QUD: Unsupervised Knowledge Distillation for Deep Face Recognition

Jan Niklas Kolf[1,2]
jan.niklas.kolf@igd.fraunhofer.de
Naser Damer[1,2]
naser.damer@igd.fraunhofer.de
Fadi Boutros[1]
fadi.boutros@igd.fraunhofer.de

[1] Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany

[2] Department of Computer Science, TU Darmstadt, Darmstadt, Germany

Abstract

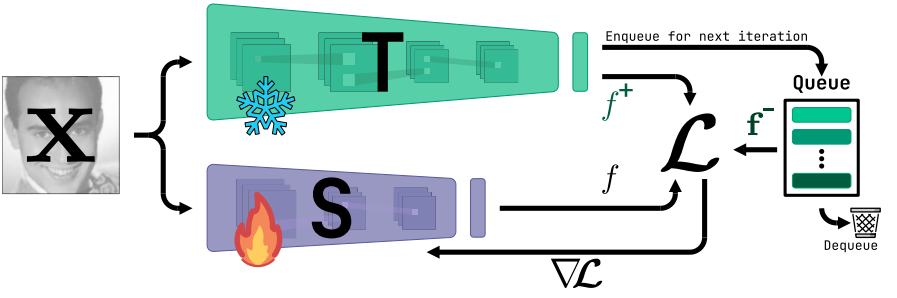
We present in this paper an unsupervised knowledge distillation (KD) approach, namely QUD, for face recognition. The proposed QUD approach utilizes a queue of features within a contrastive learning setup to guide the student model to learn a feature representation similar to its counterpart obtained from the teacher and dissimilar from the ones that are stored in a queue. This queue is updated by pushing a batch of feature representations obtained from the teacher into the queue and dequeuing the oldest ones from the queue in each training iteration. We additionally incorporate a temperature into the contrastive loss to control how sensitive contrastive learning is to samples considered negative in the queue. The proposed unsupervised QUD approach does not require accessing the same dataset used to train the teacher model or even for the data to have identity labels. The effectiveness of the proposed approach is demonstrated through several sensitivity studies on different teacher architectures and using different datasets for student training in the KD framework. Additionally, the achieved results on mainstream benchmarks by our unsupervised QUD are compared to state-of-the-art (SOTA), achieving very competitive performances and even outperforming SOTA on several benchmarks. Code and pre-trained models are available under https://github.com/jankolf/QUD.

1 Introduction

Face recognition (FR) is a well-established technology in our daily lives and is often integrated on devices with limited computational capacities, e.g. mobile phones [30, 31, 44]. However, many state-of-the-art (SOTA) FR models still rely on computationally demanding deep neural networks (DNN) [3, 49]. To maintain the performance of such models while reducing computational costs for implementation on edge devices, several techniques are adopted, such as quantization [4, 33], pruning [35] or neural architecture search [5].

A promising approach to bridge the performance gap between lightweight and highly performing models, but computationally demanding, is to utilize knowledge distillation (KD) [5, 24, 34]. In this approach introduced by Bucila *et al.* [10] and popularized by Hinton

Figure 1: Overview of the proposed unsupervised KD method QUD. Using contrastive loss, student **S** is trained so that the distance of its feature $f = \mathbf{S}(\mathbf{x})$ to teacher **T**'s positive feature $f^+ = \mathbf{T}(\mathbf{x})$ of the same input sample $\mathbf{x}$ is smaller than the distance between f and a set of negative features $f^- \in \mathbf{f}^-$ stored in a queue. The queue is filled with features of **T** from previous iterations. After each iteration, the current features of **T** are enqueued and an equal amount of the oldest features are dequeued from the queue. Only **S**'s parameters are updated.



et al. [24], a high-performing FR model with relatively high computational cost is used as a "teacher" that trains, through a KD process, a relatively compact and small "student" FR model [24, 27, 34]. KD enables the more compact student model to achieve higher recognition results than when it is trained without the support of the teacher model [24, 26]. Early works of KD [24] utilized the class logits to distil knowledge from the teacher to the student. In FR, as a feature extraction process, class logits distillation is not optimal, as discriminative features must be learned, not only the respective class [3, 17, 49]. In addition, it requires the same training data as the teacher for KD. Therefore, and beyond learning the class logits, recent works also focused on pushing the student to learn the exact features of the teacher [9, 20, 34, 45] or the relational properties between features of the teacher [27, 39, 41]. The first does not provide the tolerance for the student to adapt its feature space [34, 41] and the second typically requires labeled data [27, 39].

Toward overcoming these limitations in previous work we propose a novel unsupervised KD approach, namely QUD. In this context, unsupervised KD refers to the fact that there is no supervision through class labels and there is only self-supervision through the teacher model. Our approach leverages the properties of contrastive learning (CL) [12, 50], commonly used for unsupervised representation learning [23]. QUD uses the student feature output as anchor samples, the teacher feature output as positive samples that form positive pairs with that of the student, and the set of negative samples is the feature output of the teacher that is queued up from previous iterations, forming negative pairs with the student output. The QUD process then updates the student so that the distance between positive pairs is smaller than the distance between negative pairs. An overview of the QUD concept is visualised in Figure 1. We include a temperature into the contrastive loss to control the sensitivity of samples to the queue negative samples. The student therefore is not as strictly forced to learn the exact feature representations learned by the teacher. This all can be achieved using any pre-trained teacher model and does not require any class labels, and thus can be used with unlabeled data, even simply using generated privacy-friendly synthetic data. In detailed experiments, we analyse our design choices, prove the generalizability of QUD over different teacher models and distillation datasets, make a direct comparison to the baseline feature-based knowledge distillation, and compare to the latest SOTA works in FR KD over a wide range of evaluation benchmarks.

Table 1: Conceptual comparison on the design choices of knowledge distillation (KD) approaches using teacher **T** and student model **S** between our QUD and state-of-the-art KD methods in literature. While ReFo [34] (marked with *) does not require class labels during KD, their approach uses labels during proxy-student training.

Property	Method													
	FitNet [42]	KD [24]	DarkRank [3]	SP [43]	CCKD [41]	RKD [39]	ShrinkTeaNet [19]	Triplet-Distillation [20]	MarginKD [45]	SFTN [38]	EKD [27]	SH-KD [2]	ReFO [34]	QUD (Ours)
T and **S** require same training data	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗
Requires class labels	✓	✗	✗	✓	✓	✗	✓	✓	✓	✓	✓	✓	✗*	✗
S is required to exactly learn **T**'s feature space	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗
Limited to samples in batch	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗

2 Related Work

Seminal works can be grouped into three different groups, based on their KD approach: Response-based, feature-based, and relation-based KD.

Response-Based KD: These approaches [2, 24, 38], try to match the class output probabilities, i.e. logits, of the teacher with that of the student model, e.g. through Kullback-Leibler divergence [24].

Feature-Based KD: These approaches utilize intermediate feature representations of the models to train students. ShrinkTeaNet [19] extracts intermediate representations of the student model and passes them through the remaining layers in the teacher. The resulting teacher-student feature and the teacher feature are compared by feature direction. FitNet [42] utilized intermediate features from the student and teacher model for distillation, adding an additional network to match both feature dimensions. Other approaches adapted FR loss functions for KD [20, 45]. TripletDistillation [20] adopted triplet loss used in FR together with a dynamic margin. MarginDistillation [45] used a teacher model with fixed class centers and trained a student model using margin-penalty softmax loss and fixed class centers as targets. ReFO [34] suggests bridging the intrinsic gap between a teacher and student model by first training a compact model. This compact model acts as a guide in a reverse distillation process with a teacher model. This teacher model is then used to train a compact student model using feature-level KD. AdaDistill [9] utilizes an adaptive distillation procedure in which the student model is trained using softmax loss with class labels that are distilled from the teacher model used.

Relation-Based KD: This group of approaches utilizes relations between layers and features of teacher and student models or between individual samples. DarkRank [13] introduced cross-sample similarities and a ranking-based KD approach based on feature similarity in the current batch. RKD [39] presented a KD approach that transfers structural relations of samples in a batch from the teacher to the student using distance-wise or angle-wise losses. CCKD [41] proposed a kernel-based method that improves the capturing of correlations between instances in a given batch and additionally proposed sampling strategies for batches. EKD [27] presented a rank-based approach that selects relations that minimize the difference in FR performance metrics between teacher and student.

Towards QUD: Our method introduced in this work, QUD, utilizes relations between the feature output from the student, the teacher output from the same input (positive pair) and the teacher output from previous batches (negative pairs) to learn relations between features. An overview of the comparison of QUD with other KD methods is given in Table 1.

Unlike FitNet [42], KD [24], MarginKD [45], SFTN [38], and SH-KD [2], our KD approach does not require the same training data for KD that was used for training teacher

T. Other than FitNet [42], SP [46], CCKD [41], ShrinkTeaNet [19], Triplet-Distillation [20], MarginKD [45], SFTN [38], EKD [27], and SH-KD [2], our approach does not require any labeled datasets (ReFO [34] does not require class labels during KD, but during proxy-student training). Other than FitNet [42], MarginKD [45], and ReFO [34], the student **S** in our proposed QUD is not required to learn **T**'s feature space. Unlike any other KD approach, QUD is not limited to samples in the current batch to extract relations or information for KD.

3 Methodology

This section first presents preliminaries on conventional KD followed by the concept of unsupervised contrastive learning. Finally, it introduces our QUD approach, which enables the transfer of knowledge from the teacher to its student without the need to access the class labels of the KD training dataset. The student in our QUD is trained so that the distance between its feature representation of the input sample $\mathbf{x}$ and the counterpart obtained from the teacher is smaller than the distance between its feature representation and the ones retrieved from the queue. The queue contains a set of feature representations obtained from the teacher and is updated after each training iteration by enqueuing the current batch and dequeuing the oldest feature representation. An overview of our QUD framework is presented in Figure 1.

3.1 Preliminary: Knowledge Distillation

To distill knowledge from a teacher model **T** to a student model **S** using dataset $\mathcal{D}$, a general loss [34] for face knowledge distillation uses three losses with weight terms α, β and λ and it can be defined as:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{KD}} + \beta\mathcal{L}_{\text{feat}} + \lambda\mathcal{L}_{\Lambda}. \tag{1}$$

The loss $\mathcal{L}_{\text{KD}}$ (response-based KD) is used to drive **S** to match the classification output distribution (e.g. softmax layer) of **T**, using e.g. Kullback-Leibler divergence [24]. It requires access to the classification layer of both pre-trained **T** and **S** and it is mostly used in conjunction with the same dataset used to train **T** [24]. This imposes an additional requirement, namely having access to the dataset used to train the teacher. Furthermore, in the case of FR, matching classification output directly is suboptimal [49] for face KD, as the aim of **S** and **T** is to learn discriminative feature representations for face verification rather than learning accurate sample classification [8, 17, 49]. Thus, recent SOTA KD approaches for FR proposed learning to match the feature representations between **S** and **T** [27, 34, 41, 49].

The loss $\mathcal{L}_{\text{feat}}$ (feature-based KD) is used for feature distillation [34], where for each sample $\mathbf{x} \in \mathcal{D}$, the student is trained to minimize the distance between the teacher feature $f^t = \mathbf{T}(\mathbf{x})$ and its feature $f^s = \mathbf{S}(\mathbf{x})$ [34, 41]. This ensures that **S** mimics the feature space of **T** [27, 34]. Feature distillation approaches [34] do not require access to the **T** training dataset or the classification layer of **T**. This, along with direct learning of the features, addresses the mentioned shortcomings of response-based KD for FR. However, one remaining issue in feature-based KD is that **S** has a significantly reduced capacity compared to **T** [14, 26, 34, 48]. The feature space of **T** might be too complex for **S** [26, 48]. For this reason, several KD approaches have proposed either various techniques for feature distillation [20, 34, 45], e.g. multiple training iterations or additional margins, or specially designed losses that aim to overcome the disadvantages and limitations mentioned before, e.g. $\mathcal{L}_{\Lambda}$ [27, 41].

The $\mathcal{L}_{\Lambda}$ loss specifies the main task-specific loss, e.g., classification. In FR, this is mostly achieved using margin-penalty softmax losses [45] or triplet loss [20].

Current face KD approaches [2, 20, 27, 45] added task-specific loss functions that require partially class labels and consider only samples within the batch for face KD. Therefore, in this paper, we propose a novel unsupervised KD approach based on contrastive learning. This proposed solution does not require identity labels and inherently considers the relationships to samples outside of the current batch. Our novel approach named QUD uses only $\mathcal{L}_\Lambda$ with $\alpha = 0$ and $\beta = 0$. In the following subsections, a short introduction to unsupervised contrastive learning is given, followed by a detailed explanation of our QUD.

3.2 Unsupervised Contrastive Learning

Unsupervised contrastive learning (UCL) is a learning approach that aims to optimize a feature space where the features of the same class (positive pairs) are close to each other while the features of different classes (negative pairs) are far from each other[15, 50].

When considering UCL, e.g. for FR [7], positive pairs can be generated by heavily augmenting input images, and negative pairs are created by sampling different images from the dataset, e.g. using the current input batch [7, 23, 50]. As a large number of negative pairs is crucial for learning discriminative feature representations[52], recent methods [23, 54] utilize a queue [12, 54] of n negative samples. The queue size n determines the number of samples that are considered negative. At the end of each iteration, the current samples in the batch are enqueued into the queue and an equal number of samples are dequeued from it.

Given a feature f of sample $\mathbf{x}$, a positive feature f^+, n negative features f^- in a queue $\mathbf{f}^-$, $f^- \in \mathbf{f}^-$, dot product as similarity metric, and temperature parameter τ [54], the contrastive loss InfoNCE [47] is defined as [23, 47]

$$\mathcal{L}_{\mathrm{CL}} = -\log\left(\frac{\exp(f \cdot f^+/\tau)}{\exp(f \cdot f^+/\tau) + \sum_{f^- \in \mathbf{f}^-} \exp(f \cdot f^-/\tau)}\right). \tag{2}$$

The parameter τ scales the loss function and influences its sensitivity [7, 57] to the samples from the queue that are considered negative. A lower τ value emphasizes differences between the features, which leads to a higher loss and stronger gradients [57]. On the other hand, a higher τ value tends to be more tolerant to differences between the features [12, 50, 57]. Tuning τ is crucial for the UCL and must be adjusted for the respective task [12, 57]. The size n of the queue $\mathbf{f}^-$ is an additional hyperparameter, as its size influences the number of features that the sample is compared to [23]. The impact on the performance of both parameters is shown later in this work.

3.3 Deriving QUD Approach

In the previous subsections, we have introduced the concepts of KD in the context of FR. KD shows great promise in FR, although it typically relies on labeled data [27, 41]. In contrast, UCL offers the advantage of operating on datasets without class labels. Building upon these insights, we propose a novel unsupervised KD method, called QUD. In this context, unsupervised KD refers to the fact that there is no supervision through class labels, and there is only self-supervision through the teacher model.

To build positive and n-negative pairs required by queue-based UCL, QUD utilizes pretrained teacher model $\mathbf{T}$ with frozen weights. For any input sample $\mathbf{x}$ in the training dataset $\mathcal{D}$, student feature $f = f^s = \mathbf{S}(\mathbf{x})$ and teacher feature $f^+ = f^t = \mathbf{T}(\mathbf{x})$ form the positive pair. The queue of negative features $\mathbf{f}^-$ is incrementally built during training and contains

n features $f^- = f^t = \mathbf{T}(\mathbf{x}')$ from the previous batches that occurred during training. In the first training iteration, the dictionary is empty. Initially, the queue is randomly initialized, and it is updated over the training iterations by pushing the teacher features resulting from the current batch into the queue and dequeuing the oldest ones.

The loss used is $\mathcal{L}_\Lambda = \mathcal{L}_{\text{CL}}$ (Equation 2), the training object for **S** is to predict feature f so that the distance between f and teacher feature f^+, predicted by **T** is smaller than the distance between f and the n features predicted by **T** in previous iterations. The temperature τ relaxes the need of **S** to learn the exact feature space of **T**.

Following this, our approach does not require that **T** and **S** share the same training data, it does not require class labels, **S** is not required to exactly learn **T**'s feature space, and is the first method to consider n-pair relations and not only samples from the current batch.

4 Experimental Setup

This section presents the experimental setups followed in this work.

4.1 Datasets

Train Sets: For fair comparison with previous KD approaches, we utilize MS1MV2 [17, 21] to train our student networks. It contains 5.8M images from 85k different identities. MS1MV2 is based on MS-Celeb-1M [21] which was refined by [17]. To show the success of our proposed method on unlabeled data, we use StyleGAN2 [28] and generate 2M unlabeled images for student training, following [28]. In this work, this dataset is referred to as StyleGAN2-2M. Additionally, the synthetic dataset IDiff-Face [6] is used. It contains 10k identities with 500k images that are generated by a conditional latent diffusion model. For ablation studies, CASIA-Webface [65] is used. It contains 494,414 images of 10,575 identities that were collected from the web. During student training, no identity labels are used in any of the experiments.

Test Sets: Following common benchmarks, model performance is reported on a set of diverse benchmarks: Labeled Faces in the Wild (LFW) [25], AgeDB-30 [37], Celebrities in Frontal-Profile in the Wild (CFP-FP) [43], Cross-Age LFW (CA-LFW) [59], Cross-Pose LFW (CP-LFW) [58], ICCV21-MFR [18], IARPA Janus Benchmark–B (IJB-B) [53] and Benchmark–C (IJB-C) [36], MegaFace [29], and refined MegaFace dataset (MegaFace (R)) [17, 29].

Data Preprocessing: All face images are aligned and cropped to 112×112 pixels and normalized to [-1,1], following [17, 27, 34]. Alignment is performed with five landmarks that are extracted using Multi-Task Cascaded Convolutional Networks (MTCNN) [56] using the procedure described in [17].

4.2 Experimental Settings

Models: For comparison with other SOTA KD methods [20, 27, 34, 41], MobileFaceNet [11] (referred to as MFN, 1.19M parameters and 0.45 GFLOPs) is used as a student. We follow the common setup [27, 34, 45] and utilize a pre-trained ResNet-50 (43.59M parameters and 13.64 GFLOPs) [22] as teacher model. We also reported the results of our models using two additional pre-trained teacher models, ResNet-100 [22] (65.15M parameters and 24.21 GFLOPs) and transformer-based Transface-B (124.47M parameters and 21.92

GFLOPs) [16]. All teacher models are trained on MS1MV2 [17, 21] using ArcFace (margin $m = 0.5$, scale $s = 64$), following [17, 27, 34]. All models use a feature size of 512 and their weights are frozen during KD.

Ablation Study: For training of our QUD, the queue size and temperature parameters presented in Section 3 are experimentally selected. In Section 5 we present an in-depth sensitivity study on both hyperparameters using teacher ResNet-50 [22], student MFN [11] and CASIA-Webface [65] as training dataset. Additionally, to examine the influence of different training datasets not previously seen by the teacher (including synthetic ones), we investigate the use of CASIA-Webface [65], our generated StyleGAN2-2M [28], and IDiff-Face [6].

Training Setup: The code is implemented using Pytorch [40] and the models are trained on two NVIDIA A100 GPUs. Student models are trained using a Stochastic Gradient Descent optimizer, using a batch size of 512, weight decay of $5e^{-4}$, momentum of 0.9 and initial learning rate of 0.1, following [17, 27, 34]. Input images are augmented using horizontal flip with a probability of 0.5, following [17, 27, 34]. The number of training epochs are 26 for MS1MV2 [17, 21] and StyleGAN2-2M (following [3, 17]), and 40 for CASIA-Webface [65] and IDiff-Face [6] (following [6, 32]), respectively. The learning rate is divided by 10 in the 8th, 14th, 20th and 25th epochs for MS1MV2 [17, 21] and StyleGAN2-2M, and in the 22nd, 30th and 40th epochs for CASIA-Webface [65] and IDiff-Face [6], following similar setups in [3, 6, 17, 32].

Evaluation Metrics: We followed the evaluation protocols and metrics of each of the evaluation benchmarks. For LFW [25], AgeDB-30 [37], CFP-FP [43], CA-LFW [59], CP-LFW [58] the verification performance on their respective evaluation protocol is given in verification accuracy ([%]). For IJB-B [63] and IJB-C [36] the true acceptance rates (TAR) at false acceptance rates (FAR) of 10^{-4} and 10^{-5} are given for the 1:1 mixed verification protocol [36, 63]. Following the ICCV-MFR [18] protocols for the Mask, Children and Multi-Racial (MR-all) challenges, we report performance for Mask and Children with TAR at FAR$= 10^{-4}$ and for MR-all with TAR at FAR$= 10^{-6}$, following [1, 18, 34]. For MegaFace [29] and MegaFace(R) [17, 29] the rank-1 identification rate is reported using 1M distractors, the verification accuracy is reported as TAR at FAR$= 10^{-6}$, following [17, 27, 29, 34].

5 Results

This section presents an in-depth discussion of the experimental behaviour of the proposed solution. First, by studying the sensitivity and our choice of the parameters temperature τ and queue size n, then by analysing the applicability to variations in both teacher models and training datasets. All this is performed on the small-scale benchmarks LFW [25], CFP-FP [43], AgeDB-30 [37], CA-LFW [59], CP-LFW [58], and their average performance, as well as large-scale IJB-C [36] and performance is reported as defined in 4. Finally, a detailed comparison with the latest SOTA FR KD approaches is presented. In all experiments, MobileFaceNet (MFN) [11] is used as a student and all teacher models are trained using MS1MV2 [17, 21]. Results for experiments **I) - V)** are shown in Table 2 and the comparison to SOTA (experiment **VI)**) is shown in Table 3.

I) Effects of temperature: As discussed in Section 3, the temperature parameter τ specifies the scaling of the loss and therefore the accentuation of feature similarity and dissimilarity [57]. The results of this ablation study with fixed queue size $n = 1024$, fixed **S** (MFN [11]) and **T** (ResNet-50 [22]) architectures, fixed distillation data (CASIA-Webface [65]),

Table 2: Studies on parameter selection, KD method validation, teacher architectures and training datasets discussed in Section 5. The performance for each ablation study is reported on benchmarks LFW, CFP-FP, AgeDB-30, CA-LFW, CP-LFW (verification accuracy [%]), and on IJB-C (TAR at FAR of 10^{-4} and 10^{-5}. The baseline results of MFN trained without KD are presented in the second row of the Table. All teacher models are trained on MS1MV2.

Study	Method	Distillation Dataset	LFW	CFP-FP	AgeDB-30	CA-LFW	CP-LFW	Avg.	IJB-C 1e-4	IJB-C 1e-5
	ResNet-50 (Teacher)	-	99.80	97.63	97.92	96.10	92.43	96.77	96.05	93.96
	MFN (Student)	-	99.52	91.66	95.82	95.12	87.93	94.01	89.13	81.65
I) Temperature τ	MFN + QUD ($\tau = 0.06$, n= 1024)	CASIA-Webface [55]	99.17	93.26	94.12	93.87	89.30	93.94	88.69	75.75
	MFN + QUD ($\tau = 0.08$, n= 1024)		99.43	94.00	94.95	94.87	90.40	94.73	89.65	73.26
	MFN + QUD ($\tau = 0.1$, n= 1024)		99.55	94.14	95.67	94.78	90.50	94.93	**90.28**	**74.96**
	MFN + QUD ($\tau = 0.2$, n= 1024)		99.50	94.41	96.03	95.22	90.13	95.06	88.87	60.93
	MFN + QUD ($\tau = 0.4$, n= 1024)		99.52	94.20	95.77	94.93	89.63	94.81	89.03	65.72
	MFN + QUD ($\tau = 0.6$, n= 1024)		99.58	94.16	95.80	95.00	89.78	94.86	89.43	69.41
II) Queue Size n	MFN + QUD ($\tau = 0.1$, n= 512)	CASIA-Webface [55]	99.53	94.33	95.93	95.13	89.97	94.98	89.74	69.32
	MFN + QUD ($\tau = 0.1$, n= 1024)		99.55	94.14	95.67	94.78	90.50	94.93	**90.28**	**74.96**
	MFN + QUD ($\tau = 0.1$, n= 4096)		99.52	94.29	95.92	95.07	90.00	94.96	89.77	72.30
	MFN + QUD ($\tau = 0.1$, n= 8192)		99.53	94.36	95.60	95.03	90.45	94.99	90.10	71.01
	MFN + QUD ($\tau = 0.1$, n= 16384)		99.57	94.30	95.82	95.07	90.33	95.02	89.32	66.17
III) KD Method	ResNet-50 (Teacher)	-	99.80	97.63	97.92	96.10	92.43	96.77	96.05	93.96
	MFN + Feature-based KD	MS1MV2 [6, 55]	99.57	93.77	96.97	95.70	89.01	95.00	91.29	79.79
	MFN + QUD ($\tau = 0.1$, n= 1024)	MS1MV2 [6, 55]	99.58	93.89	96.73	95.65	90.47	95.26	**91.92**	**81.60**
IV) Teacher Architectures	ResNet-50 (Teacher)	-	99.80	97.63	97.92	96.10	92.43	96.77	96.05	93.96
	MFN + QUD ($\tau = 0.1$, n= 1024)	MS1MV2 [6, 55]	99.58	93.89	96.73	95.65	90.47	95.26	91.92	81.60
	ResNet-100 (Teacher)	-	99.83	98.40	98.33	96.13	93.22	97.19	96.39	94.58
	MFN + QUD ($\tau = 0.1$, n= 1024)	MS1MV2 [6, 55]	99.68	93.71	97.18	95.62	90.32	95.30	92.19	85.00
	TransFace-B (Teacher)	-	99.85	99.17	98.53	96.20	92.92	97.33	96.55	94.15
	MFN + QUD ($\tau = 0.1$, n= 1024)	MS1MV2 [6, 55]	99.58	93.36	96.65	95.72	90.15	95.09	92.80	87.45
V) Datasets	ResNet-50 (Teacher)	-	99.80	97.63	97.92	96.10	92.43	96.77	96.05	93.96
	MFN (Student)	-	99.52	91.66	95.82	95.12	87.93	94.01	89.13	81.65
	MFN + Feature-based KD	MS1MV2 [6, 55]	99.57	93.77	96.97	95.70	89.01	95.00	91.29	79.79
	MFN + QUD ($\tau = 0.1$, n= 1024)	MS1MV2 [6, 55]	99.58	93.89	96.73	95.65	90.47	95.26	**91.92**	**81.60**
	MFN (Student)	-	98.97	93.21	91.40	91.27	86.42	92.25	77.46	57.75
	MFN + Feature-based KD	CASIA-Webface [55]	99.45	93.90	95.82	94.95	89.48	94.72	**90.63**	**79.54**
	MFN + QUD ($\tau = 0.1$, n= 1024)	CASIA-Webface [55]	99.55	94.14	95.67	94.78	90.50	94.93	90.28	74.96
	MFN (Student)	-	97.05	79.16	81.88	89.37	78.10	85.11	22.19	3.43
	MFN + Feature-based KD	IDiff-Face [8]	99.30	88.79	91.93	93.53	85.40	91.79	60.01	17.31
	MFN + QUD ($\tau = 0.1$, n= 1024)	IDiff-Face [8]	99.33	89.57	92.50	93.73	85.65	92.16	**64.96**	**29.08**
	MFN + Feature-based KD	StyleGAN2-2M	99.32	88.30	92.67	93.82	84.52	91.73	22.56	3.02
	MFN + QUD ($\tau = 0.1$, n= 1024)	StyleGAN2-2M	99.42	90.21	93.88	94.13	86.13	92.76	**49.95**	**19.35**

and using different values for τ are shown in Table 2. Considering the average performance across the smaller benchmarks and the IJB-C [36] benchmark, $\tau = 0.1$ achieved the best overall performance, with values around it also resulting in relatively high performance that deteriorates comparably when moving away from $\tau = 0.1$.

II) Effects of queue size: The queue size determines the number of samples that are considered negative to which the current sample is compared. By increasing the number of comparisons, the student could learn more discriminative features compared to using smaller queue sizes. However, when training datasets contain a large number of samples per class, the queue has a higher probability of being filled with multiple samples from the same class. This, if it occurs more frequently, could lead to the student being trained to maximize the distance from features of the same class, which might impact the performance of the student. This is investigated in this ablation, where $\tau = 0.1$, training dataset (CASIA-WebFace [55]), **S** (MFN [11]) and **T** (ResNet-50 [22]) architectures are fixed, and queue size n is altered. The results are shown in Table 2, where all queue sizes achieve very close results on small-scale benchmarks. While on the large-scale benchmark IJB-C, $n = 1024$ achieves the best performance.

III) Effects of QUD: To investigate whether our proposed QUD is boosting the KD performance, we compare MFN [11] students trained using the baseline feature-based KD using mean-squared-error loss and trained with QUD ($\tau = 0.1$, $n = 1024$) both using teacher model

ResNet-50 [22] on MS1MV2 [17, 21]. As shown in Table 2, the proposed QUD achieved a higher average accuracy on small-scale benchmarks (95.26% average accuracy vs. 95.00%) and higher TAR at all FAR of IJB-C [36], especially at FAR=10^{-5}, in comparison to conventional feature-based KD.

IV) Different teacher models: The applicability of QUD using different teacher models, ResNet-50 [22], ResNet-100 [22], and TransFace-B [16], is investigated. As shown in Table 2, the performance of MFN [11] always increases when KD is performed using QUD ($\tau = 0.1$ and $n = 1024$), compared to MFN [11] trained without knowledge distillation (top of Table 2). In common small-scale benchmarks, ResNet-100 [22] achieves the best average verification accuracy, followed by ResNet-50 [22] and TransFace-B [16], respectively. On the large-scale benchmark IJB-C [36], TransFace-B [16] achieves the best TAR at the presented FAR, with ResNet-100 [22] placed second and ResNet-50 [22] at third place.

V) Effects of training datasets: We validate that the proposed method is generalizing across different distillation datasets, especially when the teacher training data is different from the distillation data. These results are presented in the last group of results in Table 2.

The MFN [11] trained with ArcFace [17] ($m = 0.5$, $s = 64$, without KD) on either CASIA-Webface [55], IDiff-Face [6], is compared to the MFN [11] after using QUD ($\tau = 0.1$ and $n = 1024$) with ResNet-50 [22] (trained on MS1MV2 [17, 21]) as a teacher. The distillation is performed on CASIA-Webface [55] and IDiff-Face [6]. In both cases, using QUD ($\tau = 0.1$ and $n = 1024$) instead of training from scratch using ArcFace, consistently enhances MFN performance.

The performance also improved in most cases when using QUD ($\tau = 0.1$ and $n = 1024$) compared to the baseline of MFN [11] trained using feature-based KD. Due to a significant amount of noisy labels in CASIA-Webface [51] the performance on large-scale IJB-C is slightly impacted when using QUD. On all other utilized datasets, QUD outperforms feature-based KD trained MFN on IJB-C.

The performance of MFN [11] using QUD ($\tau = 0.1$ and $n = 1024$) and ResNet-50 [22] teacher distilled on the label-free generated StyleGAN2-2M dataset is also shown (here, no model trained without KD on that data is possible as it is not labeled). In this case, QUD achieved higher average verification accuracy on the common benchmarks when compared to MFN [11] trained on CASIA-Webface [55] or IDiff-Face [6] using ArcFace [17]. On the large-scale benchmark IJB-C [36], StyleGAN2-2M trained MFN [11] with QUD ($\tau = 0.1$ and $n = 1024$) achieved higher TAR at all FAR compared to MFN [11] (ArcFace [17]) trained on the synthetic IDiff-Face [6] dataset.

VI) Comparison to SOTA: Our proposed QUD ($\tau = 0.1$, $n = 1024$) is compared to SOTA KD methods on all ten benchmarks described in Section 4.1, and the results are shown in Table 3. QUD and all SOTA KD methods in Table 3 are using MS1MV2 [17, 21] as distillation dataset, ResNet-50 [22] as teacher and MFN [11] as student. Not all methods were proposed for FR [13, 24, 38, 39, 46] or evaluated on all used benchmarks [2, 20, 41, 42] in the respective works. When available, results are obtained from [19, 27, 34, 45].

The following observations can be made from the results:

Small-Scale Benchmarks: On LFW, CFP-FP and AgeDB30 our QUD achieved competitive results to recent KD methods. On CA-LFW and CP-LFW and on average accuracy, our QUD ranked first.

Large-Scale Benchmarks IJB-B and IJB-C: QUD ranked first in IJB-B and achieved competitive results on IJB-C on TAR at FAR=10^{-4}.

ICCV21-MFR: Our QUD ranked first on the Mask subset of the ICCV21-MFR challenge and achieved competitive results on Children and MF-all subsets of the ICCV21-MFR

Table 3: Overview of the evaluation performance achieved by QUD and SOTA KD methods on the benchmarks and metrics described in Section 4.1 and Section 4.2, respectively. It is evident that our QUD achieved competitive results on all common benchmarks including MegaFace and TPR at FPR= 10^{-5} on IJB-C, when compared to the other methods.

Method	LFW	CFP-FP	AgeDB-30	CA-LFW	CP-LFW	Avg.	IJB-C		IJB-B		ICCV21-MFR			MegaFace			
							10^{-4}	10^{-5}	10^{-4}	10^{-5}	MF-all	Children	Mask	Id(R)	Ver(R)	Id	Ver
ResNet-50 (Teacher)	99.80	97.63	97.92	96.05	92.50	96.78	95.16	92.66	93.45	88.65	75.48	49.41	54.50	98.14	98.34	80.62	96.83
MFN (Student)	99.52	91.66	95.82	95.12	87.93	94.01	89.13	81.65	87.07	74.63	53.43	24.71	27.90	90.91	92.71	75.52	90.80
FitNet [8]	99.47	91.30	96.18	95.12	88.30	94.07	87.76	73.71	86.35	70.19	54.46	26.62	28.47	91.16	92.34	75.88	90.64
KD [8]	99.50	91.71	95.93	95.03	87.85	94.00	88.37	80.39	86.08	74.30	50.77	26.36	25.74	90.40	92.00	75.81	90.07
DarkRank [8]	99.55	91.84	95.60	95.07	87.77	93.97	89.28	81.62	86.76	73.75	56.82	28.84	30.07	90.76	92.41	75.80	90.66
SP [8]	99.53	92.33	96.17	95.07	88.45	94.31	88.43	78.13	86.34	72.85	54.44	26.63	29.75	91.25	92.41	75.37	90.62
CCKD [8]	99.47	91.90	95.83	95.22	88.48	94.18	87.99	78.75	85.63	72.38	55.64	27.65	30.22	91.17	92.76	75.73	90.63
RKD [8]	99.58	92.13	96.18	95.25	87.97	94.22	89.65	83.21	87.27	75.17	53.92	27.91	27.94	91.44	92.92	75.73	91.21
ShrinkTeaNet [8]	99.47	91.97	96.00	94.98	88.52	94.19	87.80	79.78	85.31	75.23	55.28	27.73	30.24	90.73	92.32	75.55	90.56
Trip.Dist. [8]	99.55	93.14	95.53	94.97	88.03	94.24	84.57	76.65	81.88	70.51	-	-	-	86.52	88.75	71.93	91.35
MarginKD [8]	*99.61*	92.01	96.55	95.13	88.03	94.27	85.71	75.00	82.97	66.25	50.73	25.14	28.54	91.70	92.96	76.34	91.31
SFTN [8]	99.48	92.77	96.30	-	-	-	90.96	82.67	-	-	55.50	28.51	29.66	91.69	93.38	-	-
EKD [8]	99.60	94.33	96.48	*95.37*	*89.35*	95.03	90.48	84.00	88.35	76.60	56.60	28.95	32.14	91.02	93.08	75.54	91.42
SH-KD [8]	99.47	*94.67*	96.53	-	-	-	91.75	85.76	-	-	*57.69*	30.15	32.01	**92.51**	**93.93**	-	-
ReFO [8]	99.55	94.51	**96.92**	-	-	-	*92.23*	*87.55*	-	-	56.63	**33.36**	31.88	92.38	93.80	-	-
ReFO+ [8]	**99.65**	**94.77**	96.42	-	-	-	**92.41**	**87.80**	-	-	***59.17***	*32.80*	*32.24*	92.41	93.75	-	-
QUD (Ours)	99.58	93.89	*96.73*	**95.65**	**90.47**	**95.26**	91.92	81.60	**90.13**	**78.08**	52.74	28.85	**33.92**	*92.42*	*93.84*	**76.46**	**92.19**

challenge.

Large-Scale Benchmark MegaFace: On MegaFace, QUD ranked first on the original benchmark and second on the refined benchmark (R), for both MegaFace evaluation protocols.

VII) Limitations and social impact: QUD indeed boosted the SOTA performances of computationally compact FR models, but it still falls behind in comparison to computationally demanding solutions [3, 17]. This issue is especially sensitive in security and safety-sensitive use cases. The training, as well as the KD process, of FR models commonly involves using, sharing and collecting authentic biometric data that if managed wrongly, can contradict with proper user consent. With the rise in synthetic data usage [8], this work takes a leap forward and uses synthetically generated data to validate its applicability in the QUD training.

6 Conclusion

This paper presented a novel unsupervised FR KD solution, QUD. This approach drives the student to optimize its features so that they are placed close to the teacher features of the same samples, but far away from queued samples from previous distillation iterations. This results in not requiring labeled data, not forcing the student to strictly learn the exact teacher features, and considering samples beyond the current batch. Detailed experimental analyses were presented to build design choices and prove the applicability over different teacher architectures and distillation datasets, even unlabeled synthetic ones. A wide comparison with recent SOTA FR KD approaches has shown our QUD as very competitive on a large set of benchmarks, including being the top-performing solution on many of these benchmarks.

Acknowledgments

This research work has been funded by the German Federal Ministry of Education and Research and the Hessian State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[1] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, and Ying Fu. Partial FC: training 10 million identities on a single machine. In *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021*, pages 1445–1449. IEEE, 2021. doi: 10.1109/ICCVW54120.2021.00166. URL https://doi.org/10.1109/ICCVW54120.2021.00166.

[2] Emanuel Ben Baruch, Matan Karklinsky, Yossi Biton, Avi Ben-Cohen, Hussam Lawen, and Nadav Zamir. It's all in the head: Representation knowledge distillation through classifier sharing. *CoRR*, abs/2201.06945, 2022. URL https://arxiv.org/abs/2201.06945.

[3] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, pages 1577–1586. IEEE, 2022. doi: 10.1109/CVPRW56347.2022.00164. URL https://doi.org/10.1109/CVPRW56347.2022.00164.

[4] Fadi Boutros, Naser Damer, and Arjan Kuijper. Quantface: Towards lightweight face recognition by synthetic data low-bit quantization. In *26th International Conference on Pattern Recognition, ICPR 2022, Montreal, QC, Canada, August 21-25, 2022*, pages 855–862. IEEE, 2022. doi: 10.1109/ICPR56361.2022.9955645. URL https://doi.org/10.1109/ICPR56361.2022.9955645.

[5] Fadi Boutros, Patrick Siebke, Marcel Klemt, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Pocketnet: Extreme lightweight face recognition network using neural architecture search and multistep knowledge distillation. *IEEE Access*, 10:46823–46833, 2022. doi: 10.1109/ACCESS.2022.3170561. URL https://doi.org/10.1109/ACCESS.2022.3170561.

[6] Fadi Boutros, Jonas Henry Grebe, Arjan Kuijper, and Naser Damer. Idiff-face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19650–19661, October 2023.

[7] Fadi Boutros, Marcel Klemt, Meiling Fang, Arjan Kuijper, and Naser Damer. Unsupervised face recognition using unlabeled synthetic data. In *17th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2023, Waikoloa Beach, HI, USA, January 5-8, 2023*, pages 1–8. IEEE, 2023. doi: 10.1109/FG57933.2023.10042627. URL https://doi.org/10.1109/FG57933.2023.10042627.

[8] Fadi Boutros, Vitomir Struc, Julian Fiérrez, and Naser Damer. Synthetic data for face recognition: Current state and future prospects. *Image Vis. Comput.*, 135:104688, 2023. doi: 10.1016/J.IMAVIS.2023.104688. URL https://doi.org/10.1016/j.imavis.2023.104688.

[9] Fadi Boutros, Vitomir Struc, and Naser Damer. Adadistill: Adaptive knowledge distillation for deep face recognition. *CoRR*, abs/2407.01332, 2024. doi: 10.48550/ARXIV.2407.01332. URL https://doi.org/10.48550/arXiv.2407.01332.

[10] Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In Tina Eliassi-Rad, Lyle H. Ungar, Mark Craven, and Dimitrios Gunopulos, editors, *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, pages 535–541. ACM, 2006. doi: 10.1145/1150402.1150464. URL https://doi.org/10.1145/1150402.1150464.

[11] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In Jie Zhou, Yunhong Wang, Zhenan Sun, Zhenhong Jia, Jianjiang Feng, Shiguang Shan, Kurban Ubul, and Zhenhua Guo, editors, *Biometric Recognition - 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings*, volume 10996 of *Lecture Notes in Computer Science*, pages 428–438. Springer, 2018. doi: 10.1007/978-3-319-97909-0_46. URL https://doi.org/10.1007/978-3-319-97909-0_46.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020. URL http://proceedings.mlr.press/v119/chen20j.html.

[13] Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *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*, pages 2852–2859. AAAI Press, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17147.

[14] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4793–4801. IEEE, 2019. doi: 10.1109/ICCV.2019.00489. URL https://doi.org/10.1109/ICCV.2019.00489.

[15] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/63c3ddcc7b23daa1e42dc41f9a44a873-Abstract.html.

[16] Jun Dan, Yang Liu, Haoyu Xie, Jiankang Deng, Haoran Xie, Xuansong Xie, and Baigui Sun. Transface: Calibrating transformer training for face recognition from a data-centric perspective. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 20585–20596. IEEE,

2023. doi: 10.1109/ICCV51070.2023.01887. URL https://doi.org/10.1109/ICCV51070.2023.01887.

[17] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4690–4699. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00482.

[18] Jiankang Deng, Jia Guo, Xiang An, Zheng Zhu, and Stefanos Zafeiriou. Masked face recognition challenge: The insightface track report. In *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021*, pages 1437–1444. IEEE, 2021. doi: 10.1109/ICCVW54120.2021.00165. URL https://doi.org/10.1109/ICCVW54120.2021.00165.

[19] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, and Ngan Le. Shrinkteanet: Million-scale lightweight face recognition via shrinking teacher-student networks. *CoRR*, abs/1905.10620, 2019. URL http://arxiv.org/abs/1905.10620.

[20] Yushu Feng, Huan Wang, Haoji Roland Hu, Lu Yu, Wei Wang, and Shiyan Wang. Triplet distillation for deep face recognition. In *IEEE International Conference on Image Processing, ICIP 2020, Abu Dhabi, United Arab Emirates, October 25-28, 2020*, pages 808–812. IEEE, 2020. doi: 10.1109/ICIP40778.2020.9190651. URL https://doi.org/10.1109/ICIP40778.2020.9190651.

[21] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, volume 9907 of *Lecture Notes in Computer Science*, pages 87–102. Springer, 2016. doi: 10.1007/978-3-319-46487-9_6. URL https://doi.org/10.1007/978-3-319-46487-9_6.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00975. URL https://doi.org/10.1109/CVPR42600.2020.00975.

[24] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. URL http://arxiv.org/abs/1503.02531. NIPS 2014 Deep Learning Workshop.

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

[26] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from A stronger teacher. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/da669dfd3c36c93905a17ddba01eef06-Abstract-Conference.html.

[27] Yuge Huang, Jiaxiang Wu, Xingkun Xu, and Shouhong Ding. Evaluation-oriented knowledge distillation for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18719–18728. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01818. URL https://doi.org/10.1109/CVPR52688.2022.01818.

[28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8107–8116. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00813. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Karras_Analyzing_and_Improving_the_Image_Quality_of_StyleGAN_CVPR_2020_paper.html.

[29] Ira Kemelmacher-Shlizerman, Steven M. Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4873–4882. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.527. URL https://doi.org/10.1109/CVPR.2016.527.

[30] Jan Niklas Kolf, Fadi Boutros, Florian Kirchbuchner, and Naser Damer. Lightweight periocular recognition through low-bit quantization. In *IEEE International Joint Conference on Biometrics, IJCB 2022, Abu Dhabi, United Arab Emirates, October 10-13, 2022*, pages 1–12. IEEE, 2022. doi: 10.1109/IJCB54206.2022.10007980. URL https://doi.org/10.1109/IJCB54206.2022.10007980.

[31] Jan Niklas Kolf, Fadi Boutros, Jurek Elliesen, Markus Theuerkauf, Naser Damer, Mohamad Alansari, Oussama Abdul Hay, Sara Alansari, Sajid Javed, Naoufel Werghi, Klemen Grm, Vitomir Struc, Fernando Alonso-Fernandez, Kevin Hernandez-Diaz, Josef Bigün, Anjith George, Christophe Ecabert, Hatef Otroshi-Shahreza, Ketan Kotwal, Sébastien Marcel, Iurii Medvedev, Bo Jin, Diogo Nunes, Ahmad Hassanpour, Pankaj Khatiwada, Aafan Ahmad Toor, and Bian Yang. Efar 2023: Efficient face recognition competition. In *IEEE International Joint Conference on Biometrics, IJCB 2023, Ljubljana, Slovenia, September 25-28, 2023*, pages 1–12. IEEE, 2023. doi: 10.1109/IJCB57857.2023.10448917. URL https://doi.org/10.1109/IJCB57857.2023.10448917.

[32] Jan Niklas Kolf, Tim Rieber, Jurek Elliesen, Fadi Boutros, Arjan Kuijper, and Naser Damer. Identity-driven three-player generative adversarial network for synthetic-based

face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 806–816. IEEE, 2023. doi: 10.1109/CVPRW59228.2023.00088. URL https://doi.org/10.1109/CVPRW59228.2023.00088.

[33] Jan Niklas Kolf, Jurek Elliesen, Naser Damer, and Fadi Boutros. Mixquantbio: Towards extreme face and periocular recognition model compression with mixed-precision quantization. *Engineering Applications of Artificial Intelligence*, 137: 109114, 2024. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2024.109114. URL https://www.sciencedirect.com/science/article/pii/S0952197624012727.

[34] Jingzhi Li, Zidong Guo, Hui Li, Seungju Han, Ji-won Baek, Min Yang, Ran Yang, and Sungjoo Suh. Rethinking feature-based knowledge distillation for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20156–20165, June 2023.

[35] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=rJlnB3C5Ym.

[36] Brianna Maze, Jocelyn C. Adams, James A. Duncan, Nathan D. Kalka, Tim Miller, Charles Otto, Anil K. Jain, W. Tyler Niggel, Janet Anderson, Jordan Cheney, and Patrick Grother. IARPA janus benchmark - C: face dataset and protocol. In *2018 International Conference on Biometrics, ICB 2018, Gold Coast, Australia, February 20-23, 2018*, pages 158–165. IEEE, 2018. doi: 10.1109/ICB2018.2018.00033. URL https://doi.org/10.1109/ICB2018.2018.00033.

[37] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: The first manually collected, in-the-wild age database. In *2017 IEEE CVPRW, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1997–2005. IEEE Computer Society, 2017. doi: 10.1109/CVPRW.2017.250. URL https://doi.org/10.1109/CVPRW.2017.250.

[38] Dae Young Park, Moon-Hyun Cha, Changwook Jeong, Daesin Kim, and Bohyung Han. Learning student-friendly teacher networks for knowledge distillation. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 13292–13303, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/6e7d2da6d3953058db75714ac400b584-Abstract.html.

[39] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3967–3976. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00409. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Park_Relational_Knowledge_Distillation_CVPR_2019_paper.html.

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learni pdf.

[41] Baoyun Peng, Xiao Jin, Dongsheng Li, Shunfeng Zhou, Yichao Wu, Jiaheng Liu, Zhaoning Zhang, and Yu Liu. Correlation congruence for knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5006–5015. IEEE, 2019. doi: 10.1109/ICCV.2019.00511. URL https://doi.org/10.1109/ICCV.2019.00511.

[42] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6550.

[43] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Domingo Castillo, Vishal M. Patel, Rama Chellappa, and David W. Jacobs. 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*, pages 1–9. IEEE Computer Society, 2016. doi: 10.1109/WACV.2016.7477558. URL https://doi.org/10.1109/WACV.2016.7477558.

[44] Riccardo Spolaor, QianQian Li, Merylin Monaro, Mauro Conti, Luciano Gamberini, and Giuseppe Sartori. Biometric authentication methods on smartphones: A survey. *PsychNology Journal*, 14(2), 2016.

[45] David Svitov and Sergey Alyamkin. Margindistillation: distillation for margin-based softmax. *CoRR*, abs/2003.02586, 2020. URL https://arxiv.org/abs/2003.02586.

[46] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 1365–1374. IEEE, 2019. doi: 10.1109/ICCV.2019.00145. URL https://doi.org/10.1109/ICCV.2019.00145.

[47] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748.

[48] Chaofei Wang, Qisen Yang, Rui Huang, Shiji Song, and Gao Huang. Efficient knowledge distillation from model checkpoints. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information*

Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/03e0712bf85ebe7cec4f1a7fc53216c9-Abstract-Conference.html.

[49] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021. doi: 10.1016/J.NEUCOM.2020.10.081. URL https://doi.org/10.1016/j.neucom.2020.10.081.

[50] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR, 2020. URL http://proceedings.mlr.press/v119/wang20k.html.

[51] Xiaobo Wang, Shuo Wang, Hailin Shi, Jun Wang, and Tao Mei. Co-mining: Deep face recognition with noisy labels. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9357–9366. IEEE, 2019. doi: 10.1109/ICCV.2019.00945. URL https://doi.org/10.1109/ICCV.2019.00945.

[52] Lilian Weng. Contrastive representation learning. *lilianweng.github.io*, May 2021. URL https://lilianweng.github.io/posts/2021-05-31-contrastive/.

[53] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn C. Adams, Tim Miller, Nathan D. Kalka, Anil K. Jain, James A. Duncan, Kristen Allen, Jordan Cheney, and Patrick Grother. 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*, pages 592–600. IEEE Computer Society, 2017. doi: 10.1109/CVPRW.2017.87.

[54] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3733–3742. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00393. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Wu_Unsupervised_Feature_Learning_CVPR_2018_paper.html.

[55] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch. *CoRR*, abs/1411.7923, 2014. URL http://arxiv.org/abs/1411.7923.

[56] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[57] Oliver Zhang, Mike Wu, Jasmine Bayrooti, and Noah D. Goodman. Temperature as uncertainty in contrastive learning. *CoRR*, abs/2110.04403, 2021. URL https://arxiv.org/abs/2110.04403.

[58] T. Zheng and W. Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. Technical Report 18-01, Beijing University of Posts and Telecommunications, February 2018.

[59] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. *CoRR*, abs/1708.08197, 2017. URL http://arxiv.org/abs/1708.08197.