feature dimension between teacher and student networks. Unfortunately, the performance of logits distillation is inferior in large-scale face recognition. One major challenge is that the light-weight student network has difficulty fitting the target logits due to its low model capacity, which is attributed to the great number of identities in face recognition. Therefore, we seek to probe the target logits to extract the primary knowledge related to face identity and discard the others, to make the distillation more achievable for the student network. Specifically, we can see from Fig. 1 that there is a tail group with near-zero values in the softened prediction, which contains minor knowledge for distillation and wastes the learning capabilities of student models.

To provide a clear perspective for the impact of the tail group, we partition the output logits into two groups via the cumulative probability threshold of the student's softened prediction, i.e., primary group and secondary group. We argue that secondary group contains minor knowledge and can not be covered by the student network with low model capacity. To more distinctly embody the role of secondary group in the classic KD, we reorganize the classical KD loss into three parts, i.e., Primary-KD, Secondary-KD, and Binary-KD. Primary-KD distills the most discriminative knowledge from the teacher, Secondary-KD aims at distilling minor knowledge embedded in the tail group, and Binary-KD ensures the consistency of knowledge distribution between teacher and student. The extensive experiments prove that (1) Primary-KD and Binary-KD are indispensable elements for KD, and (2) Secondary-KD is the culprit restricting KD at the bottleneck. Therefore, we propose a Grouped Knowledge Distillation (GKD), which remains Primary-KD and Binary-KD in the ultimate loss calculation. The extensive comparisons with current SOTA methods on several popular face benchmarks demonstrate the superiority of the proposed GKD.

Overall, the contributions of this paper are summarized as follows:

- We propose to find an achievable distillation method for the student network to bridge the performance gap between teacher and student models. Specifically, we introduce the grouped logits to partition the logits into two groups, i.e., primary group and secondary group, via cumulative probability threshold of corresponding softened prediction.
- Given the grouped logits, we propose reorganizing the classical KD loss into three parts, i.e., Primary-KD, Secondary-KD, and Binary-KD. Specifically, we experimentally analyze and prove the individual impacts of these three components; that is, Primary-KD and Binary-KD are indispensable for KD, and Secondary-KD is the culprit restricting KD at the bottleneck. On the basis of the reformulation of KD loss, we propose a Grouped Knowledge Distillation that retains the Primary-KD and Binary-KD in the KD loss computation.
- The extensive experiments on popular face recognition benchmarks demonstrate the superiority of the proposed GKD over the state-of-the-art methods.

Related Work

Loss Functions. There are two types of loss functions applied for face recognition. The first is the verification loss function. Contrastive loss (Chopra, Hadsell, and LeCun 2005; Sun et al. 2014) optimizes pairwise Euclidean distance in feature space. Triplet loss (Schroff, Kalenichenko, and Philbin 2015; Hoffer and Ailon 2015) makes up triplets to separate the positive pair from the negative pair by a distance margin. The second is the softmax-based loss function, which is mostly adopted by current SOTA deep face recognition methods. To learn angularly discriminative features, the SphereFace (Liu et al. 2017) was proposed by introducing the angular SoftMax function (i.e., A-SoftMax), which imposes discriminative constraints on a hypersphere manifold. Also, the CosFace (Wang et al. 2018) with a large margin cosine loss was proposed further to maximize the decision margin in the angular space. The additive angular margin loss ArcFace loss was designed to obtain highly discriminative features for FR (Deng et al. 2019a). CurricularFace (Huang et al. 2020) embeds the idea of curriculum learning into the loss function. Softmthatsed loss functions combined with heavy neural networks are demonstrated to obtain satisfactory performance (Deng et al. 2019a), but they cannot be well deployed in the light-weight network in the mobile sciences (Deng et al. 2019b). Specifically, a performance gap exists between heavy and mobile models, which requires the application of knowledge distillation.

Knowledge Distillation. Knowledge distillation was first proposed by Hinton *et al.* (Hinton, Vinyals, and Dean 2015), which refers to a model compression method to transfer the knowledge of a heavy teacher model to a light-weight student network. Later variants of distillation strategies are proposed to utilize diverse information from the teacher model, which can be divided into two types, i.e., logits distillation and feature distillation. As for the feature distillation, FitNet (Romero et al. 2015) bridges the middle layers of the student and teacher networks and adopted L_2 loss to further supervised the output of the student. ShrinkTeaNet (Duong et al. 2019) proposes minimizing the angle between teacher and student embedding vectors. TripletDistillation (Feng et al. 2020) improves the triplet loss with dynamic margins by utilizing the similarity structures among different identities in the teacher network. MarginDistillation (Svitov and Alyamkin 2020) utilizes class prototypes from the teacher network for the student network. Additionally, SP (Tung and Mori 2019) adopt the pairwise similarities of the outputs. EKD introduces the evaluation-oriented method to optimize

the student model's critical relations (Huang et al. 2022). Most feature distillation strategies could perform better than logits distillation but introduce more computational and storage costs.

Different from feature-based distillation methods, logits distillation does not require the student to mimic the teacher's representation space, but rather to preserve the high semantic consistency with the teacher. The classical KD (Hinton, Vinyals, and Dean 2015) proposes to minimize the Kullback-Leibler Divergence of softened class probabilities between the teacher and student. Besides, DML (Zhang et al. 2018) suggests mutual learning to train multiple networks simultaneously. Mirzadeh *et al.* (Mirzadeh et al. 2020) introduce multi-step knowledge distillation, which employs an intermediate-sized network to bridge the gap between the student and the teacher. Li *et al.* propose a nested collaborative learning structure (Li et al. 2022). DKD (Zhao et al. 2022) firstly proposes to decouple the classical KD Loss into target class knowledge distillation and non-target class knowledge distillation. However, we argue that DKD has inferior performance over large-scale face recognition because considerable logits will dramatically increase the difficulty of distillation.

Method

In this section, we first introduce the primary group and secondary group. Then, we reformulate the classical KD Loss into three parts, i.e., Primary Knowledge Distillation, Secondary Knowledge Distillation, and Binary Knowledge Distillation, and describe the proposed Grouped Knowledged Distillation (GKD) loss.

Grouped Logits

From Figure 1, we find that the prediction of the student network includes a long tail group with several near-zero values, which contains minor knowledge but increases the difficulty of distillation. Therefore, we seek to probe the target logits to extract the primary knowledge related to face identity, to make the distillation more viable for the student network. Specifically, we propose partitioning the original logits into two groups, i.e., primary group and secondary group. We denote the training dataset including n facial images of y identities as $D = \{x_t, y_t\}$, where x_t refers to t-th sample and y_t indicates its identity label. For a sample of i-th identity, the softened probabilities is formulated as $\mathbf{p} = \{p_i\}, i = 1, 2, \cdots, C$, where C denotes the number of identities. For each $p_i \in \mathbf{p}$:

$$p_i = \frac{exp(z_i)}{\sum_{j=1}^{C} exp(z_j)}, \tag{1}$$

where z_i indicates the logit output from the network. According to the above formulation and the results shown in Figure 1, there is a tail group with near-zero values in the prediction, which includes minor knowledge but increases the difficulty of distillation for the student with low model capacity. To further explore the effects of the tail group, we propose to separate the tail group from the original logits via the cumulative probability threshold of the softened prediction, as shown in Figure 2. Specifically, given a ranked logits $\mathbf{z} = [z_1, z_2, \cdots, z_C]$ for sample x, we first obtain its ranked prediction $\widetilde{\mathbf{p}}$. Then, we partition the original logits into two groups, i.e., primary group and secondary group, and respectively denote them as $\mathbf{z}_\Phi$ and $\mathbf{z}_\Psi$, which can be formulated as follow:

$$\begin{aligned} \mathbf{z}_\Phi &= TopK(\mathbf{z}, k), \\ \mathbf{z}_\Psi &= \mathbf{z} \setminus \mathbf{z}_\Phi. \end{aligned} \tag{2}$$

We analyze that the primary group contains most of the discriminative knowledge for distillation where elements are predicted with large values from the model and vice versa for secondary group. To adaptively determine the value of k, we utilize a cumulative probability threshold of the student's prediction, which can be formulated as follows:

$$k = \arg\min_k |\sum_{i=1}^{k} \widetilde{p}_i - \tau|, \tag{3}$$

The τ denotes the cumulative probability threshold. Generally, as the value of τ increases, the model distills more knowledge, i.e., fitting a larger proportion of target logits from the teacher network. The effects of τ are further studied in the "Effects of Threshold τ" section.

Grouped Knowledge Distillation

The secondary group (tail group) contains minor knowledge for distillation, which we argue wastes the learning capability of the student model. To figure out the effects of secondary group, we review the formulation of the classical KD loss that refers to the Kullback-Leibler Divergence of the softened prediction between teacher and student networks. The classical KD loss can be formulated as follows:

$$KL(\mathbf{p}^T||\mathbf{p}^S) = \sum_{i=1}^{C} p_i^T log(\frac{p_i^T}{p_i^S}), \tag{4}$$

where $\mathbf{p}^T$ and $\mathbf{p}^S$ denotes the prediction of teacher and student, where the elements p_i^T and p_i^S are formulated as Equation 1. Let $\mathbf{\Phi}$, $\mathbf{\Psi}$, and $\mathbf{U}$ denote the corresponding identity sets of primary group, secondary group, and whole logits, respectively. Then, we can obtain primary group prediction $\hat{\mathbf{p}}$ and secondary group prediction $\check{\mathbf{p}}$, which are denoted as $\hat{\mathbf{p}} = \{\hat{p}_i\}, i = 1, 2, \cdots, C_\Phi$, and $\check{\mathbf{p}} = \{\check{p}_j\}, j = 1, 2, \cdots, C_\Psi$, respectively. C_Φ and C_Ψ represent the length of $\hat{\mathbf{p}}$ and $\check{\mathbf{p}}$, and the elements can be denoted as follows:

$$\begin{aligned} \hat{p}_i &= \frac{exp(z_i)}{\sum_{l\in\Phi} exp(z_l)} \mid i \in \Phi, \\ \check{p}_j &= \frac{exp(z_j)}{\sum_{l\in\Psi} exp(z_l)} \mid j \in \Psi. \end{aligned} \tag{5}$$

We define the notion $\mathbf{p}_b$ to represent the binary logits that indicates the knowledge distribution, which is denoted as $\mathbf{p}_b = [p_\Phi, p_\Psi]$. Additionally, two notions $\mathbf{p}_\Phi$and $\mathbf{p}_\Psi$ are denoted to represent the binary probability of primary group and secondary group, which can be computed as follows:

$$p_\Phi = \frac{\sum_{l\in\Phi} exp(z_l)}{\sum_{l\in\mathbf{U}} exp(z_l)}, \quad p_\Psi = \frac{\sum_{l\in\Psi} exp(z_l)}{\sum_{l\in\mathbf{U}} exp(z_l)}. \tag{6}$$

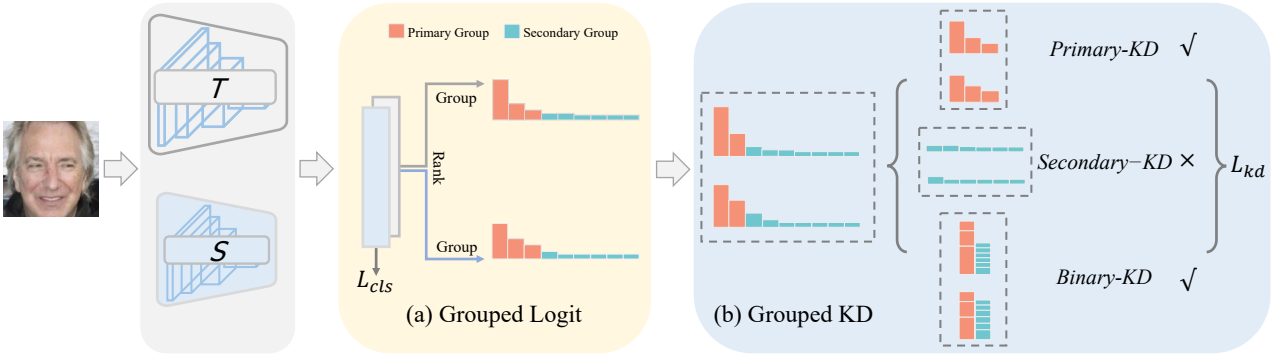

Figure 2: The main framework of the proposed Grouped Knowledge Distillation (GKD). First, a facial image is fed into teacher and student networks to obtain the corresponding logits output. Then, (a) both the logits of teacher and student are ranked and partitioned into two groups, i.e., primary group and secondary group, via a cumulative probability threshold τ of student prediction. (b) With the grouped logits, the classical Knowledge Distillation loss is reorganized and divided into three parts, i.e., Primary-KD, Secondary-KD, and Binary-KD. The extensive experiments prove that (1) Primary-KD and Binary-KD are indispensable for KD, and (2) Secondary-KD is the culprit restricting KD at the bottleneck. Therefore, the proposed Grouped Knowledge Distillation (GKD) retains the Primary-KD and Binary-KD but omits Secondary-KD in the ultimate KD loss calculation. Additionally, CosFace loss is utilized as the classification loss to maintain the intra-class and inter-class relations.

According to Equation 5 and Equation 6, we can reorganize the formulation of KD loss in Equation 4 with the grouped logits, which can be formulated as follows:

$$
\begin{aligned}
KL(\mathbf{p}^T||\mathbf{p}^S) &= \sum_{i=1}^{C} p_i^T log(\frac{p_i^T}{p_i^S}) \\
&= \sum_{i\in\Phi} p_i^T log(\frac{p_i^T}{p_i^S}) + \sum_{j\in\Psi} p_j^T log(\frac{p_j^T}{p_j^S}) \\
&= \sum_{i\in\Phi} \hat{p_i}^T p_\Phi^T log(\frac{\hat{p_i}^T p_\Phi^T}{\hat{p_i}^S p_\Phi^S}) + \sum_{j\in\Psi} \check{p_j}^T p_\Psi^T log(\frac{\check{p_j}^T p_\Psi^T}{\check{p_j}^S p_\Psi^S}) \quad (7) \\
&= p_\Phi^T \sum_{i\in\Phi} \hat{p_i}^T log(\frac{\hat{p_i}^T}{\hat{p_i}^S}) + p_\Phi^T log(\frac{p_\Phi^T}{p_\Phi^S}) \sum_{i\in\Phi} \hat{p_i}^T \\
&+ p_\Psi^T \sum_{j\in\Psi} \check{p_j}^T log(\frac{\check{p_j}^T}{\check{p_j}^S}) + p_\Psi^T log(\frac{p_\Psi^T}{p_\Psi^S}) \sum_{j\in\Psi} \check{p_j}^T,
\end{aligned}
$$

where $\mathbf{p}_b = [p_\Phi, p_\Psi]$ represents the binary logits that indicates the knowledge distribution. From Equation 5, we have:

$$
\sum_{i\in\Phi} \hat{p_i}^T = \sum_{j\in\Psi} \check{p_j}^T = 1. \quad (8)
$$

Therefore, the KD loss can be further formulated as follows:

$$
\begin{aligned}
KL(\mathbf{p}^T||\mathbf{p}^S) &= p_\Phi^T \sum_{i\in\Phi} \hat{p_i}^T log(\frac{\hat{p_i}^T}{\hat{p_i}^S}) \\
&+ p_\Psi^T \sum_{j\in\Psi} \check{p_j}^T log(\frac{\check{p_j}^T}{\check{p_j}^S}) + [p_\Phi^T log(\frac{p_\Phi^T}{p_\Phi^S}) + p_\Psi^T log(\frac{p_\Psi^T}{p_\Psi^S})] \\
&= p_\Phi^T \cdot KL(\hat{\mathbf{p}}^T||\hat{\mathbf{p}}^S) + p_\Psi^T \cdot KL(\check{\mathbf{p}}^T||\check{\mathbf{p}}^S) + KL(\mathbf{p}_b^T||\mathbf{p}_b^S). \quad (9)
\end{aligned}
$$

From the Equation 9, we reformulate the KD loss into three parts, i.e., Primary-KD, Secondary-KD, and Binary-KD, which can be denoted as follows:

$$
\begin{aligned}
\text{Primary-KD} &= KL(\hat{\mathbf{p}}^T||\hat{\mathbf{p}}^S), \\
\text{Secondary-KD} &= KL(\check{\mathbf{p}}^T||\check{\mathbf{p}}^S), \quad (10) \\
\text{Binary-KD} &= KL(\mathbf{p}_b^T||\mathbf{p}_b^S).
\end{aligned}
$$

We analyze that Primary-KD refers to distilling the primary knowledge from the teacher, Secondary-KD aims at distilling minor knowledge but increases the difficulty of distillation simultaneously, and Binary-KD ensures the consistency of knowledge distribution between teacher and student. To reduce the difficulty of fitting large-scale target logits for a light-weight student network and make the distillation task more achievable, we retain the Primary-KD and Binary-KD as distillation loss, and assign proper weight to Primary-KD, which is calculated as follows:

$$
\mathcal{L}_{kd} = \lambda_1 * KL(\hat{\mathbf{p}}^T||\hat{\mathbf{p}}^S) + \lambda_2 * KL(\mathbf{p}_b^T||\mathbf{p}_b^S). \quad (11)
$$

Additionally, we utilize ArcFace (Deng et al. 2019a) as our classification loss function to maintain intraclass differentiation and interclass aggregation, which is formulated as:

$$
\mathcal{L}_{cls} = \frac{1}{N_i} \sum_i -log \frac{e^{s(cos(\theta_{y_i}+m,i))}}{e^{s(cos(\theta_{y_i}+m,i))} + \sum_{j\neq y_i} e^{scos(\theta_j,i)}}, \quad (12)
$$

The overall loss function is denoted as:

$$
\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{kd}. \quad (13)
$$

Experiments

Dataset

Training Set. We utilize the refined MS1MV2 (Deng et al. 2019a) as our training set for fair comparisons with other SOTA methods. MS1MV2 consists of 5.8M facial images of 85K individuals.

Testing Set. We evaluate our method on several popular face benchmarks, including LFW (Huang et al. 2008), CFP-FP (Sengupta et al. 2016), CPLFW (Zheng and Deng 2018), AgeDB (Moschoglou et al. 2017), CALFW (Zheng, Deng, and Hu 2017), IJB-B (Whitelam et al. 2017), IJB-C (Maze et al. 2018). LFW is the most commonly utilized face verification dataset, which consists of 13,233 facial images of 5,749 individuals. Cross-Age LFW (CALFW) and Cross-Pose LFW (CPLFW) databases are constructed based on the LFW database, to emphasize similar-looking challenges, cross-age and cross-pose challenges. CFP-FP database is built for facilitating large pose variation and the AgeDB-30 database is a manually collected cross-age database. The IJB-B and IJB-C are two challenging public template-based benchmarks for face recognition. The IJB-B dataset contains 1,845 subjects with 21.8K still images and 55K frames from 7,011 videos. The IJB-C dataset is a further extension of IJB-B, which contains about 3,500 identities with a total of 31,334 images and 11,7542 unconstrained video frames. MegaFace Challenge (Kemelmacher-Shlizerman et al. 2016) consists of the gallery set including 1M images of 690K subjects and the probe set including 100K photos of 530 individuals from FaceScrub.

Implementation Details

Data Processing. We adopt the preprocess settings in ArcFace (Deng et al. 2019a). The input facial images are cropped to 112×112 and normalized by subtracting 127.5 and dividing by 128. For the data augmentation, we apply the horizontal flip with a probability of 50%.

Training. We utilize IResnet-50 as the teacher model that is trained by ArcFace (Deng et al. 2019a), which is pretrained and frozen for all face recognition model training. Additionally, we adopt two student network architectures, e.g., IResnet-18 (Deng et al. 2019a) and MobileFacenet(Chen et al. 2018), following the network architecture settings of EKD (Huang et al. 2022). To show the generality of our method, we use two neural student networks, i.e., MobileFaceNet (Chen et al. 2018) and IResnet-18 (Deng et al. 2019a). We set the batch size to 128 for each GPU in all experiments, and train models on 8 NVIDIA Tesla V100 (32GB) GPUs. We apply the SGD optimization method and divide the initial learning rate (0.1) at 10, 18, and 24 epochs. The momentum is set to 0.9, and the weight decay is 5e-4. The hyper-parameters λ_1 and λ_2 are set to 8.0 and 1.0, respectively. For ArcFace, we follow the common setting with s = 64 and margin m = 0.5.

Testing. We follow the evaluation protocol and network settings proposed by EKD (Huang et al. 2022) to report the performance on LFW, CFP-FP, CPLFW, AgeDB, CALFW, IJB-B, IJB-C, and MegaFace Challenge.

Ablation Study

In this section, we first conduct the ablation experiments on Primary-KD, Secondary-KD, and Binary-KD. Then, we explore the effects of different cumulative probability threshold τ and that of hyper-parameters λ_1 and λ_2. Additionally, we investigate the generalization capability of our method with different student network structures, i.e., IResNet-18 and MobileFacenet. All the experiments are evaluated on five popular face benchmarks, i.e., LFW, CFP-FP, CPLFW, AgeDB, and CALFW.

P-KD	B-KD	S-KD	CFP-FP	CPLFW	AgeDB	CALFW
✓	✓	✓	91.71	87.85	95.93	95.03
✓			94.08	90.56	96.73	95.45
✓	✓		**94.35**	**90.86**	**97.25**	**95.78**

Table 1: Ablation experiments of Primary-KD (P-KD), Secondary-KD (S-KD), and Binary-KD (B-KD).

Ablation on GKD. We divide the classical KD loss into three parts, i.e., Primary-KD, Secondary-KD, and Binary-KD. To investigate their effects, we conduct the ablation experiments on them, and we can see from Table 1 that the classical KD loss achieves inferior performance, and the recognition performance achieves a significant improvement on five testing sets when we omit the Secondary-KD while retaining the Primary-KD and Secondary-KD. This phenomenon indicates the effectiveness of distilling primary knowledge and attributes Secondary-KD as the culprit for the bottleneck of knowledge distillation. We analyze that liberalizing the requirements for the student network to fit secondary knowledge can reduce the difficulty of distillation, thus bridging the performance gap between teacher and student. Additionally, Binary-KD also brings a slight performance improvement, since it specifies the knowledge distribution of model output and keeps consistent knowledge distribution between teacher and student networks to avoid the overfitting of the student model. From the results shown in Table 1, we experimentally find that (1) Primary-KD and Binary-KD are indispensable for KD, and (2) Secondary-KD is the culprit restricting KD at the bottleneck. Therefore, we propose a Grouped Knowledge Distillation (GKD), which retains the Primary-KD and Binary-KD but omits Secondary-KD in the ultimate KD loss calculation.

Method	LFW	CFP-FP	CPLFW	AgeDB	CALFW
Student	99.52	91.66	87.93	95.82	95.12
KD ($\tau = 1$)	99.50	91.71	87.85	95.93	95.03
$\tau = 0.99$	99.46	92.15	87.98	96.16	95.41
$\tau = 0.97$	99.53	93.81	90.65	96.90	95.53
$\tau = 0.95$	99.55	94.25	90.43	96.88	95.60
$\tau = \mathbf{0.93}$	**99.61**	**94.35**	**90.86**	**97.25**	**95.78**
$\tau = 0.91$	99.52	94.00	89.95	96.93	95.65

Table 2: Extensive ablation on cumulative threshold τ.

Effects of Threshold τ. As described in Equation 2, we partition the original logits into two groups, i.e., primary group and secondary group, via a cumulative probability threshold τ. To explore the effects of τ, we conduct comparisons for different thresholds. Generally, a larger threshold indicates more knowledge for a student to fit and refers to KD loss when $\tau = 1$. Moreover, we can see from Ta-

ble 2 that the model trained with $\tau = 0.99$ has a similar performance in comparison to the classical KD loss. As τ decreases, less knowledge is required to be fitted for the student network, corresponding to achievable distillation tasks and better performance. Additionally, the best recognition performance comes when τ equals 0.93, and we set $\tau = 0.93$ as the default value of τ in our subsequent ablation and comparison experiments.

Method	LFW	CFP-FP	CPLFW	AgeDB	CALFW
Student	99.52	91.66	87.93	95.82	95.12
$\lambda_1 = p_\Phi^T$	99.58	94.25	89.85	97.21	95.65
$\lambda_1 = 2$	99.46	94.07	90.63	96.80	95.60
$\lambda_1 = 4$	99.52	94.23	90.75	96.95	95.50
$\lambda_1 = 8$	**99.61**	**94.35**	**90.86**	**97.25**	**95.78**
$\lambda_1 = 10$	99.48	94.03	90.58	97.22	95.66
$\lambda_2 = 0$	99.51	94.08	90.56	96.73	95.45
$\lambda_2 = 1$	99.61	**94.35**	**90.86**	97.25	**95.78**
$\lambda_2 = 2$	**99.64**	94.23	90.68	**97.28**	95.69

Table 3: Ablation on different λ_1 and λ_2, which correspond to weights for Primary-KD and Secondary-KD, respectively.

Effects of λ_1 and λ_2. The table 3 reports the recognition accuracy of student networks with different λ_1 and λ_2. First, we can find the default value of $\lambda_1 = \mathbf{p}_\Phi^T$ can bring reasonable performance improvement. Then, we conduct comparisons among different values of it and find the model achieves better performance when λ_1 increases. We analyze that the default weight $\mathbf{p}_\Phi^T$ cannot highlight the significance of Primary-KD, which can be enhanced by a larger weight. Additionally, we find that different weights do not introduce significant performance improvement since we discard Secondary-KD, which disguisedly enhances Primary-KD learning. The best performance is obtained when $\lambda_1 = 8$. Moreover, based on the setting ($\lambda_1 = 8$), we perform the ablation on different values of λ_2. From Table 3, we find that different values of λ_2 have a minor effect on the final result, as long as Binary-KD is introduced. The best performance achieves when $\lambda_2 = 1.0$.

Method	IJB-C	
	1e-4	1e-5
MobileFaceNet	89.13	81.65
MobileFaceNet+EKD	90.48	84.00
MobileFaceNet+Ours	**94.34**	**91.01**
IR18	91.96	86.01
IR18+EKD	92.74	88.84
IR18+Ours	**94.93**	**92.29**

Table 4: Ablation studies of student networks, which involves IResNet-18 and MobileFacenet. TPR@FPR=1e-4 and TPR@FPR=1e-5 on IJB-C are reported.

Effects of Student Network Architecture. We investigate the generalization capability of our method for different student network structures. Table 4 shows the results of two structures, i.e., IResNet-18 and MobileFaceNet. Although the performance improvement on the two network structures is different, our method generally performs better than directly training the student network from scratch. Additionally, we keep the consistent student network (IResNet-18 and MobileFacenet) and conduct a comparison with EKD (Huang et al. 2022). The results of Table 4 show that our method improves the student model.

Comparison with SOTA

In this section, we compare our proposed GKD with several Sate-Of-The-Art (SOTA) knowledge distillation methods, including the methods proposed for other tasks (KD (Hinton, Vinyals, and Dean 2015), FitNet (Romero et al. 2015), DarkRank (Chen, Wang, and Zhang 2018), SP (Tung and Mori 2019), CCKD (Park et al. 2019) and RKD (Park et al. 2019)), and methods designed for face recognition (ShrinkTeaNet (Duong et al. 2019), Triplet Distillation (Feng et al. 2020), MarginDistillation (Svitov and Alyamkin 2020) and EKD (Huang et al. 2022)). The results of six methods designed for other tasks are reproduced and reported on face recognition benchmarks. For the methods designed for face recognition, we replicate the results from EKD.

Results on LFW, CFP-FP, CPLFW, AgeDB, and CALFW. To evaluate the superiority of our proposed GKD, we first evaluate the recognition performance on five widely-used face benchmarks to compare with other SOTA methods. As shown in Table 5, most of the knowledge distillation methods are better than the student network that is directly trained from scratch (i.e., MobileFaceNet), but the performance improvement is limited. Additionally, feature-based distillation, e.g., FitNet and RKD, seem to show better performance than the logits-based methods (KD), in comparison with all the competitors, while performing inferior to MarginDistillation. By contrast, the SOTA method EKD brings significant improvement in comparison to other methods. Compared with the feature-based EKD, our method follow the logits distillation and achieves explicit improvements on five face benchmarks, which further bridges the performance gap by alleviating the difficulty of distillation.

Results on IJB-B and IJB-C. In Table 5, we extensively conduct the comparisons of the 1:1 verification (TPR@FPR=1e-4 and TPR@FPR=1e-5) with the previous SOTA methods on the IJB-B and IJB-C. Surprisingly, the results of our proposed method achieve a significant verification performance improvement, which is different from the results in the small test datasets. Additionally, most of the knowledge distillation methods bring little performance improvement or are even worse than the baseline on these two large-scale test datasets. Compared with previous methods, both RKD and EKD bring supervising improvements, but our method further improves 3.86% and 7.01% verification performance on IJB-C and brings 4.17% and 9.46% improvements on IJB-B (TPR@FPR=1e-4 and TPR@FPR=1e-5), which indicates the effectiveness of our method in the large-scale verification scenarios.

Method	IJB-C	IJB-B	MegaFace	LFW	CFP-FP	CPLFW	AgeDB	CALFW
	1e-4/1e-5	1e-4/1e-5	Id/Ver	ACC	ACC	ACC	ACC	ACC
Teacher-IR50 (upper bound)	95.16/92.66	93.45/88.65	98.14/98.34	99.80	97.63	92.50	97.92	96.05
MobileFaceNet (student)	89.13/81.65	87.07/74.63	90.91/92.71	99.52	91.66	87.93	95.82	95.12
FitNet (Romero et al. 2015)	87.76/73.71	86.35/70.19	91.16/92.34	99.47	91.30	88.30	96.18	95.12
KD (Hinton, Vinyals, and Dean 2015)	88.37/80.39	86.08/74.30	90.40/92.00	99.50	91.71	87.85	95.93	95.03
Dark (Chen, Wang, and Zhang 2018)	89.28/81.62	86.76/73.75	90.76/92.41	99.55	91.84	87.77	95.60	95.07
SP (Tung and Mori 2019)	88.43/78.13	86.34/72.85	91.25/92.41	99.53	92.33	88.45	96.17	95.07
CCKD (Park et al. 2019)	87.99/78.75	85.63/72.38	91.17/92.76	99.47	91.90	88.48	95.83	95.22
RKD (Park et al. 2019)	89.65/83.21	87.27/75.17	91.44/92.92	99.58	92.13	87.97	96.18	95.25
ShrinkTeaNet (Duong et al. 2019)	87.80/79.78	85.31/75.23	90.73/92.32	99.47	91.97	88.52	96.00	94.98
Triplet (Feng et al. 2020)	84.57/76.65	81.88/70.51	86.52/88.75	99.55	93.14	88.03	95.33	94.97
Margin (Svitov and Alyamkin 2020)	85.71/75.00	82.97/66.25	91.70/92.96	99.61	92.01	88.03	96.55	95.13
EKD (Huang et al. 2022)	90.48/84.00	88.35/76.60	91.02/93.08	99.60	94.33	89.35	96.48	95.37
Ours	**94.34/91.01**	**92.52/86.06**	**95.48/95.87**	**99.61**	**94.35**	**90.86**	**97.25**	**95.78**

Table 5: Verification comparison with SOTA methods on LFW, two pose benchmarks: CFP-FP and CPLFW, two age benchmarks: AgeDB and CALFW, and large-scale benchmarks: IJB-B, IJB-C and MegaFace.

Results on MegaFace. We test our model on MegaFace Challenge 1 using FaceScrub as the probe set. "Id" refers to Rank-1, and Ver refers to TPR@FPR=1e-6. As shown in Table 5, the proposed GKD outperforms other methods on MegaFace.

Method	Student	FitNet	KD	EKD	GKD
Training Time(s)	0.088	0.120	0.116	0.133	0.121

Table 6: Training time comparison for each batch under the same experimental setting.

Time Complexity. As shown in Table 6, we assess the training time of students including 1) student (w/o distillation loss), 2) KD, 3) FitNet, 4) EKD, and 5) our method for each batch under the same experimental setting. Specifically, we conduct 4000 complete iterations including forward and backward propagation on one NVIDIA Tesla-V100, and calculate the average as the training time. The batch size is set to 8 and the Pytorch version is 1.7.1.

Method	CALFW	CPLFW	AgeDB
	M-M/M-N	M-M/M-N	M-M/M-N
Teacher (IR100)	95.80/95.58	91.68/92.51	96.30/97.00
MobileNetV3-L	92.65/93.33	85.91/87.23	90.30/93.46
MobileNetV3-L+Ours	**94.41/94.66**	**89.46/90.48**	**94.50/95.83**

Table 7: Ablation studies on Masked Face Recognition challenge with different student-teacher network settings, which involves in MobileNetV3-large300 (Student) and IResNet-100 (Teacher). The testing scenarios M-M and M-N indicate mask vs. mask and mask vs. non-mask, respectively.

Generalization on Masked Face Recognition

In this section, we study the generalization of our method to other recognition tasks, e.g., Masked Face Recognition. Knowledge distillation is a general method to bridge the performance gap between teacher and student networks, and we think its effectiveness should be verified on other tasks, e.g., Masked Face Recognition (MFR). We conduct ablation experiments on MFR with different network settings, i.e., MobileNetV3-large (Student) (Howard et al. 2019) and IResNet-100 (Teacher) (Deng et al. 2019a). For the training dataset, we adopt the FaceX-Zoo (Wang et al. 2021) to generate the masked data from MS1M (Deng et al. 2019a), which consists of approximately 10M images of 9.3K identities. We utilize CosFace (Wang et al. 2018) loss with margin $m = 0.4$ and scale $s = 64$ as the loss function. There are two scenarios for MFR, i.e., Mask vs. Mask (M-M), and Mask vs. Non-mask (M-N). For the testing sets and evaluation protocol, we keep consistent with MaskInv (Huber et al. 2021). From Table 7, we can see that our method outperforms the baseline method, which demonstrates the generalization of our model on masked face recognition.

Conclusion

This paper proposes Grouped Knowledge Distillation (GKD) to probe the target logits to extract the primary knowledge that is related to face identity, and discard the others, to make the distillation more achievable for the student network. Specifically, there is a tail group that has near-zero values in the prediction, including minor knowledge of distillation. Therefore, we first partition the logits into two groups, i.e., primary group and secondary group, via the cumulative probability of the softened prediction. Then, we reorganize the distillation loss into three parts, i.e., Primary-KD, Secondary-KD, and Binary-KD. We experimentally found that Primary-KD and Binary-KD are indispensable for KD, and Secondary KD is the culprit restricting KD at the bottleneck. Extensive experimental results on popular face recognition benchmarks demonstrate the superiority of proposed GKD over state-of-the-art methods. Moreover, the experiments conducted on masked face recognition tasks demonstrate the generalization of our method as well.

Acknowledgments

This work was supported in part by the National Key Research & Development Program (No. 2020YFC2003901), Chinese National Natural Science Foundation Projects (61876178, 62276254, 62176256, 62106264, 62206280, 61871378, U2003111), the Youth Innovation Promotion Association CAS (#Y2021131), Defense Industrial Technology Development Program (Grant JCKY2021906A001) and the InnoHK program.

References

Chen, S.; Liu, Y.; Gao, X.; and Han, Z. 2018. MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices. In Zhou, J.; Wang, Y.; Sun, Z.; Jia, Z.; Feng, J.; Shan, S.; Ubul, K.; and Guo, Z., eds., *Biometric Recognition - 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings*, volume 10996 of *Lecture Notes in Computer Science*, 428–438. Springer.

Chen, Y.; Wang, N.; and Zhang, Z. 2018. DarkRank: Accelerating Deep Metric Learning via Cross Sample Similarities Transfer. In McIlraith, S. A.; and Weinberger, K. Q., eds., *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 2852–2859. AAAI Press.

Chopra, S.; Hadsell, R.; and LeCun, Y. 2005. Learning a Similarity Metric Discriminatively, with Application to Face Verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, 539–546. IEEE Computer Society.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019a. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 4690–4699. Computer Vision Foundation / IEEE.

Deng, J.; Guo, J.; Zhang, D.; Deng, Y.; Lu, X.; and Shi, S. 2019b. Lightweight Face Recognition Challenge. In *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, 2638–2646. IEEE.

Duong, C. N.; Luu, K.; Quach, K. G.; and Le, N. 2019. ShrinkTeaNet: Million-scale Lightweight Face Recognition via Shrinking Teacher-Student Networks. *CoRR*, abs/1905.10620.

Feng, Y.; Wang, H.; Hu, H. R.; Yu, L.; Wang, W.; and Wang, S. 2020. Triplet Distillation For Deep Face Recognition. In *IEEE International Conference on Image Processing, ICIP 2020, Abu Dhabi, United Arab Emirates, October 25-28, 2020*, 808–812. IEEE.

Hinton, G. E.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *CoRR*, abs/1503.02531.

Hoffer, E.; and Ailon, N. 2015. Deep metric learning using Triplet network. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.

Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. 2019. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1314–1324.

Huang, G. B.; Mattar, M.; Berg, T.; and Learned-Miller, E. 2008. 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*.

Huang, Y.; Wang, Y.; Tai, Y.; Liu, X.; Shen, P.; Li, S.; Li, J.; and Huang, F. 2020. CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 5900–5909. Computer Vision Foundation / IEEE.

Huang, Y.; Wu, J.; Xu, X.; and Ding, S. 2022. Evaluation-oriented Knowledge Distillation for Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18740–18749.

Huber, M.; Boutros, F.; Kirchbuchner, F.; and Damer, N. 2021. Mask-invariant Face Recognition through Template-level Knowledge Distillation. In *16th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2021, Jodhpur, India, December 15-18, 2021*, 1–8. IEEE.

Kemelmacher-Shlizerman, I.; Seitz, S. M.; Miller, D.; and Brossard, E. 2016. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4873–4882.

Lei, Z.; Pietikäinen, M.; and Li, S. Z. 2014. Learning Discriminant Face Descriptor. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(2): 289–302.

Li, J.; Tan, Z.; Wan, J.; Lei, Z.; and Guo, G. 2022. Nested Collaborative Learning for Long-Tailed Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6949–6958.

Li, S. Z.; and Jain, A. K., eds. 2011. *Handbook of Face Recognition, 2nd Edition*. Springer. ISBN 978-0-85729-931-4.

Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; and Song, L. 2017. SphereFace: Deep Hypersphere Embedding for Face Recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 6738–6746. IEEE Computer Society.

Maze, B.; Adams, J. C.; Duncan, J. A.; Kalka, N. D.; Miller, T.; Otto, C.; Jain, A. K.; Niggel, W. T.; Anderson, J.; Cheney, J.; and Grother, P. 2018. IARPA Janus Benchmark - C: Face Dataset and Protocol. In *2018 International Conference on Biometrics, ICB 2018, Gold Coast, Australia, February 20-23, 2018*, 158–165. IEEE.

Mirzadeh, S.; Farajtabar, M.; Li, A.; Levine, N.; Matsukawa, A.; and Ghasemzadeh, H. 2020. Improved Knowledge Distillation via Teacher Assistant. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 5191–5198. AAAI Press.

Moschoglou, S.; Papaioannou, A.; Sagonas, C.; Deng, J.; Kotsia, I.; and Zafeiriou, S. 2017. AgeDB: The First Manually Collected, In-the-Wild Age Database. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, 1997–2005. IEEE Computer Society.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational Knowledge Distillation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 3967–3976. Computer Vision Foundation / IEEE.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2015. FitNets: Hints for Thin Deep Nets. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 815–823. IEEE Computer Society.

Sengupta, S.; Chen, J.; Castillo, C. D.; Patel, V. M.; Chellappa, R.; and Jacobs, D. W. 2016. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Lake Placid, NY, USA, March 7-10, 2016*, 1–9. IEEE Computer Society.

Sun, Y.; Chen, Y.; Wang, X.; and Tang, X. 2014. Deep Learning Face Representation by Joint Identification-Verification. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 1988–1996.

Svitov, D.; and Alyamkin, S. 2020. MarginDistillation: distillation for margin-based softmax. *CoRR*, abs/2003.02586.

Tung, F.; and Mori, G. 2019. Similarity-Preserving Knowledge Distillation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 1365–1374. IEEE.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018. CosFace: Large Margin Cosine Loss for Deep Face Recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 5265–5274. Computer Vision Foundation / IEEE Computer Society.

Wang, J.; Liu, Y.; Hu, Y.; Shi, H.; and Mei, T. 2021. FaceX-Zoo: A PyTorch Toolbox for Face Recognition. In Shen, H. T.; Zhuang, Y.; Smith, J. R.; Yang, Y.; Cesar, P.; Metze, F.; and Prabhakaran, B., eds., *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, 3779–3782. ACM.

Whitelam, C.; Taborsky, E.; Blanton, A.; Maze, B.; Adams, J. C.; Miller, T.; Kalka, N. D.; Jain, A. K.; Duncan, J. A.; Allen, K.; Cheney, J.; and Grother, P. 2017. IARPA Janus Benchmark-B Face Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, 592–600. IEEE Computer Society.

Zhang, Y.; Xiang, T.; Hospedales, T. M.; and Lu, H. 2018. Deep Mutual Learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 4320–4328. Computer Vision Foundation / IEEE Computer Society.

Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; and Liang, J. 2022. Decoupled Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11953–11962.

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

Zheng, T.; Deng, W.; and Hu, J. 2017. Cross-Age LFW: A Database for Studying Cross-Age Face Recognition in Unconstrained Environments. *CoRR*, abs/1708.08197.