# Rethinking Feature-based Knowledge Distillation for Face Recognition

Jingzhi Li[†,*,1], Zidong Guo[*,1], Hui Li[1], Seungju Han[2], Ji-won Baek[2], Min Yang[1], Ran Yang[1], Sungjoo Suh[2]

[1]Samsung R&D Institute China Xi'an (SRCX)

[2]Samsung Advanced Institute of Technology (SAIT), South Korea

jingzhi.li, zidong.guo, hui01.li, sj75.han, jw0328.baek, min16.yang, ran01.yang, sungjoo.suh@samsung.com

## Abstract

*With the continual expansion of face datasets, feature-based distillation prevails for large-scale face recognition. In this work, we attempt to remove identity supervision in student training, to spare the GPU memory from saving massive class centers. However, this naive removal leads to inferior distillation result. We carefully inspect the performance degradation from the perspective of intrinsic dimension, and argue that the gap in intrinsic dimension, namely the intrinsic gap, is intimately connected to the infamous capacity gap problem. By constraining the teacher's search space with reverse distillation, we narrow the intrinsic gap and unleash the potential of feature-only distillation. Remarkably, the proposed reverse distillation creates universally student-friendly teacher that demonstrates outstanding student improvement. We further enhance its effectiveness by designing a student proxy to better bridge the intrinsic gap. As a result, the proposed method surpasses state-of-the-art distillation techniques with identity supervision on various face recognition benchmarks, and the improvements are consistent across different teacher-student pairs.*

## 1. Introduction

Despite the unceasing emergence of larger and more powerful models for face recognition (FR), industrial deployment continues to demand for accurate and lightweight solutions. Among other compression techniques like pruning [27] and quantization [21], knowledge distillation (KD) has been proven to be effective in producing high-performing compact model from well-trained teacher. Unlike classic KD [17] and its variants [14, 24, 43, 44] who distill on logits, most of the existing works on FR distill on features [11, 13] or feature-relations [8, 20, 35]. One key reason is that the massive and still growing number of identities (IDs) in FR datasets, such as the 2 million IDs in WebFace42M [45], make it too expensive to save extra teacher's class centers for logits distillation.

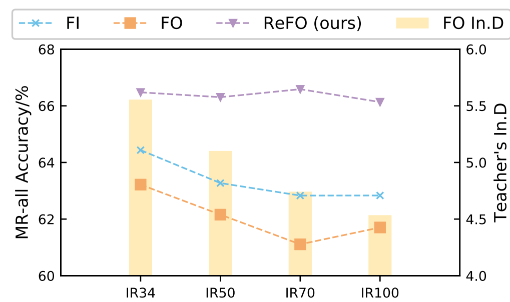



Figure 1. IResNet18 (IR18) is distilled by four different teachers. Feature-only distillation (FO) shows performance degradation comparing to feature-based distillation with ID supervision (FI). The proposed method (ReFO) significantly uplifts the performance of FO distillation. For both FI and FO, the student performance drops with larger teachers of lower intrinsic dimension. In line plot: student performance (%) on MR-all benchmark [9]. In bar plot: teacher's intrinsic dimension (In.D).

The ground truth supervision from ID labels, which we call ID supervision, is still retained when training student models for better distillation results. Nonetheless, it is not only non-trivial to find the right balancing weight [15, 33], the obtained class centers are also not needed during inference in an open-set FR problem. This motivates the complete removal of class centers in the student training for a number of benefits: 1) speed, the student distillation breaks free from the need of keeping any class center, providing further training speed-up with even lower GPU memory occupancy; 2) access to unlabeled dataset, removing the dependency on ID labels conveniently opens the door to the vast quantity of unlabeled or uncleaned face images like WebFace260M [45]; and 3) better focus on feature space, which is what really matters in an open-set problem. Hence, in this work, we are motivated to investigate feature distilla-

*Equal contribution. † Corresponding author.

tion for face recognition without ID supervision, which we call **feature-only (FO) distillation**.

The capacity gap problem is widely observed in various KD applications [7, 19, 30, 37], where the student finds it increasingly difficult to learn from more powerful teacher due to larger mismatch in network capacity. In FO distillation, the naive removal of ID supervision degrades student performance with more severe capacity gap problem. As shown in Fig. 1, comparing to the conventional feature distillation with ID supervision (FI distillation), the IResNet18 (IR18) students trained by four other teachers all experience drops in performance when ID supervision is removed.

Pertinent works commonly agree that differing model sizes cause the capacity gap issue [7, 20, 30, 40]. Some remedies were proposed to mitigate the problem such as early stopping [7] and training teacher assistants as intermediate agents [30]. Liu et al. [26] further proved the importance of teacher-student structural compatibility. For a given teacher, their best student from Neural Architecture Search outperformed other candidates of similar model size in the search space. However, recent works like [3, 32] showed that teachers of the same structure, same parameter size and comparable accuracy can also have differing distillation results on the same student. Hence, there must be other factors contributing to the capacity gap problem other than model size and model structure.

In this work, we argue that the teacher-student gap in intrinsic dimension, namely the **intrinsic gap**, plays a part. The intrinsic dimension [2, 16, 36] of a feature space is the minimum number of variables needed to unambiguously describe all points in the feature space. Specifically for a model, lower intrinsic dimension is often associated with better generalization power and better performance for both general classification [2] and face recognition [16]. In Fig. 1, as the teacher gets stronger with lower intrinsic dimension, we observe a drop in student performance with wider intrinsic gap for both FI distillation and FO distillation. If narrower intrinsic gap is related to better distillation result, can the capacity gap problem be mitigated by closing the intrinsic gap? This sparkles the idea that whether it is possible to narrow the intrinsic gap by raising teacher's intrinsic dimension for easier student-learning, neither changing its model size nor model structure.

Firstly, we revisit FO distillation and point out the intrinsic gap as another factor that could cause ineffective distillation. Then a reverse distillation strategy is proposed to solve the problem by injecting knowledge about higher intrinsic dimensional feature space into the teacher training. With reverse-distilled teachers, students trained with just FO distillation loss like mean-square-error (MSE) show performance on par or even better than competitors trained by sophisticatedly designed distillation loss with ID supervision [20, 35]. The proposed method is thus fast and versatile, it can be online or offline and easily portable to unlabeled datasets. On top of that, we further improve the distillation results by allowing the teacher to learn from more light-weight student proxies. This better closes the intrinsic gap and we are able to obtain state-of-the-art (SOTA) student models on popular face recognition benchmarks.

To summarize, the contribution of this work includes:

- We reconsider the capacity gap issue in FO distillation and provide an alternative view from the perspective of the intrinsic dimension. The gap in the intrinsic dimension between the teacher and the student is found to be related to the distillation performance.
- We propose a novel training scheme that narrows the teacher-student intrinsic gap via reverse distillation in the teacher training. Furthermore, we enhance its effectiveness by designing light-weight student proxies as the reverse distillation targets. Students trained by the new teachers show consistent performance improvement on FO distillation.
- Our method pushes the limit of FO distillation with easier-to-learn teacher. With only feature distillation loss, resulting students are shown to be superior than students trained by other SOTA distillation techniques with ID supervision.

## 2. Related Works

**Feature-based Knowledge Distillation.** Over the past decade, numerous distillation techniques emerged studying where, what and how to distill. For face recognition, feature-based distillation techniques are the most relevant. FitNets [38] proposed to distill the intermediate feature maps with the help of a regressor for dimension matching. AT [25] encouraged the attention maps of the teacher and the student to be similar. Works like FT [23] further studied how to transform teacher features and student features for efficient distillation. These methods focus on individual data point and are usually referred as instance-level distillation. From another perspective, relation-based distillations focus on preserving the structural information between features. RKD [33] proposed to transfer mutual relations in a mini-batch via pair-wise distance loss and triplet-wise angle loss on embeddings. CCKD [35] used the batch feature correlation matrix as the medium for knowledge transfer.

Works specialized in face recognition are also worth mentioning. ShrinkTeaNet [11] proposed to minimize the angle between each teacher-student embedding pair. MarginDistillation [8] reused teacher's class weights in the student training and forced the student to have the same sample-to-prototype margin as the teacher. TripletDistillation [13] followed triplet-based training scheme and encouraged the student margin to be simialr to the teachers.

EKD [20] introduced a novel rank-based loss to select key pair-relations to be distilled to the student.

The above mentioned methods all put emphasis on student learning and neglect the teacher's compatibility to the student. Although some relational methods like EKD try to make learning easier by imposing less stringent constraints on the student, effective knowledge transfer can still be challenging with exceedingly difficult teacher.

**Knowledge Distillation with Customized Teachers.** Dealing with the notorious capacity gap problem, many works have also attempted to solve the issue from the teacher side. Mirzadeh et al. [30] proposed multi-step distillation via teacher assistant to bridge the gap, while Cho et al. [7] discovered that early stopping of the teacher training mitigates the problem. However, their effectiveness heavily depends on choice of the right intermediate network structure or the right epoch for early stopping. More recently, SH-KD in [3] proposed to freeze the student classifier weights for the teacher training. SFTN [32] trained teachers to optimize the student branches jointly with ID supervision, providing a snapshot of the student in the teacher training. It needs special design of the joint-training position and the distillation has to be online, requiring the teacher backbone running multiple forward inferences during the student distillation. This adversely affects distillation efficiency since teacher model tends to be large.

These works all used ID supervision in the student training. In the proposed method, the student is distilled with just feature distillation loss. In our training of student-aware teachers, we do not introduce any additional module and there is no special design in the training loss.

## 3. Method

In this section, we first review the capacity gap problem in FO distillation. A connection is established between the teacher-student intrinsic gap and the student's inability to reproduce the teacher's feature space. Reverse distillation is then proposed as a remedy to the problem. Moreover, we improve the strategy by designing more light-weight student proxies used in reverse distillation, and further enhance the distillation result with narrower intrinsic gap.

### 3.1. Feature-only Distillation and the Intrinsic Gap

The general loss function used in KD can be written as:

$$L = \gamma L_{cls} + \alpha L_{logit} + \beta L_{feat}, \tag{1}$$

where $L_{cls}$ denotes the classification loss with ground truth label, $L_{logit}$ and $L_{feat}$ refer to the distillation loss on logits and features respectively.

For FO distillation, $\gamma$ and $\alpha$ are both zero, concerning only with the design of the $L_{feat}$ term. For face recognition, the prevalent choice is to take certain distance metric on the network embeddings. Following common practices [3, 11, 35], we use MSE loss on normalized embeddings as shown in Eq. (2).

$$L_{emb} = L_{feat}(\boldsymbol{f}_s, \boldsymbol{f}_t) = \frac{1}{N}\sum_{i=1}^{N}\left\|\frac{\boldsymbol{f}_s^i}{\|\boldsymbol{f}_s^i\|_2} - \frac{\boldsymbol{f}_t^i}{\|\boldsymbol{f}_t^i\|_2}\right\|_2^2, \tag{2}$$

where $\boldsymbol{f}_s$ and $\boldsymbol{f}_t$ refer to student embedding vector and teacher embedding vector respectively, $N$ is the batch size. This is conceptually equivalent to matching embeddings on the unit hypersphere or minimizing their angular distances.

Beyer et al. [4] proposed to view distillation as a pure function matching task, where the student model is trained to reproduce every output of the teacher model. They removed $L_{cls}$ and performed function matching on logits. Similarly, our feature-only distillation is essentially a function matching task on the feature space of the embeddings.

Function matching in the feature space, however, is a much more stringent constraint than function matching on the logits. The later only specifies comparative similarities to the class prototypes, which allows the student model to establish its own preferred feature distribution as long as the sample-to-prototype relationships hold. Feature-based function matching, on the other hand, forces the student to mimic the entire teacher's feature space which can be too ambitious to handle. When ID supervision signal is available, the points that are challenging to imitate can be guided to attainable positions that satisfy the relational constraints imposed by ID supervision. In the absence of ID supervision, the student loses guidance for free exploration and relies solely on its ability to mimic the teacher.

The student's inability to mimic the teacher's feature space now lies at the center of the problem. As inspired by existing works on intrinsic dimension [2, 16], we estimate the intrinsic dimension of common face recognition models using the TwoNN [12] method as applied in [2]. The results are listed in Tab. 1, which show that, in general, weaker model inherently converges to a feature space of higher intrinsic dimension.

Geometrically, intrinsic dimension describes the compactness of feature manifold and often indicates model performance [2,16,28]. It represents the model's ability to generalize against noise and non-discriminating variables for the task. The lower the intrinsic dimension, the less non-relevant noise in the feature space. In the process of FO dis-

Table 1. The intrinsic dimension (In.D) of common face recognition models. Details of the calculation can be found in Sec.2 of the supplementary material.

| Model | MFN | ires18 | ires34 | ires50 | ires100 |
|---|---|---|---|---|---|
| In.D | 8.645 | 6.792 | 5.559 | 5.105 | 4.539 |

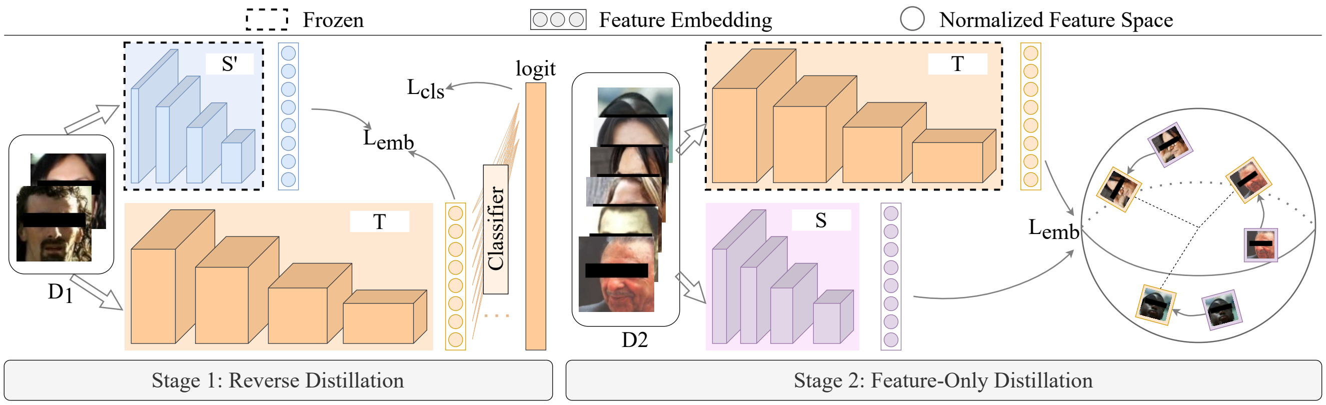



Figure 2. The proposed *ReFO* training scheme. $S'$ is a student model trained with standard supervision on dataset $D_1$. It is frozen to extract embeddings to guide the training of teacher $T$ with $L_{emb}$, and $T$ is additionally trained by $L_{cls}$ on $D1$. $T$ is then frozen to extract embeddings on $D_2$ which acts as the sole supervision for training final student $S$ with $L_{emb}$.

tillation, students learn to remove redundant information, transforming towards more compact and teacher-like manifold. Intrinsic gap essentially quantifies the complexity of the required transform hence the distillation difficulty.

### 3.2. Reverse Distillation

Based on the above interpretation, if the student has reached its bottleneck to mimic the teacher with lower intrinsic dimension, can the teacher raise its intrinsic dimension instead, to bridge the intrinsic gap and enabling easier student learning? Note that the intrinsic dimension is not an absolute performance predictor[1]. It is theoretically possible to obtain model with higher intrinsic dimension under additional constraint without compromising its performance.

In this section, we propose to solve the aforementioned problem by injecting knowledge about higher intrinsic dimensional feature space into the teacher training. As shown in Fig. 2, the overall distillation process can be achieved by a two-stage training scheme which we call **Reverse distillation empowered Feature-Only (*ReFO*) distillation.**

The first stage is the reverse distillation from the student to the teacher. First of all, an initial student $S'$ is trained on dataset $D1$ with ID supervision $L_{cls}$. The parameters of $S'$ are frozen to obtain its embeddings on $D1$. The teacher $T$ is then trained on $D1$. Besides $L_{cls}$, its optimization is guided with the embedding distillation loss $L_{emb}$ by the initial student $S'$. This essentially constrains the teacher's search space on higher intrinsic dimension, closer to the innate disposition of the student. We refer to the teacher as being **tailored** to $S'$, represented by $T \leftarrow S'$. In the second stage of FO distillation, we freeze the teacher's parameters to obtain its embeddings on dataset $D2$. These embeddings are used for the training of the final target student $S$. Finally, $S$ is trained only by the embedding distillation loss $L_{emb}$ with embeddings from $T$.

Formally, the proposed *ReFO* distillation is described in Algorithm 1. The distillation can be offline, where the features obtained in step 2 and 4 are saved in advance to avoid multiple forward inferences during training. For online distillation, these features can be generated on-site, providing consistent distillation view across data augmentation [4].

Intrinsic dimension ultimately depends on the embedding distribution in the feature space as it is estimated from distances between neighboring points (Suppl. Eq.1). Reverse distillation encourages the teacher's embedding distribution to resemble the student's, and essentially constrains the teacher to optimize in restricted search space of higher intrinsic dimension. Experiments in Sec. 4.3 show that this design is able to raise the teacher's intrinsic dimension and brings consistent improvements to students trained by FO distillation. The students generally converge faster and attain much lower MSE loss, finding the new student-aware feature space easier to learn.

Since the teacher training dataset $D1$ and the student training dataset $D2$ are independent and no ID supervision is required in the student training, the proposed method can easily exploit abundant unlabeled datasets as $D2$ to reap additional performance gains.

### 3.3. Further Bridging the Intrinsic Gap

Encouraged by the effectiveness of *ReFO*, we continue the pursuit of pushing the limit FO distillation. It is observed in Sec. 4.3.2 that teacher tailored to a specific student shows universal improvements on other students. For example, IResNet100 (IR100) tailored to IR18 brings 3% of improvement on IR34 as well on MR-all [9]. We are wondering if it is possible guide teacher's optimization with a student of even higher intrinsic dimension, so that the intrinsic gap can be better bridged. Observing that smaller models usually have higher intrinsic dimension, we propose

[1] E.g. the VGGs in Fig.4 of [2] does not compare meaningfully with the ResNets but the trend still holds within the VGG family.

**Algorithm 1** ***ReFO*** **Knowledge Distillation**

1: Train a student model $S'$ on dataset $D_1$ with standard classification loss for face recognition.
$$L = L_{cls}.$$
2: Obtain features of model $S'$ on dataset $D_1$.
3: Train tailored teacher model $T$ on dataset $D_1$ with classification loss and embedding distillation loss using features from step 2.
$$L = L_{cls} + \beta_1 L_{emb}(\boldsymbol{f}_t, \boldsymbol{f}_{s'}).$$
4: Obtain features of model $T$ on dataset $D_2$.
5: Train final student model $S$ on dataset $D_2$ with embedding distillation loss using features from step 4.
$$L = \beta_2 L_{emb}(\boldsymbol{f}_s, \boldsymbol{f}_t).$$

to design a light-weight student proxy as the target for reverse distillation.

Rather than using exactly the same student structure for stage 1 and 2, we propose to train a half-depth student proxy as $S'$ in step 1 of Algorithm 1. Specifically, for block-based network structure like IResNet and MobileFaceNet (MFN) [5], we reduce the number of blocks in each block group according to a pre-set ratio $S_d = 0.5$. When non-integer block number is incurred, we always round it down but ensuring it is at least 1. For example, the block number of an official MFN is $[4, 6, 3]$. If scale $S_d = 0.5$, the block number is set to $[2, 3, 1]$. The rest of the distillation procedure is the same as *ReFO*. This revised training scheme with student proxy is referred as **Enhanced-ReFO (*ReFO+*)**.

Besides depth reduction, there are many other ways to design light-weight student proxies. We examine a few choices in Sec. 4.4 and discover that there is a limit to how small the student can be. The optimal student proxy structure may vary for each target student network, and we leave the search for the optimal structure for future work. The intention of this work is to show that using a more light-weight student proxy can better close the intrinsic gap and bring further improvement to the student.

# 4. Experiments

## 4.1. Datasets

**Training.** We use MS1MV2 [10] as the standard training data for fair comparisons. Additionally, Glint360k [1] is used without ID labels to show our effectiveness on unlabeled dataset. MS1MV2 contains about 5.8M images of 85k individuals, while Glint360k contains 17M images.

**Testing.** Test results are reported on popular face benchmarks, including LFW [18], CFP-FP [39], AgeDB [31], IJB-C [29], MegaFace [22] and the newly proposed ICCV21-MFR [9]. The first three are typical face verification test sets. IJB-C is a challenging template-based benchmark with 3.5k IDs from images and wild video frames. MegaFace evaluates face recognition (FR) accuracy on 100k images belonging to 530 IDs under the 1M distractors images from 690k IDs. The largest and the most recently introduced ICCV21-MFR is a comprehensive large-scale benchmark for FR, containing the following three tracks: Mask, Children, and Multi-racial (MR-all). Specifically, Mask set contains 7k IDs, and Children set includes 14k IDs. The largest MR-all set contains 4.69M positive pairs and 2.6 trillion negative pairs, composed of 1.6M images involving 242k IDs. We adopt ICCV21-MFR as the primary criterion in design selection and ablation study.

## 4.2. Experimental Settings

**Network input & output.** We follow [10] to preprocess the data with five landmarks [42]. Network inputs have the size of $112 \times 112$ and are normalized to $[-1, 1]$. The output embedding size is 512.

**Teacher-student pair.** IResNet100-IResNet18 (IR100-IR18) and IResNet50-MobileFaceNet (IR50-MFN) are the two default teacher-student pairs. Various other networks are also investigated. All models from the IResNet family follow the original design in [10]. The standard MobileFaceNet (MFN) is used with the default channel scale. For MobileNetV2 (MNv2), the last Conv layer of the backbone is modified for embedding size consistency.

**Training.** All experiments are conducted on 8 NVIDIA Tesla V100 GPU with Pytorch [34]. All models are trained from scratch using SGD with 20 epochs. The batch size is 512 on each GPU, and the learning rate starts at 0.4 with poly scheduler. The momentum is 0.9 and the weight decay is $5e^{-4}$. The default weights for $\beta_1$ and $\beta_2$ are 0.5 and 5. The Arcface [10] loss with default settings is used as the ID supervision. Random flip with a probability of 0.5 is the only data augmentation strategy.

**Testing.** We follow prevailing test protocols in reporting model performance. Specifically, 10-fold validation is used for LFW, CFP-FP, AgeDB. For MegaFace, performance is reported with provided refinement. For 1:$N$ verification, track identification(*Id*) is reported for the rank-1 face identification accuracy with 1M distractors. For 1:1 verification, track verification(*Ver*) is reported for the face verification TAR at $1e^{-6}$ FAR. For IJB-C, we follow common test procedure as in [9, 10]. For ICCV21-MFR [9], we report the performance in all three tracks with official setting[2].

## 4.3. Results on ReFO

*ReFO* turns out to be surprisingly effective for FO distillation. As shown in Tab. 2, tailored teachers all have lower intrinsic gaps with the students. The narrower intrinsic gaps

[2]True Positive Rate (TPR) @ False Positive Rate (FPR) = $1e^{-6}$ for MR-all, TPR@FPR=$1e^{-4}$ for Children and Mask

Table 2. The Intrinsic gap and student performance of various students with IR100 as the common teacher. ReFO boosts all students' performance (%), evaluated on MR-all. Corresponding intrinsic gaps ($w.r.t.$ IR100) are found to be narrower.

| Student | Intrinsic Gap | | MR-all/% | |
|---|---|---|---|---|
| | FO | ReFO | FO | ReFO |
| MFN | 4.10 | 3.75 | 53.86 | 57.27 |
| MNv2 | 3.53 | 3.08 | 58.33 | 63.04 |
| IR18 | 2.25 | 2.03 | 61.70 | 66.13 |
| IR34 | 1.02 | 0.97 | 73.17 | 75.07 |

are manifested as better distillation results, and improvements are observed for multiple teacher-student pairs. On average, the students taught by the tailored teachers outperform their peers by 3.6%.

The students also converge faster during training and achieve lower MSE loss. As shown in Fig. 3a, the IR18 student trained by the tailored IR100 teacher settles at around half of the training loss compared to the one trained by the original IR100 teacher. The faster convergence and lower final loss suggest better and easier imitation of the teacher's feature space, which proves the effectiveness of *ReFO* and confirms that better student performance comes from an easier-to-learn feature space.

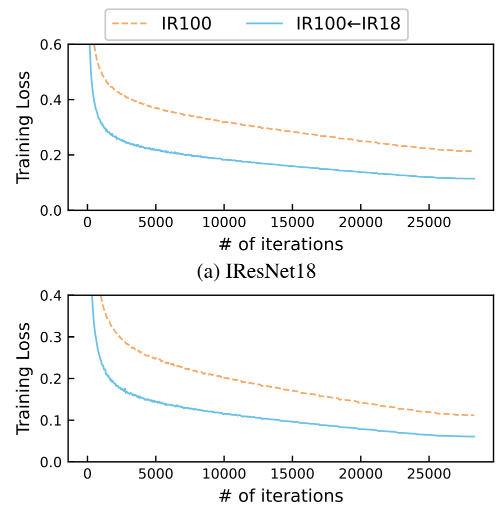
#### 4.3.1 Impact on Teacher

Tailored teachers have higher intrinsic dimension as shown in Tab. 3. This is observed for all teacher-student pairs, and smaller student model produces teacher with the higher intrinsic dimension. The absolute change may not appear significant, but the relative change in the intrinsic gap ranges from 4.9% to 12.7% with reference to Tab. 2. With the interpretation of intrinsic gap representaion distillation difficulty, the relative change in intrinsic gap matters more.

Moreover, teachers' accuracies are shown to be comparable or even better than the baseline trained with standard ID supervision. This sets *ReFO* apart from methods like early stopping [7] that lowers the teacher's performance to bridge the capacity gap. In contrast, reverse distillation pushes the teacher's intrinsic dimension higher by imposing extra constraints without compromising its accuracy. The teacher is able to find a solution in the higher intrinsic dimensional space that is of the similar level of performance.

Table 3. Reverse distillation raises teacher's intrinsic dimension (In.D) without lowering its performance (%), evaluated on MR-all.

| | Original IR100 | IR100 tailored to | | | |
|---|---|---|---|---|---|
| | | MFN | MNv2 | IR18 | IR34 |
| In.D | 4.53 | 4.91 | 4.98 | 4.75 | 4.58 |
| MR-all/% | 79.06 | 80.07 | 81.30 | 81.67 | 81.53 |




(a) IResNet18

(b) IResNet34

Figure 3. Training loss evolution of FO distillation with MSE loss. The students trained by tailored teachers show faster convergence and lower final loss. (a): IR18 model. (b): IR34 model. IR100: the student trained by original IR100. IR100←IR18: the student trained by IR100 tailored to IR18.

#### 4.3.2 Universally Friendly ReFO Teacher

Interestingly, the benefit from reverse distillation is observed to be non-exclusive to the student structure the teacher being tailored to. As shown in Tab. 4, it is clear that the teacher tailored to one student shows improvement on another. For instance, the MNv2 model trained by IR100 tailored to MFN enjoys a boost in performance by 4.9%. We repeat the experiment with teacher tailoring to IR18 and the same phenomenon is observed.

By raising the teacher's intrinsic dimension, reverse distillation has changed the teacher feature space in a generic way in favor of FO distillation for all students. Fig. 3b also shows that the IR34 model enjoys similar lowered training loss as IR18 when trained by the IR100 tailored to IR18.

Table 4. Teachers tailored to MFN and IR18 show universal improvements on students' accuracies (%) with FO distillation, evaluated on MR-all. A←B refers to A that is reverse-distilled by B.

| Teacher | Student | | | |
|---|---|---|---|---|
| | MFN | MNv2 | IR18 | IR34 |
| IR100 | 53.86 | 58.32 | 61.70 | 73.16 |
| IR100←MFN | 57.27 | 63.22 | 66.58 | 76.18 |
| IR100←IR18 | 56.53 | 63.57 | 66.13 | 76.01 |

### 4.4. Ablation Studies on ReFO+

#### 4.4.1 Ablation on Different Student Proxy

Inspired by the results in Tab. 4 that IR18 actually benefits more from IR100 tailored to MFN. It is natural to wonder if we can push the limit of FO distillation by designing a lightweight student proxy $S'$. The effectiveness of the proposed half-depth proxy of IR18 and MFN are presented as *ReFO+* in Tab. 6 and Tab. 7 respectively.

In this section, we further investigate the effect of different layer scale ratio $S_d$ using MFN as an example. In addition, we include experiments on channel slimming as an alternative option for designing student proxy. For channel slimming, we proportionally reduce the number of all channels to a pre-set ratio $S_c$ (0.25, 0.5, and 0.75), except for the final embedding which remains constant as 512.

As shown in Fig. 4, both designs bring extra performance boost compared to the *ReFO* baseline (found at scaling ratio $= 1.0$). Depth reduction shows superior performance overall. For depth reduction, the best result is obtained at $S_d = 0.5$ with the lowest intrinsic gap. When the $S_d$ continues to drop to 0.25, leaving the student with only 140k parameters, we observe a drop in student performance. This shows that there is a limit to how small the student can be. Notice that this drop in student performance is accompanied by a rise in intrinsic gap as $S_d$ changes from 0.5 to 0.25.

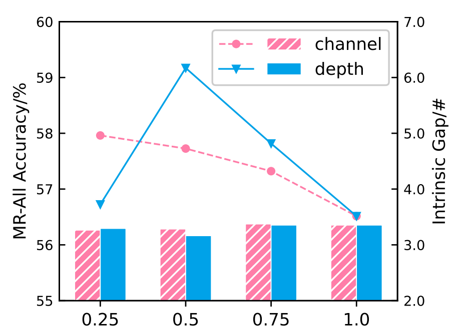



Figure 4. Effects of channel slimming and depth reduction on MFN. In line plot with markers: student performance (%) evaluated on MR-all. In bar plot: the teacher-student intrinsic gap.

#### 4.4.2 Ablation on Training Specification

We investigate the sensitivity of *ReFO+* with respect to a few training settings on the IR100-IR18 teacher-student pair. With reference to Tab. 5, *Norm* indicates whether L2 normalization is performed on embeddings during reverse distillation. $L^{Reverse}_{emb}$ and $L^{FO}_{emb}$ refer to the type of distance metric used in the reverse distillation and the FO distillation respectively. We perform experiments on a combination of these settings in Tab. 5. The student performance appear to be comparable, exhibiting good robustness against changes in training loss. The slightly better option in ***bold italic*** is used as the final training setting.

In the last row, we add back ID supervision on top of the optimal setting with $\gamma = 1.0$, and are surprised to find a slight drop in the student performance. This may be a result and inappropriate weight proportion in the loss function, which testifies the sensitivity of the balancing weight for ID supervision as mentioned in Sec. 1.

Table 5. Ablation of training specifics on IR18 shows robustness against changes in training loss. $L^{Reverse}_{emb}$ and $L^{FO}_{emb}$ are the embedding distance metric for stage 1 and 2 respectively. *SmL1* refers to Smooth L1 loss, and $L_{cls}$ is arcface loss with default settings.

| Stage 1 | | Stage 2 | | |
|---|---|---|---|---|
| Norm | $L^{Reverse}_{emb}$ | $L^{FO}_{emb}$ | $L_{cls}$ | MR-all |
| ✓ | MSE | MSE | × | 68.518 |
| × | MSE | MSE | × | 68.500 |
| × | ***SmL1*** | ***MSE*** | × | ***68.563*** |
| × | SmL1 | SmL1 | × | 68.239 |
| × | SmL1 | MSE | ✓ | 68.269 |

### 4.5. Comparison with SOTA Methods

In this section, we compare our methods with several SOTA competitors on various benchmarks using two teacher-student pairs (IR100-IR18 & IR50-MFN). For our method, we report the offline performance for *ReFO* and *ReFO+* with the settings specified in Sec. 4.2 and Sec. 4.4.

Additionally, since FO distillation can be easily extended to **U**nlabeled **D**ataset. We further present *ReFO+* (UD) as an example to demonstrate the amount of improvement we can obtain from a larger unlabeled face dataset.

In Tab. 6 and Tab. 7, we compare with general KD methods [6,17,33,35,38,41], FR specific KD methods [8,11,20] and student-aware KD methods [3, 32]. They are further grouped into three categories with horizontal rules for easier comparison. The first group are student-centric, where teachers are not given any information about students. The second group contains two recent student-aware methods. The third group encapsulates our *ReFO* variants which are also student-aware. When available, we cite the results from [8, 20]. Results that we additionally reproduced are labeled with $*$.

On the three small benchmarks (LFW, CFP-FP and AgeDB) most distillation techniques show comparable performances for both teacher-student pair. While EKD [20] and SH-KD [3] produce better results than the rest, our methods show the best performance. IJB-C and MegaFace evaluate model performance on 1:1 verification and 1:N identification. EKD [20], designed specially for 1:1 metric, and student-aware methods [3, 32] generally show better performance on these two benchmarks. Our methods are also among the top performers with comparable results.

The last 3 columns report the results on the largest and most comprehensive ICCV21-MFR benchmarks. Student-aware methods, SFTN, SH-KD and our *ReFO* variants, show clear advantage in this track for both teacher-student pairs. *ReFO+* demonstrates best performance overall, surpassing the best competitor on MR-all by 1.3% (IR18-IR100) and 1.48% (IR50-MFN).

With unlabeled data, *ReFO+ (UD)* easily outperforms *ReFO* and *ReFO+* by a significant margin on almost all benchmarks. On MR-all, it brings 3.79% (IR18-IR100) and

Table 6. Comparison with SOTA methods, the IR100-IR18 pair. $L_{cls}$: whether ID supervision is used. *ReFO+* attains overall SOTA performance and shows great advantage on the most comprehensive ICCV21-MFR benchmark. With unlabeled dataset, *ReFO+ (UD)* significantly outperforms *ReFO+*. The best and second best results excluding *ReFO+ (UD)* are in **bold** and *italic* respectively.

| Method | $L_{cls}$ | LFW* | CFP-FP* | AgeDB* | IJB-C* | | MegaFace* | | ICCV21-MFR* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $1e-4$ | $1e-5$ | Id(R) | Ver(R) | MR-all | Children | Mask |
| IR100 (teacher) | ✓ | 99.78 | 98.40 | 98.27 | 96.39 | 94.58 | 98.73 | 98.98 | 79.07 | 48.57 | 59.43 |
| IR18 (student) | ✓ | 99.67 | 94.60 | 97.33 | 93.99 | 91.14 | 96.22 | 96.66 | 65.97 | 38.44 | 45.41 |
| KD ('15) [17] | ✓ | **99.72** | 94.11 | 97.35 | 93.89 | 89.90 | *96.44* | 96.83 | 63.70 | 39.11 | 45.38 |
| FitNet ('15) [38] | ✓ | 99.68 | 95.07 | 97.60 | 94.18 | 91.21 | *96.44* | 96.72 | 65.53 | 40.04 | 44.55 |
| DarkRank ('18) [6] | ✓ | 99.65 | 94.84 | 97.70 | 94.22 | **91.31** | 96.42 | 96.86 | 66.23 | 37.95 | 45.80 |
| SP ('19) [41] | ✓ | 99.67 | 94.99 | 97.57 | 93.90 | 91.20 | 96.11 | 96.39 | 63.96 | 38.79 | 44.31 |
| CCKD ('19) [35] | ✓ | 99.70 | 93.57 | 97.33 | 93.58 | 89.85 | 96.01 | 96.51 | 61.19 | 33.01 | 42.89 |
| RKD ('19) [33] | ✓ | 99.52 | 93.46 | 97.00 | 93.56 | 90.20 | 95.87 | 96.31 | 63.69 | 38.96 | 44.14 |
| EKD ('22) [20] | ✓ | 99.63 | *95.95* | 97.73 | **94.37** | 90.60 | 96.23 | **97.17** | 65.45 | 39.98 | 46.01 |
| SFTN ('21) [32] | ✓ | 99.61 | 94.76 | 97.52 | 94.02 | 90.87 | 96.36 | 96.72 | 66.18 | *40.66* | 45.22 |
| SH-KD ('22) [3] | ✓ | 99.65 | 95.33 | *97.80* | *94.34* | 90.93 | **96.54** | *97.06* | *67.26* | 40.19 | 45.61 |
| **ReFO (ours)** | × | 99.65 | 95.79 | 97.63 | 94.31 | 90.90 | 96.42 | 96.96 | 66.13 | 40.46 | *47.09* |
| **ReFO+ (ours)** | × | **99.72** | **96.23** | **97.83** | 94.28 | **91.31** | 96.42 | *97.06* | **68.56** | **41.49** | **48.72** |
| **ReFO+ (UD) (ours)** | × | 99.65 | 97.39 | 97.70 | 94.95 | 92.47 | 96.90 | 97.33 | 72.35 | 43.54 | 53.78 |

Table 7. Comparison with SOTA methods, the IR50-MFN pair. $L_{cls}$: whether ID supervision is used. In general, *ReFO* or *ReFO+* are found among the top two performers. On the largest ICCV21-MFR benchmark, *ReFO+* demonstrates clear superiority. With unlabeled dataset, *ReFO+ (UD)* boosts the performance of *ReFO+* by a large margin for all benchmarks. The best and second best results excluding *ReFO+ (UD)* are in **bold** and *italic* respectively.

| Method | $L_{cls}$ | LFW | CFP-FP | AgeDB | IJB-C* | | MegaFace | | ICCV21-MFR* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $1e-4$ | $1e-5$ | Id(R) | Ver(R) | MR-all | Children | Mask |
| IR50 (teacher) | ✓ | 99.80 | 97.63 | 97.92 | 96.05 | 93.96 | 98.14 | 98.34 | 75.48 | 49.41 | 54.50 |
| MFN (student) | ✓ | 99.52 | 91.66 | 95.82 | 92.16 | 85.83 | 90.91 | 92.71 | 53.43 | 24.71 | 27.90 |
| KD ('15) [17] | ✓ | 99.50 | 91.71 | 95.93 | 86.96 | 69.98 | 90.40 | 92.00 | 50.77 | 26.36 | 25.74 |
| FitNet ('15) [38] | ✓ | 99.47 | 91.30 | 96.18 | 91.73 | 86.07 | 91.16 | 92.34 | 54.46 | 26.62 | 28.47 |
| DarkRank ('18) [6] | ✓ | 99.55 | 91.84 | 95.60 | 92.15 | 86.28 | 90.76 | 92.41 | 56.82 | 28.84 | 30.07 |
| SP ('19) [41] | ✓ | 99.53 | 92.33 | 96.17 | 91.79 | 87.22 | 91.25 | 92.41 | 54.44 | 26.63 | 29.75 |
| CCKD ('19) [35] | ✓ | 99.47 | 91.90 | 95.83 | 91.73 | 85.75 | 91.17 | 92.76 | 55.64 | 27.65 | 30.22 |
| RKD ('19) [33] | ✓ | 99.58 | 92.13 | 96.18 | 89.36 | 81.88 | 91.44 | 92.92 | 53.92 | 27.91 | 27.94 |
| ShrinkTeaNet ('19) [11] | ✓ | 99.47 | 91.97 | 96.00 | 91.50 | 86.23 | 90.73 | 92.32 | 55.28 | 27.73 | 30.24 |
| MarginKD ('21) [8] | ✓ | *99.61* | 92.01 | *96.55* | 91.02 | 83.39 | 91.70 | 92.96 | 50.73 | 25.14 | 28.54 |
| EKD ('22) [20] | ✓ | 99.60 | 94.33 | 96.48 | 92.28 | 86.47 | 91.02 | 93.08 | 56.60 | 28.95 | *32.14* |
| SFTN* ('21) [32] | ✓ | 99.48 | 92.77 | 96.30 | 90.96 | 82.67 | 91.69 | 93.38 | 55.50 | 28.51 | 29.66 |
| SH-KD* ('22) [3] | ✓ | 99.47 | *94.67* | 96.53 | 91.75 | 85.76 | **92.51** | **93.93** | *57.69* | 30.15 | 32.01 |
| **ReFO (ours)** | × | 99.55 | 94.51 | **96.92** | *92.23* | *87.55* | 92.38 | *93.80* | 56.63 | **33.36** | 31.88 |
| **ReFO+ (ours)** | × | **99.65** | **94.77** | 96.42 | **92.41** | **87.80** | *92.41* | 93.75 | **59.17** | *32.80* | **32.24** |
| **ReFO+ (UD) (ours)** | × | 99.67 | 95.61 | 97.07 | 93.51 | 89.41 | 93.23 | 94.16 | 63.32 | 33.21 | 37.72 |

$4.15\%$ (IR50-MFN) improvements on top of *ReFO+*.

## 5. Conclusion

In this work, we re-examined the capacity gap problem in FO distillation in the context of face recognition. Besides model size and model structure, we offered a new view on capacity gap from the perspective of teacher-student intrinsic gap. We proposed to narrow the intrinsic gap by incorporating reverse distillation in teacher training. The resulting teacher turned out to have universally easier-to-learn feature space for various student models. By designing more light-weight student proxies used in reverse distillation, the intrinsic gap was better bridged, yielding better performing student. With the proposed *ReFO+*, students trained by only MSE loss outperformed competitors trained by other advanced techniques with ID supervision.

## 6. Social Imapct and Limitation

Advocating for performance boosts with *ReFO+ (UD)*, this work may encourage the collection of large-scale face datasets, and possibly induces the unauthorized or inappropriate use of these highly personal identifiable images.

There are still many limitations to our understanding of intrinsic dimension in this work. We have yet to methodologically design the optimal student proxy and explore more effective method to close the intrinsic gap to fill the still significant performance gap.

## References


[1] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1445–1449, 2021. 5

[2] Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3, 4

[3] 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. *arXiv preprint arXiv:2201.06945*, 2022. 2, 3, 7, 8

[4] Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10925–10934, 2022. 3, 4

[5] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Chinese Conference on Biometric Recognition*, pages 428–438. Springer, 2018. 5

[6] Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 7, 8

[7] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802, 2019. 2, 3, 6

[8] Svitov David and Alyamkin Sergey. Margindistillation: Distillation for face recognition neural networks with margin-based softmax. *International Journal of Computer and Information Engineering*, 15(3):206–210, 2021. 1, 2, 7, 8

[9] Jiankang Deng, Jia Guo, Xiang An, Zheng Zhu, and Stefanos Zafeiriou. Masked face recognition challenge: The insightface track report. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1437–1444, 2021. 1, 4, 5

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 5

[11] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, and Ngan Le. Shrinkteanet: Million-scale lightweight face recognition via shrinking teacher-student networks. *arXiv preprint arXiv:1905.10620*, 2019. 1, 2, 3, 7, 8

[12] Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):1–8, 2017. 3

[13] Yushu Feng, Huan Wang, Haoji Roland Hu, Lu Yu, Wei Wang, and Shiyan Wang. Triplet distillation for deep face recognition. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 808–812. IEEE, 2020. 1, 2

[14] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR, 2018. 1

[15] Mengya Gao, Yujun Shen, Quanquan Li, Junjie Yan, Liang Wan, Dahua Lin, Chen Change Loy, and Xiaoou Tang. An embarrassingly simple approach for knowledge distillation. *arXiv preprint arXiv:1812.01819*, 2018. 1

[16] Sixue Gong, Vishnu Naresh Boddeti, and Anil K Jain. On the intrinsic dimensionality of image representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3987–3996, 2019. 2, 3

[17] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 1, 7, 8

[18] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 5

[19] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *arXiv preprint arXiv:2205.10536*, 2022. 2

[20] Yuge Huang, Jiaxiang Wu, Xingkun Xu, and Shouhong Ding. Evaluation-oriented knowledge distillation for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18740–18749, 2022. 1, 2, 3, 7, 8

[21] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713, 2018. 1

[22] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 5

[23] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems*, 31, 2018. 2

[24] Youmin Kim, Jinbae Park, YounHo Jang, Muhammad Ali, Tae-Hyun Oh, and Sung-Ho Bae. Distilling global and local logits with densely connected relations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6290–6300, 2021. 1

[25] Nikos Komodakis and Sergey Zagoruyko. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017. 2

[26] Yu Liu, Xuhui Jia, Mingxing Tan, Raviteja Vemulapalli, Yukun Zhu, Bradley Green, and Xiaogang Wang. Search to distill: Pearls are everywhere but not the eyes. In *Proceed-*

*ings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7539–7548, 2020. 2

[27] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. *International Conference on Learning Representations*, 2019. 1

[28] Xingjun Ma, Yisen Wang, Michael E Houle, Shuo Zhou, Sarah Erfani, Shutao Xia, Sudanthi Wijewickrema, and James Bailey. Dimensionality-driven learning with noisy labels. In *International Conference on Machine Learning*, pages 3355–3364. PMLR, 2018. 3

[29] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 international conference on biometrics (ICB)*, 2018. 5

[30] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198, 2020. 2, 3

[31] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017. 5

[32] Dae Young Park, Moon-Hyun Cha, Daesin Kim, Bohyung Han, et al. Learning student-friendly teacher networks for knowledge distillation. *Advances in Neural Information Processing Systems*, 34:13292–13303, 2021. 2, 3, 7, 8

[33] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019. 1, 2, 7, 8

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[35] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5007–5016, 2019. 1, 2, 3, 7, 8

[36] Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. *International Conference on Learning Representations*, 2021. 2

[37] Zengyu Qiu, Xinzhu Ma, Kunlin Yang, Chunya Liu, Jun Hou, Shuai Yi, and Wanli Ouyang. Better teacher better student: Dynamic prior knowledge for knowledge distillation. *arXiv preprint arXiv:2206.06067*, 2022. 2

[38] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015. 2, 7, 8

[39] Soumyadip Sengupta, Jun-Cheng Chen, Carlos 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. 5

[40] Wonchul Son, Jaemin Na, Junyong Choi, and Wonjun Hwang. Densely guided knowledge distillation using multiple teacher assistants. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9395–9404, 2021. 2

[41] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1365–1374, 2019. 7, 8

[42] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 2016. 5

[43] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4320–4328, 2018. 1

[44] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11953–11962, 2022. 1

[45] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021. 1