Exclusivity-Consistency Regularized Knowledge Distillation for Face Recognition

Xiaobo Wang[1⋆], Tianyu Fu[1⋆], Shengcai Liao[2], Shuo Wang[1], Zhen Lei[3,4†], and Tao Mei[1]

[1] JD AI Research, Beijing, China
[2] Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, UAE
[3] CBSR&NLPR, Institute of Automation, Chinese Academy of Science
[4] School of Artificial Intelligence, University of Chinese Academy of Science
wangxiaobo2015cbsr@gmail.com, zlei@nlpr.ia.ac.cn, tmei@jd.com

Abstract. Knowledge distillation is an effective tool to compress large pre-trained Convolutional Neural Networks (CNNs) or their ensembles into models applicable to mobile and embedded devices. The success of which mainly comes from two aspects: the designed student network and the exploited knowledge. However, current methods usually suffer from the low-capability of mobile-level student network and the unsatisfactory knowledge for distillation. In this paper, we propose a novel position-aware exclusivity to encourage large diversity among different filters of the same layer to alleviate the low-capability of student network. Moreover, we investigate the effect of several prevailing knowledge for face recognition distillation and conclude that the knowledge of feature consistency is more flexible and preserves much more information than others. Experiments on a variety of face recognition benchmarks have revealed the superiority of our method over the state-of-the-arts.

Keywords: Face Recognition; Knowledge Distillation; Weight Exclusivity; Feature Consistency

1 Introduction

Convolutional neural networks (CNNs) have gained impressive success in the recent advanced face recognition systems [47, 24, 12, 13, 53, 45, 48, 44]. However, the performance advantages are driven at the cost of training and deploying resource-intensive networks with millions of parameters. As face recognition shifts toward mobile and embedded devices, the computational cost of large CNNs prevents them from being deployed to these devices. It motivates research of developing compact yet still discriminative models. Several directions such as model pruning, model quantization and knowledge distillation have been suggested to make the model smaller and cost-efficient. Among them, knowledge distillation is being actively investigated. The distillation process aims to learn a compact

⋆ Equal contribution
† Corresponding author

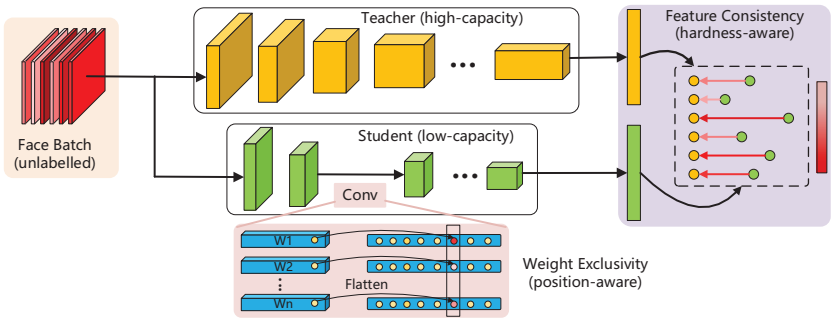

Fig. 1. Overview of our exclusivity-consistency regularized knowledge distillation. The target student network is trained with unlabelled face data by the position-aware weight exclusivity and the hardness-aware feature consistency.

network (student) by utilizing the knowledge of a larger network or its ensemble (teacher) as supervision. Unlike other compression methods, it can downsize a network regardless of the structural difference between teacher and student.

For face recognition model compression, there have been several attempts [16, 40, 51, 26, 19, 52, 18, 35, 9, 10] in literatures to distil large CNNs, so as to make their deployments easier. Hinton *et al.* [16] propose the first knowledge distillation based on the soften probability consistency, where a temperature parameter is introduced in the softmax function to disclose the similarity structure of data. Wang *et al.* [40] use both the soften and one-hot probability consistency knowledge for face recognition and alignment. Luo *et al.* [31] propose a neuron selection method by leveraging the essential characteristics (domain knowledge) of the learned face representation. Karlekar *et al.* [19] simultaneously exploit the one-hot probability consistency and the feature consistency for knowledge transfer between different face resolutions. Yan *et al.* [52] employ the one-hot probability consistency to guide the network training and design a recursive knowledge distillation strategy to relieve the discrepancy between the teacher and student models. Peng *et al.* [35] use the knowledge of probability consistency to transfer not only the instance-level information, but also the correlation between instances. Although current knowledge distillation methods can achieve more promising results than directly training the mobile-level student network, most of them are limited because of the low-capability of pre-defined student network and the inflexible probability consistency knowledge.

In practice, the common dilemma is that we only have a teacher model at hand and do not know how it was trained (including training sets, loss functions and training strategies of teacher *etc*). The task of knowledge distillation is to distil a mobile-level student model from the pre-given teacher. However, as the student network is much more smaller than the teacher, it usually suffer from low-capability for achieving a good performance. Moreover, what kind of knowledge should be used under such dilemma is an open issue. To address

these problems, this paper proposes a novel exclusivity-consistency regularized knowledge distillation namely EC-KD, to simultaneously exploit the weight exclusivity and the feature consistency into one framework for face recognition model compression. Figure 1 gives an illustration of our proposed EC-KD. To sum up, the contributions of this paper can be summarized as follows:

- We propose a novel position-aware exclusivity regularization to encourage large diversity among different filters of the same convolutional layer to alleviate the low-capability of student network.
- We investigate several knowledge for face recognition model distillation and demonstrate that the knowledge of feature consistency is more flexible and powerful than others in face recognition. Moreover, a hardness-aware feature consistency term is developed for fitting the teacher knowledge well.
- We conduct extensive experiments on a variety of face recognition benchmarks, including LFW [17], CALFW [59], CPLFW [60], SLLFW[8], AgeDB [32], CFP [38], RFW [43], MegaFace [20], Trillion-Pairs [6] and IQIYI-VID [30] have verified the superiority of our approach over the state-of-the-arts. Our code is available at `http://www.cbsr.ia.ac.cn/users/xiaobowang/`.

2 Related Work

Knowledge Distillation. Many studies have been conducted since Hinton *et al.* [16] proposed the first knowledge distillation based on the soften class probabilities. Romero *et al.* [36] used the hidden layer response of a teacher network as a hint for a student network to improve knowledge distillation. Zagoruyko and Komodakis *et al.* [56] found the area of activated neurons in a teacher network and transferred the activated area to a student network. Luo *et al.* [31] resorted to the top hidden layer as the knowledge and used the attributes to select the important neurons. Karlekar *et al.* [19] simultaneously exploited one-hot labels and feature vectors for the knowledge transfer between different face resolutions. Heo *et al.* utilized an adversarial attack to discover supporting samples [14] and focused on the transfer of activation boundaries formed by hidden neurons [15] for knowledge distillation. Some studies [23, 2, 40, 4, 1, 54] extended knowledge distillation to other applications.

Deep Face Recognition. Face recognition is an essential open-set metric learning problem, which is different from the closed-set image classification. Specifically, rather than the traditional softmax loss, face recognition is usually supervised by margin-based softmax losses [28, 24, 42, 47, 7, 46], metric learning losses [37] or both [39]. Moreover, the training set used in face recognition is often with larger identities than image classification. To achieve better performance, large CNNs like ResNet [7] or AttentionNet [46] are usually employed, which makes them hard to deploy on mobile and embedded devices. Some works [3, 50] start to design small networks, but the balance between inference time and performance is unsatisfactory, which motivates us to use the knowledge distillation tool for further model compression.

3 Proposed Formulation

3.1 Weight Exclusivity

It is well-known that larger CNNs exhibit higher capability than smaller ones. For face recognition model compression, this phenomenon is more obvious. To this end, we need take some steps to improve the capability of the target student network. In this paper, we try to exploit the diverse information among different filters. To achieve this, several methods [29, 27, 5] have been proposed. However, they are all value-aware criteria and require to normalize the filters (fixed magnitude), which is in contradiction to the weight decay (dynamic magnitude) thus may not address the diversity well. Alternatively, we define a novel position-aware exclusivity. Specifically, assume that all filters in a convolutional layer are a tensor $\mathcal{W} \in \mathbb{R}^{N\times M\times K_1\times K_2}$, where N and M are the numbers of filters and input channels, K_1 and K_2 are the spatial height and width of the filters, respectively. Usually, $K_1 = K_2 = K$. Suppose the tensor $\mathcal{W} \in \mathbb{R}^{N\times M\times K_1\times K_2}$ is reshaped as vectors $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_N]^T \in \mathbb{R}^{N\times D}$, where $D = MK_1K_2$. We define a new measure of diversity, *i.e.*, Exclusivity.

Definition 1. *(**Weight Exclusivity**) Exclusivity between two filter vectors $\boldsymbol{w}_i \in \mathbb{R}^{1\times D}$ and $\boldsymbol{w}_j \in \mathbb{R}^{1\times D}$ is defined as $\mathcal{H}(\boldsymbol{w}_i, \boldsymbol{w}_j) := \|\boldsymbol{w}_i \odot \boldsymbol{w}_j\|_0 = \sum_{k=1}^{D}(\boldsymbol{w}_i(k) \cdot \boldsymbol{w}_j(k) \neq 0)$, where the operator $\odot$ designates the Hadamard product (i.e., element-wise product), and $\|\cdot\|_0$ is the ℓ_0-norm.*

From the definition, we can observe that the exclusivity encourages two filter vectors to be as diverse as possible. Ideally, if the position k of $\boldsymbol{w}_i$ (*i.e.*, $\boldsymbol{w}_i(k)$) is not equal to zero, then the exclusivity term encourages the same position k of $\boldsymbol{w}_j$ (*i.e.*, $\boldsymbol{w}_j(k)$) to be zero. In other words, the same position from different filters are competing to survive and the winner positions are set to large values while the loser ones are set to zeros. Consequently, we can say that the defined exclusivity term is *position-aware*. Compared with *value-aware* regularizations, *e.g.*, the orthonormal regularization [29] to minimize $\|\boldsymbol{W}\boldsymbol{W}^T - \boldsymbol{I}\|_F^2$ and the hyperspherical diversity [27], our *position-aware* exclusivity has the following two advantages. One is that *value-aware* criteria are often based on the normalized weights (*i.e.*, fixed magnitude by setting $||\boldsymbol{w}_i||_2^2 = 1$), which is in contradiction to the weight decay (*i.e.*, dynamic magnitude by regularizing the norm of weights) thus may not address the diversity well in practice. Our *position-aware* exclusivity has no such restriction. The other one is that our weight exclusivity can be seamlessly incorporated into the traditional weight decay (please see the section 3.3). Nevertheless, the non-convexity and discontinuity of ℓ_0-norm make our exclusivity hard to optimize. Fortunately, it is known that ℓ_1-norm is the tightest convex relaxation of ℓ_0-norm [49], thus we have the following relaxed exclusivity.

Definition 2. *(**Relaxed Weight Exclusivity**) Relaxed exclusivity between two filters $\boldsymbol{w}_i \in \mathbb{R}^{1\times D}$ and $\boldsymbol{w}_j \in \mathbb{R}^{1\times D}$ is defined as $\mathcal{H}(\boldsymbol{w}_i, \boldsymbol{w}_j) := \|\boldsymbol{w}_i \odot \boldsymbol{w}_j\|_1 = \sum_{k=1}^{D} |\boldsymbol{w}_i(k)| \cdot |\boldsymbol{w}_j(k)|$, where $|\cdot|$ is the absolute value.*

Consequently, our final **weight exclusivity** is formulated as:

$$\mathcal{L}_{\mathrm{WE}}(\mathcal{W}) := \sum_{1\leq j\neq i\leq N} \|\boldsymbol{w}_i \odot \boldsymbol{w}_j\|_1 = \sum_{1\leq j\neq i\leq N} \sum_{k=1}^{D} |\boldsymbol{w}_i(k)| \cdot |\boldsymbol{w}_j(k)|. \tag{1}$$

3.2 Feature Consistency

In face recognition knowledge distillation, the common dilemma is that we only have a teacher model at hand and do not know how it was trained (including training sets, loss functions and training strategies *etc*). But the task is to obtain a student network with satisfactory performance as well as can be applicable to mobile and embedded devices. As a result, we have the following cases:

One-hot Labels. If the training set of student network is well-labelled, we can directly train the target student network with one-hot labels. Obviously, this manner does not utilize the knowledge of teacher.

Probability Consistency (PC). Let's denote the final softmax output as $\boldsymbol{z}$, the soft label for teacher model T can be defined as $P_{\mathrm{T}}^{\tau} = \mathrm{softmax}(\boldsymbol{z}_{\mathrm{T}}/\tau)$, where τ is the temperature parameter. Similarly, the soft label for student network S is $P_{\mathrm{S}}^{\tau} = \mathrm{softmax}(\boldsymbol{z}_{\mathrm{S}}/\tau)$. Prevailing approaches usually exploit the popular **probability consistency** as follows:

$$\mathcal{L}_{\mathrm{PC}} := \mathcal{L}(P_{\mathrm{T}}^{\tau}, P_{\mathrm{S}}^{\tau}) = \mathcal{L}(\mathrm{softmax}(\boldsymbol{z}_{\mathrm{T}}/\tau), \mathrm{softmax}(\boldsymbol{z}_{\mathrm{S}}/\tau)), \tag{2}$$

where $\mathcal{L}$ is the cross entropy loss between P_{T}^{τ} and P_{S}^{τ}. However, the formulation of PC is inflexible due to the potential discrepancies between teacher and student networks. For example, 1) If the training classes of teacher are different from student's or the teacher model was pre-trained by metric learning losses (*e.g.*, contrastive or triplet losses), $P_{\mathrm{T}}^{\tau} = \mathrm{softmax}(\boldsymbol{z}_{\mathrm{T}}/\tau)$ can not be computed. 2) If the training set of student network contains noisy labels, the performance is not guaranteed because of the unreliable $P_{\mathrm{S}}^{\tau} = \mathrm{softmax}(\boldsymbol{z}_{\mathrm{S}}/\tau)$. To sum up, all these point to: the probability consistency knowledge is not flexible and powerful for face recognition.

Feature Consistency (FC). In face recognition, we can also use the feature layer as hint to train the student network. The **feature consistency** can be formulated as follows:

$$\mathcal{L}_{\mathrm{FC}} := \mathcal{H}(F_{\mathrm{S}}, F_{\mathrm{T}}) = ||F_{\mathrm{S}} - F_{\mathrm{T}}||, \tag{3}$$

where $\mathcal{H}$ is the L2 loss, F_{S} and F_{T} are the features from student and teacher. From the formulation, it can be concluded that FC is flexible for training because it is not restricted by the discrepancies between the unknown teacher and the target student. Moreover, to make full use of feature consistency knowledge, we further develop a hardness-aware one. Intuitively, for face samples that are far away from their teachers, they should be emphasized. As a consequence, we define a re-weighted softmax function, $s_i = \frac{e^{\mathcal{H}_i}}{\sum_{j=1}^{m} e^{\mathcal{H}_j}}$, where m is the batch size and our **hardness-aware feature consistency** is simply formulated as:

$$\mathcal{L}_{\mathrm{HFC}} := (1 + s_i)\mathcal{H}(F_{\mathrm{S}}, F_{\mathrm{T}}). \tag{4}$$

Algorithm 1: Exclusivity-Consistency Regularized Knowledge Distillation (EC-KD)

Input: Unlabelled training data; Pre-trained teacher model
;
Initialization: $e = 1$; Randomly initialize $\boldsymbol{\Theta}^{\mathrm{S}}$;
while $e \leq E$ **do**
  Shuffle the unlabelled training set $\mathcal{S}$ and fetch mini-batch $\mathcal{S}_m$;
  Forward: Compute the re-weighted matrix $\boldsymbol{G}$ and the final loss (Eq. (5));
  Backward: Compute the gradient $\frac{\partial \mathcal{L}_{\mathrm{EC-KD}}}{\partial \boldsymbol{W}} = \frac{\partial \mathcal{L}_{\mathrm{HFC}}}{\partial \boldsymbol{W}} + \lambda_1(\boldsymbol{G} \odot \boldsymbol{W})$ and Update the student model $\boldsymbol{\Theta}^{\mathrm{S}}$ by SGD.
end
Output: Target student model $\boldsymbol{\Theta}^{\mathrm{S}}$.

3.3 Exclusivity-Consistency Regularized Knowledge Distillation

Based on the above analysis, we prefer to simultaneously harness the weight exclusivity and the feature consistency, *i.e.*, Eqs. (1) and (4), together with the weight decay result in our final Exclusivity-Consistency Regularized Knowledge Distillation (EC-KD):

$$\mathcal{L}_{\mathrm{EC-KD}} = \mathcal{L}_{\mathrm{HFC}} + \lambda_1 \underbrace{\|\boldsymbol{W}\|_F^2}_{\text{weight decay}} + \lambda_2 \underbrace{\sum_{1 \leq j \neq i \leq N} \sum_{k=1}^{D} |\boldsymbol{w}_i(k)| \cdot |\boldsymbol{w}_j(k)|}_{\text{weight exclusivity}} \tag{5}$$

where λ_1 and λ_2 are the trade-off parameters. Typically, weight decay is a unary cost to regularize the norm of filters, while weight exclusivity is a pairwise cost to promote the direction of the filters. Therefore, they are complementary to each other. For simplicity, we empirically set $\lambda_2 = 2\lambda_1$. In consequence, the weight decay and the weight exclusivity can be seamlessly formulated as $\Phi(\boldsymbol{W}) :=$

$$\|\boldsymbol{W}\|_F^2 + 2 \sum_{1 \leq j \neq i \leq N} \sum_{k=1}^{D} |\boldsymbol{w}_i(k)| \cdot |\boldsymbol{w}_j(k)| = \sum_{k=1}^{D} \left(\sum_{i=1}^{N} |w_i(k)| \right)^2 = \|\boldsymbol{W}\|_{1,2}^2 \tag{6}$$

As observed from Eq. (6), it can be split into a set of smaller problems. For each column $\boldsymbol{w}_{\cdot k}$ of $\boldsymbol{W}$, according to [21], the gradient of $\Phi(\boldsymbol{W})$ with respect to $\boldsymbol{w}_{\cdot k}$ is computed as follows:

$$\frac{\partial \Phi(\boldsymbol{W})}{\partial \boldsymbol{w}_{\cdot k}} = \mathbf{g}_{\cdot k} \odot \boldsymbol{w}_{\cdot k} = \left[\frac{\|\boldsymbol{w}_{\cdot k}\|_1}{|\boldsymbol{w}_{\cdot k}(1)| + \epsilon}, \ldots, \frac{\|\boldsymbol{w}_{\cdot k}\|_1}{|\boldsymbol{w}_{\cdot k}(N)| + \epsilon} \right]^T \odot \boldsymbol{w}_{\cdot k}, \tag{7}$$

where $\epsilon \to 0^+$ (a small constant) is introduced to avoid zero denominators. Since both $\boldsymbol{g}_{\cdot k}$ and $\boldsymbol{w}_{\cdot k}$ depend on $\boldsymbol{w}_{\cdot k}$, we employ an efficient re-weighted algorithm to iteratively update $\boldsymbol{g}_{\cdot k}$ and $\boldsymbol{w}_{\cdot k}$. In each iteration, for the forward, we compute the re-weighted matrix $\boldsymbol{G} = [\boldsymbol{g}_{\cdot 1}, \ldots, \boldsymbol{g}_{\cdot D}] \in \mathbb{R}^{N \times D}$, while for the backward, we update the weight by using the gradient $\frac{\partial \Phi(\boldsymbol{W})}{\partial \boldsymbol{W}} = \boldsymbol{G} \odot \boldsymbol{W}$. For clarity, the whole scheme of our framework is summarized in Algorithm 1.

Table 1. Face datasets for training and test. (P) and (G) refer to the probe and gallery set, respectively. (V) refers to video clips.

	Datasets	#Identities	Images
Training	CASIA-WebFace-R [55]	9,809	0.39M
Test	LFW [17]	5,749	13,233
	CALFW [59]	5,749	12,174
	CPLFW [60]	5,749	11,652
	SLLFW [8]	5,749	13,233
	AgeDB [32]	568	16,488
	CFP [38]	500	7,000
	RFW [43]	11,430	40,607
	MegaFace [20]	530(P)	1M(G)
	Trillion-Pairs [6]	5,749(P)	1.58M(G)
	IQIYI-VID [30]	4,934	565,372(V)

4 Experiments

4.1 Datasets

Training Set. We use CASIA-WebFace [55] as the training set of student models unless otherwise specified. Specifically, we use the publicly available one[5].
Test Set. We use ten face recognition benchmarks, including LFW [17], CALFW [59], CPLFW [60], SLLFW[8], AgeDB [32], CFP [38], RFW[43], MegaFace [20, 33], Trillion-Pairs [6] and IQIYI-VID [30] as the test sets. LFW contains 13,233 web-collected images from 5,749 different identities. CALFW [59] was collected by crowdsourcing efforts to seek the pictures of people in LFW with age gap as large as possible on the Internet. CPLFW [60] is similar to CALFW, but from the perspective of pose difference. SLLFW [8] selects 3,000 similar-looking negative face pairs from the original LFW image collection. AgeDB [32] contains images annotated with accurate to the year, noise-free labels. CFP [38] consists of collected images of celebrities in frontal and profile views. RFW [43] is a benchmark for measuring racial bias, which consists of four test subsets, namely Caucasian, Indian, Asian and African. MegaFace [33] aims at evaluating the face recognition performance at the million scale of distractors. Trillion-Pairs [6] is a benchmark for testing the face recognition performance with trillion scale of distractors. IQIYI-VID [30] contains multiple video clips from IQIYI variety shows, films and television dramas.
Dataset Overlap Removal. In face recognition, it is very important to perform open-set evaluation [28, 42, 7], *i.e.*, there should be no overlapping identities between training set and test set. To this end, we need to carefully remove the overlapped identities between the employed training set and the test sets. For the overlap identities removal tool, we use the publicly available script provided by [42] to check whether if two names are of the same person. As a result, we

[5] `https://github.com/ZhaoJ9014/face.evoLVe.PyTorch`

remove 766 identities from the training set of the CASIA-WebFace. For clarity, we denote the refined training set as CASIA-WebFace-R. Important statistics of the datasets are summarized in Table 1. To be rigorous, all the experiments in this paper are based on the refined training dataset.

4.2 Experimental Settings

Data Processing. We detect the faces by adopting the FaceBoxes detector [58, 57] and localize five landmarks (two eyes, nose tip and two mouth corners) through a simple 6-layer CNN [11]. The detected faces are cropped and resized to 144×144, and each pixel (ranged between [0,255]) in RGB images is normalized by subtracting 127.5 and divided by 128. For all the training faces, they are horizontally flipped with probability 0.5 for data augmentation.
Pre-trained Teacher. There are many kinds of network architectures [28, 3, 41] and several loss functions [7, 46] for face recognition. Without loss of generality, we use SEResNet50-IR [7] as the teacher model, which was trained by SV-AM-Softmax loss [46]. For all the experiments in this paper, the teacher is pre-given and frozen. Here we provide the details of teacher to the competitors KD [16], FitNet [36], AB[15], BBS [14] and ONE [22].
Student. we use MobileFaceNet [3] and its variants as the student model. The feature dimension of student is 512.
Training. All the student models are trained from scratch, with the batch size of 256 on 4 P40 GPUs parallelly. All experiments in this paper are implemented by PyTorch [34]. The weight decay λ_1 is empirically set to 0.0005 and the momentum is 0.9. The learning rate is initially 0.1 and divided by 10 at the 9, 18, 26 epochs, and we finish the training process at 30 epochs.
Test. We use the learned student network to extract face features. For the evaluation metric, cosine similarity is utilized. We follow the unrestricted with labeled outside data protocol [17] to report the performance on LFW, CALFW, CPLFW, SLLFW, AgeDB, CFP and RFW. Moreover, we also report the BLUFR (TPR@FAR=1e-4) protocol [25] on LFW. On Megaface and Trillion-Pairs, both face identification and verification are conducted by ranking and thresholding the scores. Specifically, for face identification (Id.), the Cumulative Matching Characteristics (CMC) curves are adopted to evaluate the Rank-1 accuracy. For face verification (Veri.), the Receiver Operating Characteristic (ROC) curves at different false alarm rates are adopted. On IQIYI-VID, the MAP@100 is adopted as the evaluation indicator. MAP (Mean Average Precision) refers to the average accuracy rate of the videos of person ID retrieved in the test set for each person ID (as the query) in the training set.

4.3 Ablation Study and Exploratory Experiments

Feature consistency vs. other knowledge. In this part, we use MobileFaceNet [3] as the student network. For the adopted knowledge, we compare the soften probability consistency (PC) (*i.e.*, Eq .(2)), feature consistency (FC) (*i.e.*, Eq .(3)) and their combinations with the softmax-based loss [46]. The results in

Table 2. Performance(%) comparison of different knowledge.

	LFW	BLUFR	MF-Id.	MF-Veri.
SEResNet50-IR(T)	99.21	96.41	86.14	88.32
MobileFaceNet(S)	99.08	93.65	80.27	85.20
PC [Eq.(2)]	98.48	84.17	69.33	68.07
FC [Eq.(3)]	99.11	95.57	83.96	86.56
PC + Loss [46]	99.01	93.71	81.74	85.48
FC + Loss [46]	99.15	94.29	81.90	84.72
HFC [Eq.(4)]	**99.20**	**95.59**	**84.19**	**87.86**

Table 2 show that simply using the knowledge of soften probability consistency is not enough. It should combine with the softmax-based loss (one-hot labels) to achieve satisfactory performance. While the simple knowledge of feature consistency can achieve higher performance, which reveals that it preserves much more information than probability consistency. We also observe that the improvement by combining the knowledge of feature consistency with the softmax-based loss is limited. Moreover, from the results of our hardness-aware feature consistency (HFC), we can see that the feature consistency knowledge should be emphasized.

Table 3. Comparison of different filter numbers.

#Filters	Model Size	Flops	Infer Time	LFW	MF-Id.	MF-Veri.
(2×)128	16MB	1.44G	158ms	99.34	87.19	90.82
(Orig.)64	4.8MB	0.38G	84ms	99.11	83.96	87.57
(1/2)32	1.7MB	0.11G	49ms	98.55	74.32	78.71
(1/4)16	648KB	0.03G	34ms	97.60	52.60	58.69
(1/8)8	304KB	0.01G	28ms	94.29	25.32	27.04

Effect of filter numbers. We further evaluate the feature consistency knowledge with different number of filters. Specifically, we change the filter numbers of all convolutional layers from the student network (*i.e.*, MobileFaceNet) to 2, 1/2, 1/4 and 1/8 times size. The performance is reported in Table 3, from which we can conclude that smaller networks usually exhibit lower capability for face recognition. To achieve a good balance between different factors, we employ the (1/2)MobileFaceNet as the student network in the following experiments unless otherwise specified.

Position-aware exclusivity vs. value-aware regularizations. Promoting orthogonality (Orth.) [29] or minimum hyperspherical energy (MHE) [27] among filters has been a popular choice for encouraging diversity. However, their assumption of normalized weights is limited and usually lead to very slow convergence. From the values in Table 4, we can see that the weight decay (*e.g.*,

Table 4. Performance(%) comparison of different diversities.

	LFW	BLUFR	MF-Id.	MF-Veri.
SEResNet50-IR(T)	99.21	96.41	86.14	88.32
(1/2)MobileFaceNet(S)	98.64	89.32	69.54	74.93
$FC_{wd=0}$[Eq.(3)]	97.43	64.26	51.65	24.50
$FC_{wd=5e-4}$[Eq.(3)]	98.55	90.29	74.32	78.71
FC+Orth. [29]	98.30	86.19	67.80	73.86
FC+MHE [27]	**98.86**	90.44	74.58	79.39
FC+Exclusivity	98.63	**91.07**	**75.29**	**79.56**

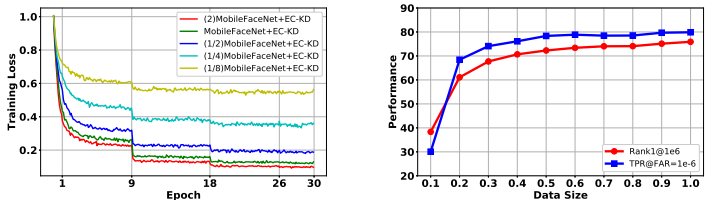

Fig. 2. Left: Convergence of EC-KD. **Right**: Effect of different data size.

wd=5e-4 vs. wd=0) to control the norm of filters is not negligible. Moreover, the diversity to determine the direction of filters is also important. The experiments show that our position-aware exclusivity can achieve better diversity and result in higher performance than previous value-aware regularizations.

Convergence. For our EC-KD method, we alternatively update the re-weighted matrix $\boldsymbol{G}$ and the filters $\boldsymbol{W}$. Specifically, we compute the re-weighted matrix $\boldsymbol{G}$ in the forward and update the filters $\boldsymbol{W}$ in the backward, which is different from the standard stochastic gradient descent (SGD) algorithm. Although the convergence of our method is not easy to be theoretically analyzed, it would be intuitive to see its empirical behavior. Here, we give the loss changes as the number of epochs increases. From the curves in the left of Figure 2, it can be observed that our method has a good behavior of convergence under various models. Moreover, we can also observe that smaller networks suffer from low-capability and their loss values are generally higher than the larger ones.

Effect of data size. Since the training set of student network can be unlabelled, we evaluate the performance gain from using different percentages of unlabelled training data in our EC-KD algorithm. Various percentages (from 10% to 100%, with step 10%) of unlabelled data are randomly emerged for training the target student network. As can be seen from the curves in the right of Figure 2, at the beginning, the performance on MegaFace goes up as the amount of unlabelled data increases, but the improvement is minor when the percentage of data is large enough. So we conclude that our method can benefit from the small scale training sets and thus can reduce the training time.

Table 5. Performance(%) comparison of different methods with different noise rates.

	LFW	BLUFR	MF-Id.	MF-Veri.
SEResNet50-IR(T)	99.21	96.41	86.14	88.32
(1/2)MobileFaceNet(S)	98.64	89.32	69.54	74.93
KD [16] (symmetric=0.1)	98.30	77.77	64.09	41.42
EC-KD (symmetric=0.1)	**98.93**	**91.11**	**75.94**	**79.87**
KD [16] (symmetric=0.2)	94.98	40.77	30.82	5.83
EC-KD (symmetric=0.2)	**98.91**	**91.87**	**75.78**	**80.45**
KD [16] (symmetric=0.3)	93.56	27.18	25.53	2.15
EC-KD (symmetric=0.3)	**98.89**	**91.03**	**75.98**	**79.64**

Effect of noisy labels. To validate the robustness of our method under the case that the training set of student network contains noisy labels, in this experiments, we use the training set CASIA-WebFace-R with different synthetic noise rates to train the student network. The symmetric noise is generated by randomly selecting a label with equal probabilities among all the classes [45]. From the values in Table 5, we can see that the probability consistency method KD [16] is very sensitive to noise rates. With the increase of noise rates, its performance decreases sharply. While our method can guarantee the performance regardless of the noise rates. The reason behind this is that most of current knowledge distillation methods depend on well-labelled training set because of the knowledge of probability consistency. Our method EC-KD resorts to feature consistency and dose not require labels. As a consequence, our method is insensitive to the noisy labels existing in the training set.

Table 6. Performance(%) comparison of different methods by using new training sets.

Training set	Method	LFW	BLUFR	MF-Id.	MF-Veri.
CASIA-WebFace-R	SEResNet50-IR(T)	99.21	96.41	86.14	88.32
IQIYI-VID-Training	(1/2)MobileFaceNet(S)	86.08	66.68	29.62	26.90
	EC-KD (**Ours**)	**98.41**	**85.18**	**64.27**	**67.47**

Generalization Capability. In practice, it is hard to know how the teacher network was pre-trained. More frequently, we only have the teacher model at hand. In this case, we may face the situation that the training set of student is different from teacher's. For example, the teacher is pre-trained by CASIA-WebFace-R dataset but we can only get a new dataset (*e.g.*, IQIYI-VID-Training [30]) to train the target student network. As shown in Table 6, it can be seen that directly training the student network from scratch (*i.e*, (1/2)MobileFaceNet) is hard to boost the performance. Current knowledge distillation methods like KD [16] are unable to train the student network because the training classes of

teacher and student are different. In contrast, our method EC-KD can not only be used for training the student network, but also be used to effectively transfer the useful knowledge and achieve higher performance.

Table 7. Performance (%) of different knowledge distillation methods on LFW, CALFW, CPLFW, SLLFW, AgeDB and CFP.

Method	LFW	CALFW	CPLFW	SLLFW	AgeDB	CFP	**Avg.**
SEResNet50-IR	99.21	90.78	84.06	97.76	93.98	93.38	93.23
(1/2)MobileFaceNet	98.64	87.79	78.03	94.39	89.91	86.52	89.21
KD [16]	98.81	89.35	76.38	95.13	90.95	85.11	89.28
FitNet [36]	**99.06**	89.33	77.51	95.41	91.21	87.01	89.92
Selection [31]	98.66	89.06	79.83	95.15	91.50	89.30	90.58
AB [15]	97.54	85.93	74.30	92.78	92.06	89.95	88.76
BSS [14]	98.98	89.18	77.28	95.51	91.78	85.14	89.64
ONE [22]	98.41	88.36	78.06	94.21	89.73	86.08	89.14
EC-KD (**Ours**)	98.96	**89.39**	**80.98**	**95.58**	**92.33**	**90.20**	**91.22**

4.4 Comparison to State-of-the-art Methods

Results on LFW, CALFW, CPLFW, SLLFW, AgeDB and CFP Table 7 shows the results of different approaches on LFW [17], CALFW [59], CPLFW [60], SLLFW [8], AgeDB [32] and CFP [38] sets. The bold values in each column represent the best result. From the numbers, we observe that most of the knowledge distillation methods are better than simply training the student network from scratch (*i.e.*, (1/2)MobileFaceNet). Among all the competitors, the Selection [31], AB [15] and BSS [14] seem to show better generalization ability than others. For our method, we boost about 2% average improvement over the baseline (1/2)MobileFaceNet. Although we cannot beat the competitors on each test set, we achieve the best average (Avg.) performance on these six test sets than the best competitor Selection [31] because of our position-aware weight exclusivity and hardness-aware feature consistency.

Results on RFW Table 8 displays the performance comparison of all the methods on the RFW test set. The results exhibit the same trends that emerged on previous test sets. Concretely, most of the knowledge distillation methods are consistently better than directly training the student network from scratch. For instance, the classical Knowledge Distillation (KD) [16] boost about 2.3% average performance than the baseline (1/2)MobileFaceNet. While for our proposed EC-KD, it can further boost the performance. Specifically, it achieves about 1.5% average improvement on the four subsets of RFW than the KD [16]. From the experiments, we can conclude the effectiveness of our weight exclusivity and hardness-aware feature consistency knowledge.

Table 8. Performance (%) of different methods on RFW.

Method	Caucasian	Indian	Asian	African	**Avg.**
SEResNet50-IR(T)	92.66	88.50	84.00	83.50	87.16
(1/2)MobileFaceNet(S)	84.16	78.33	80.83	77.66	80.24
KD [16]	88.16	80.66	81.16	**80.33**	82.57
FitNet [36]	87.50	82.16	81.83	78.00	82.37
Selection [31]	88.83	79.83	78.83	77.50	81.24
AB [15]	83.33	75.83	74.16	71.16	76.12
BSS [14]	89.16	79.66	81.83	77.16	81.95
ONE [22]	86.83	77.66	80.33	78.33	80.78
EC-KD (**Ours**)	**91.33**	**82.83**	**82.83**	79.49	**84.12**

Results on MegaFace and Trillion-Pairs Table 9 gives the identification and verification results on MegaFace [33] and Trillion-Pairs challenge. In particular, compared with the baseline, *i.e.*, (1/2)MobileFaceNet, most of the competitors have shown their strong abilities to fit the teacher knowledge and usually achieve better performance on MegaFace. While on Trillion-Pairs, these methods usually fail. The reason may be that the knowledge of probability consistency is hard to preserve at the small rank and very low false alarm rate, due to its high-dimensionality. For feature consistency knowledge, it has been proved that it can preserve much more information than probability consistency knowledge. Moreover, with the weight exclusivity to exploit the diverse information among different filters, our EC-KD can keep the performance, even at the small rank or very low false alarm rate (*e.g.* 0.54% for Id. rate and 0.55 % for Veri. rate over the best competitor AB [15] on Trillion-Pairs).

Table 9. Performance (%) of different methods on MegaFace and Trillion-Pairs.

Method	MF-Id.	MF-Veri.	TP-Id.	TP-Veri.	**Avg.**
SEResNet50-IR(T)	86.14	88.32	33.08	32.09	59.90
(1/2)MobileFaceNet(S)	69.54	74.93	16.57	4.77	41.45
KD [16]	69.86	71.67	3.60	1.08	36.55
FitNet [36]	71.74	74.12	5.96	1.24	38.26
Selection [31]	74.50	79.08	16.31	15.65	46.38
AB [15]	**77.42**	75.24	17.13	16.41	46.55
BSS [14]	71.56	73.72	5.25	1.08	37.90
ONE [22]	68.63	74.07	16.60	11.34	42.66
EC-KD (**Ours**)	75.89	**79.87**	**17.67**	**16.96**	**47.59**

Results on IQIYI-VID Table 10 shows the mean of the average accuracy rate (MAP) of different methods on IQIYI-VID test set. The performance is not

Table 10. Performance (%) of different methods on IQIYI-VID test set.

Training Set	Method	MAP(%)
CASIA-WebFace-R	SEResNet50-IR(T)	45.83
	(1/2)MobileFaceNet(S)	26.43
CASIA-WebFace-R	KD [16]	21.26
	FitNet [36]	21.74
	Selection [31]	32.76
	AB [15]	23.51
	BSS [14]	21.37
	ONE [22]	25.45
	EC-KD (**Ours**)	33.28
IQIYI-VID-Training [30]	EC-KD (**Ours**)	**38.03**

high because of the domain gap between the training set CASIA-WebFace-R and the test set IQIYI-VID. Specifically, directly training the student network from scratch can only reach MAP=26.43%. Moreover, we empirically find that most of the competitors degrade the performance in this case. While for our method, despite the large domain gap between training and test sets, we still keep the useful knowledge and can achieve higher performance (about 7% improvement over the baseline (1/2)MobileFaceNet). These experiments have also shown the generalization capability of our method. Moreover, if we can collect the training set that are similar to the test set (without needing the labeled training data), we can further improve the performance (38.03% vs. 33.28%).

5 Conclusion

In this paper, we have exploited a new measurement of diversity, *i.e.*, exclusivity, to improve the low-capability of student network. Different from the weight decay, which is a unary cost to regularize the norm of filters, our weight exclusivity is a pairwise cost to promote the direction of the filters. Therefore, these two branches are complementary to each other. Moreover, we have demonstrated that the feature consistency knowledge is more flexible and preserves much more information than others for face recognition distillation. Incorporating the weight exclusivity and the hardness-aware feature consistency together gives birth to a new knowledge distillation, namely EC-KD. Extensive experiments on a variety of face recognition benchmarks have validated the effectiveness and generalization capability of our method.

6 Acknowledgement

This work was supported in part by the National Key Research & Development Program (No. 2020YFC2003901), Chinese National Natural Science Foundation Projects #61872367 and partially supported by Beijing Academy of Artificial Intelligence (BAAI).

References

1. Aguinaldo, A., Chiang, P.Y., Gain, A., Patil, A., Pearson, K., Feizi, S.: Compressing gans using knowledge distillation. arXiv preprint arXiv:1902.00159 (2019)
2. Chen, G., Choi, W., Yu, X., Han, T., Chandraker, M.: Learning efficient object detection models with knowledge distillation. In: Advances in Neural Information Processing Systems. pp. 742–751 (2017)
3. Chen, S., Liu, Y., Gao, X., Han, Z.: Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In: Chinese Conference on Biometric Recognition. pp. 428–438. Springer (2018)
4. Chen, Y., Wang, N., Zhang, Z.: Darkrank: Accelerating deep metric learning via cross sample similarities transfer. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
5. Cogswell, M., Ahmed, F., Girshick, R., Zitnick, L., Batra, D.: Reducing overfitting in deep networks by decorrelating representations. arXiv preprint arXiv:1511.06068 (2015)
6. Deepglint: `http://trillionpairs.deepglint.com/overview` (2018)
7. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
8. Deng, W., Hu, J., Zhang, N., Chen, B., Guo, J.: Fine-grained face verification: Fglfw database, baselines, and human-dcmn partnership. Pattern Recognition **66**, 63–73 (2017)
9. Duong, C.N., Luu, K., Quach, K.G., Le, N.: Shrinkteanet: Million-scale lightweight face recognition via shrinking teacher-student networks. arXiv preprint arXiv:1905.10620 (2019)
10. Feng, Y., Wang, H., Hu, R., Yi, D.T.: Triplet distillation for deep face recognition. arXiv preprint arXiv:1905.04457 (2019)
11. Feng, Z.H., Kittler, J., Awais, M., Huber, P., Wu, X.J.: Wing loss for robust facial landmark localisation with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2235–2245 (2018)
12. Guo, J., Zhu, X., Lei, Z., Li, S.Z.: Face synthesis for eyeglass-robust face recognition. In: Chinese Conference on Biometric Recognition. pp. 275–284. Springer (2018)
13. Guo, J., Zhu, X., Zhao, C., Cao, D., Lei, Z., Li, S.Z.: Learning meta face recognition in unseen domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6163–6172 (2020)
14. Heo, B., Lee, M., Yun, S., Choi, J.Y.: Knowledge distillation with adversarial samples supporting decision boundary. arXiv preprint arXiv:1805.05532 **3** (2018)
15. Heo, B., Lee, M., Yun, S., Choi, J.Y.: Knowledge transfer via distillation of activation boundaries formed by hidden neurons. arXiv preprint arXiv:1811.03233 (2018)
16. Hinton, G., Vinyals, O., Dean., J.: Distilling the knowledge in a neural network. In: arXiv preprint arXiv:1503.02531 (2015)
17. Huang, G., Ramesh, M., Miller., E.: Labeled faces in the wild: A database for studying face recognition in unconstrained enviroments. Technical Report (2007)
18. Jin, X., Peng, B., Wu, Y., Liu, Y., Liu, J., Liang, D., Yan, J., Hu, X.: Knowledge distillation via route constrained optimization. arXiv preprint arXiv:1904.09149 (2019)

19. Karlekar, J., Feng, J., Wong, Z.S., Pranata, S.: Deep face recognition model compression via knowledge transfer and distillation. arXiv preprint arXiv:1906.00619 (2019)
20. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4873–4882 (2016)
21. Kong, D., Fujimaki, R., Liu, J., Nie, F., Ding, C.: Exclusive feature learning on arbitrary structures via $\ell_{-}\{1,2\}$-norm. In: Advances in Neural Information Processing Systems. pp. 1655–1663 (2014)
22. Lan, X., Zhu, X., Gong, S., Lan, X.: Knowledge distillation by on-the-fly native ensemble. arXiv preprint arXiv:1806.04606 (2018)
23. Li, Q., Jin, S., Yan, J.: Mimicking very efficient network for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6356–6364 (2017)
24. Liang, X., Wang, X., Lei, Z., Liao, S., Li, S.Z.: Soft-margin softmax for deep classification. In: International Conference on Neural Information Processing. pp. 413–421. Springer (2017)
25. Liao, S., Lei, Z., Yi, D., Li, S.Z.: A benchmark study of large-scale unconstrained face recognition. In: International Conference on Biometrics (2014)
26. Lin, R., Liu, W., Liu, Z., Feng, C., Yu, Z., Rehg, J.M., Xiong, L., Song, L.: Regularizing neural networks via minimizing hyperspherical energy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6917–6927 (2020)
27. Liu, W., Lin, R., Liu, Z., Liu, L., Yu, Z., Dai, B., Song, L.: Learning towards minimum hyperspherical energy. In: Advances in Neural Information Processing Systems. pp. 6222–6233 (2018)
28. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 212–220 (2017)
29. Liu, W., Zhang, Y.M., Li, X., Yu, Z., Dai, B., Zhao, T., Song, L.: Deep hyperspherical learning. In: Advances in neural information processing systems. pp. 3950–3960 (2017)
30. Liu, Y., Shi, P., Peng, B., Yan, H., Zhou, Y., Han, B., Zheng, Y., Lin, C., Jiang, J., Fan, Y., et al.: iqiyi-vid: A large dataset for multi-modal person identification. arXiv preprint arXiv:1811.07548 (2018)
31. Luo, P., Zhu, Z., Liu, Z., Wang, X., Tang, X.: Face model compression by distilling knowledge from neurons. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
32. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: the first manually collected, in-the-wild age database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 51–59 (2017)
33. Nech, A., Kemelmacher-Shlizerman, I.: Level playing field for million scale face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7044–7053 (2017)
34. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
35. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Correlation congruence for knowledge distillation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5007–5016 (2019)

36. Romero, A., Ballas, N., Kahou., S.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
37. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
38. Sengupta, S., Chen, J.C., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–9. IEEE (2016)
39. Sun, Y., Wang, X., Tang, X.: Deeply learned face representations are sparse, selective, and robust. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2892–2900 (2015)
40. Wang, C., Lan, X., Zhang, Y.: Model distillation with knowledge transfer from face classification to alignment and verification. arXiv preprint arXiv:1709.02929 (2017)
41. Wang, F., Chen, L., Li, C., Huang, S., Chen, Y., Qian, C., Change Loy, C.: The devil of face recognition is in the noise. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 765–780 (2018)
42. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Processing Letters **25**(7), 926–930 (2018)
43. Wang, M., Deng, W., Hu, J., Peng, J., Tao, X., Huang, Y.: Racial faces in-the-wild: Reducing racial bias by deep unsupervised domain adaptation. arXiv:1812.00194 (2018)
44. Wang, X., Shuo, W., Cheng, C., Shifeng, Z., Tao, M.: Loss function search for face recognition. In: Proceedings of the 37-th International Conference on Machine Learning (2020)
45. Wang, X., Wang, S., Wang, J., Shi, H., Mei, T.: Co-mining: Deep face recognition with noisy labels. In: Proceedings of the IEEE international conference on computer vision. pp. 9358–9367 (2019)
46. Wang, X., Wang, S., Zhang, S., Fu, T., Shi, H., Mei, T.: Support vector guided softmax loss for face recognition. arXiv:1812.11317 (2018)
47. Wang, X., Zhang, S., Lei, Z., Liu, S., Guo, X., Li, S.Z.: Ensemble soft-margin softmax loss for image classification. arXiv preprint arXiv:1805.03922 (2018)
48. Wang, X., Zhang, S., Wang, S., Fu, T., Shi, H., Mei, T.: Mis-classified vector guided softmax loss for face recognition. arXiv preprint arXiv:1912.00833 (2019)
49. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE transactions on pattern analysis and machine intelligence **31**(2), 210–227 (2008)
50. Wu, X., He, R., Sun, Z., Tan, T.: A light cnn for deep face representation with noisy labels. IEEE Transactions on Information Forensics and Security **13**(11), 2884–2896 (2018)
51. Xie, P., Singh, A., Xing, E.P.: Uncorrelation and evenness: a new diversity-promoting regularizer. In: International Conference on Machine Learning. pp. 3811–3820 (2017)
52. Yan, M., Zhao, M., Xu, Z., Zhang, Q., Wang, G., Su, Z.: Vargfacenet: An efficient variable group convolutional neural network for lightweight face recognition. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 0–0 (2019)
53. Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring visual relationship for image captioning. In: Proceedings of the European conference on computer vision (ECCV). pp. 684–699 (2018)

54. Yao, T., Pan, Y., Li, Y., Qiu, Z., Mei, T.: Boosting image captioning with attributes. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4894–4902 (2017)
55. Yi, D., Lei, Z., Liao, S., Li., S.Z.: Learning face representation from scratch. arXiv:1411.7923. (2014)
56. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016)
57. Zhang, S., Wang, X., Lei, Z., Li, S.Z.: Faceboxes: A cpu real-time and accurate unconstrained face detector. Neurocomputing (2019)
58. Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., Li, S.Z.: Faceboxes: A cpu real-time face detector with high accuracy. In: 2017 IEEE International Joint Conference on Biometrics (IJCB). pp. 1–9. IEEE (2017)
59. Zheng, T., Deng, W., Hu, J., Hu, J.: Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. arXiv:1708.08197 (2017)
60. Zheng, T., Deng, W., Zheng, T., Deng, W.: Cross-pose lfw: A database for studying crosspose face recognition in unconstrained environments. Tech. Rep (2018)