Sub-center ArcFace: Boosting Face Recognition by Large-scale Noisy Web Faces

Jiankang Deng [*1] Jia Guo [*2] Tongliang Liu[3]
Mingming Gong [4] Stefanos Zafeiriou[1]

[1]Imperial College [2]InsightFace [3]University of Sydney [4]University of Melbourne
{j.deng16, s.zafeiriou}@imperial.ac.uk, guojia@gmail.com
tongliang.liu@sydney.edu.au, mingming.gong@unimelb.edu.au

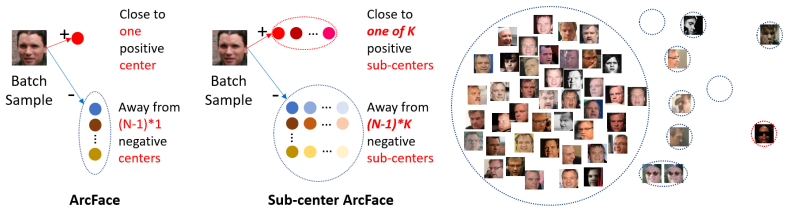
Abstract. Margin-based deep face recognition methods (e.g. SphereFace, CosFace, and ArcFace) have achieved remarkable success in unconstrained face recognition. However, these methods are susceptible to the massive label noise in the training data and thus require laborious human effort to clean the datasets. In this paper, we relax the intra-class constraint of ArcFace to improve the robustness to label noise. More specifically, we design K sub-centers for each class and the training sample only needs to be close to any of the K positive sub-centers instead of the only one positive center. The proposed sub-center ArcFace encourages one dominant sub-class that contains the majority of clean faces and non-dominant sub-classes that include hard or noisy faces. Extensive experiments confirm the robustness of sub-center ArcFace under massive real-world noise. After the model achieves enough discriminative power, we directly drop non-dominant sub-centers and high-confident noisy samples, which helps recapture intra-compactness, decrease the influence from noise, and achieve comparable performance compared to ArcFace trained on the manually cleaned dataset. By taking advantage of the large-scale raw web faces (Celeb500K), sub-center Arcface achieves state-of-the-art performance on IJB-B, IJB-C, MegaFace, and FRVT.

Keywords: Face Recognition, Sub-class, Sub-center, Large-scale, Noisy Data

1 Introduction

Face representation using Deep Convolutional Neural Network (DCNN) embedding with margin penalty [26,15,32,5] to simultaneously achieve intra-class compactness and inter-class discrepancy is the method of choice for state-of-the-art face recognition. To avoid the sampling problem in the Triplet loss [26], margin-based softmax methods [15,32,31,5] focused on incorporating margin penalty into a more feasible framework, the softmax loss, which has global sample-to-class comparisons within the multiplication step between the embedding feature and the linear transformation matrix. Naturally, each column of the linear transformation matrix is viewed as a class center representing a certain class [5].

[*] Equal contributions.
InsightFace is a nonprofit Github project for 2D and 3D face analysis.



(a) ArcFace vs. Sub-center ArcFace (b) Example of Sub-classes

Fig. 1. (a) Difference between ArcFace and the proposed sub-center ArcFace. In this paper, we introduce sub-class into ArcFace to relax the intra-class constraint, which can effectively improve robustness under noise. (b) The sub-classes of one identity from the CASIA dataset [40] after using the sub-center ArcFace loss ($K = 10$). Noisy samples and hard samples (e.g. profile and occluded faces) are automatically separated from the majority of clean samples.

Even though remarkable advances have been achieved by the margin-based softmax methods [8,15,32,31,5,39], they all need to be trained on well-annotated clean datasets [30,5], which require intensive human efforts. Wang et al. [30] found that faces with label noise significantly degenerate the recognition accuracy and manually built a high-quality dataset including 1.7M images of 59K celebrities. However, it took 50 annotators to work continuously for one month to clean the dataset, which further demonstrates the difficulty of obtaining a large-scale clean dataset for face recognition.

Since accurate manual annotations can be expensive [30], learning with massive noisy data [1] has recently attracted much attention [14,4,11,41,33]. However, computing time-varying weights for samples [11] or designing piece-wise loss functions [41] based on the current model's predictions can only alleviate the influence from noisy data to some extent as the robustness and improvement depend on the initial performance of the model. Besides, the co-mining method [33] requires to train twin networks together thus it is less practical for training large models on large-scale datasets.

As shown in Fig. 1(a), the objective of ArcFace [5] has two parts: (1) intra-class compactness: pushing the sample close to the corresponding positive center; and (2) inter-class discrepancy: pushing the sample away from all other negative centers. If a face is a noisy sample, it does not belong to the corresponding positive class. In ArcFace, this noisy sample generates a large wrong loss value, which impairs the model training. In this paper, we relax the intra-class constraint of forcing all samples close to the corresponding positive center by introducing sub-classes into ArcFace. For each class, we design K sub-centers and the training sample only needs to be close to any of the K positive sub-centers instead of the only one positive center. As illustrated in Fig. 1(b), the proposed sub-center ArcFace will encourage one dominant sub-class that contains the majority clean faces and multiple non-dominant sub-classes that include hard or noisy faces. This happens because the intra-class constraint of sub-center Arc-

[1] Generally, there are two types of label noise in face recognition [30,11,41,33]: one is open-set label noise, i.e., faces whose true labels are out of the training label set but are wrongly labeled to be within the set; and the other one is close-set label noise, i.e., faces whose true labels are in the training label set but are wrongly labeled.

Face enforces the training sample to be close to one of the multiple positive sub-classes but not all of them. The noise is likely to form a non-dominant sub-class and will not be enforced into the dominant sub-class. Therefore, sub-center ArcFace is more robust to noise. Extensive experimental results in this paper indicate that the proposed sub-center ArcFace is more robust than ArcFace [5] under massive real-world noises.

Although the proposed sub-center ArcFace can effectively separate clean data from noisy data. However, hard samples are also kept away. The existing of sub-centers can improve the robustness but also undermine the intra-class compactness, which is important for face recognition [34]. As the devil of face recognition is in the noise [30], we directly drop non-dominant sub-centers and high-confident noisy samples after the model achieves enough discriminative power. By pushing hard samples close to the dominant sub-center, we gradually recapture intra-class compactness and further improve the accuracy.

To summarise, our key contributions are as follows:

- We introduce sub-class into ArcFace to improve its robustness on noisy training data. The proposed sub-center ArcFace consistently outperforms ArcFace under massive real-world noises.
- By dropping non-dominant sub-centers and high-confident noisy samples, our method can achieve comparable performance compared to ArcFace trained on the manually cleaned dataset.
- Sub-center Arcface can be easily implemented by using the parallel toolkit and thus enjoys scalability to large-scale datasets. By taking advantage of the large-scale raw web faces (e.g. Celeb500K [1]), the proposed sub-center Arcface achieves state-of-the-art performance on IJB-B, IJB-C, MegaFace, and FRVT 1:1 Verification.

2 Related work

Face Recognition with Margin Penalty. The pioneering work [26] uses the Triplet loss to exploit triplet data such that faces from the same class are closer than faces from different classes by a clear Euclidean distance margin. Even though the Triplet loss makes perfect sense for face recognition, the sample-to-sample comparisons are local within mini-batch and the training procedure for the Triplet loss is very challenging as there is a combinatorial explosion in the number of triplets especially for large-scale datasets, requiring effective sampling strategies to select informative mini-batch [25,26] and choose representative triplets within the mini-batch [36,21,28]. Some works tried to reduce the total number of triplets with proxies [19,23], i.e., sample-to-sample comparison is changed into sample-to-proxy comparison. However, sampling and proxy methods only optimise the embedding of partial classes instead of all classes in one iteration step.

Margin-based softmax methods [15,8,32,31,5] focused on incorporating margin penalty into a more feasible framework, softmax loss, which has extensive sample-to-class comparisons. Compared to deep metric learning methods (e.g., Triplet [26], Tuplet [21,28]),

margin-based softmax methods conduct global comparisons at the cost of memory consumption on holding the center of each class. Sample-to-class comparison is more efficient and stable than sample-to-sample comparison as (1) the class number is much smaller than sample number, and (2) each class can be represented by a smoothed center vector which can be updated during training.

Face Recognition under Noise. Most of the face recognition datasets [40,9,2,1] are downloaded from the Internet by searching a pre-defined celebrity list, and the original labels are likely to be ambiguous and inaccurate [30]. Learning with massive noisy data has recently drawn much attention in face recognition [37,11,41,33] as accurate manual annotations can be expensive [30] or even unavailable.

Wu et al. [37] proposed a semantic bootstrap strategy, which re-labels the noisy samples according to the probabilities of the softmax function. However, automatic cleaning by the bootstrapping rule requires time-consuming iterations (e.g. twice refinement steps are used in [37]) and the labelling quality is affected by the capacity of the original model. Hu et al. [11] found that the cleanness possibility of a sample can be dynamically reflected by its position in the target logit distribution and presented a noise-tolerant end-to-end paradigm by employing the idea of weighting training samples. Zhong et al. [41] devised a noise-resistant loss by introducing a hypothetical training label, which is a convex combination of the original label with probability ρ and the predicted label by the current model with probability $1 - \rho$. However, computing time-varying fusion weight [11] and designing piece-wise loss [41] contain many hand-designed hyper-parameters. Besides, re-weighting methods are susceptible to the performance of the initial model. Wang et al. [33] proposed a co-mining strategy which uses the loss values as the cue to simultaneously detect noisy labels, exchange the high-confidence clean faces to alleviate the error accumulation caused by the sampling bias, and re-weight the predicted clean faces to make them dominate the discriminative model training. However, the co-mining method requires training twin networks simultaneously and it is challenging to train large networks (e.g. ResNet100 [10]) on a large-scale dataset (e.g. MS1M [9] and Celeb500K [1]).

Face Recognition with Sub-classes. Practices and theories that lead to "sub-class" have been studied for a long time [42,43]. The concept of "sub-class" applied in face recognition was first introduced in [42,43], where a mixture of Gaussians was used to approximate the underlying distribution of each class. For instance, a person's face images may be frontal view or side view, resulting in different modalities when all images are represented in the same data space. In [42,43], experimental results showed that sub-class divisions can be used to effectively adapt to different face modalities thus improve the performance of face recognition. Wan et al. [29] further proposed a separability criterion to divide every class into sub-classes, which have much less overlaps. The new within-class scatter can represent multi-modality information, therefore optimising this within-class scatter will separate different modalities more clearly and further increase the accuracy of face recognition. However, these work [42,43,29] only employed hand-designed feature descriptor on tiny under-controlled datasets.

Concurrent with our work, Softtriple [22] presents a multi-center softmax loss with class-wise regularizer. These multi-centers can capture the hidden distribution of the data better [20] due to the fact that they can capture the complex geometry of the orig-

inal data and help reduce the intra-class variance. On the fine-grained visual retrieval problem, the Softtriple [22] loss achieves better performance than the softmax loss as capturing local clusters is essential for this task. Even though the concept of "sub-class" has been employed in face recognition [42,43,29] and fine-grained visual retrieval [22], none of these work has considered the large-scale (e.g. 0.5 million classes) face recognition problem under massive noise (e.g. around 50% noisy samples within the training data).

3 The Proposed Approach

3.1 ArcFace

ArcFace [5] introduced an additive angular margin penalty into the softmax loss,

$$\ell_{\text{ArcFace}} = -\log \frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s\cos(\theta_{y_i}+m)} + \sum_{j=1,j\neq y_i}^{N} e^{s\cos\theta_j}}, \tag{1}$$

where θ_j is the angle between the embedding feature $\mathbf{x}_i \in \mathbb{R}^{512\times 1}$ of the i-th face sample and the j-th class center $W_j \in \mathbb{R}^{512\times 1}$. Given that the corresponding class label of $\mathbf{x}_i$ is y_i, θ_{y_i} represents the angle between $\mathbf{x}_i$ and the ground-truth center W_{y_i}. $m = 0.5$ is the angular margin parameter, $s = 64$ is the feature re-scale parameter, and N is the total class number. As there is a ℓ_2 normalisation step on both $\mathbf{x}_i$ and W_j, $\theta_j = arccos\left(W_j^T \mathbf{x}_i\right)$.

Taking advantage of parallel acceleration on both $\mathbf{x}_i$ and W_j, the implementation of ArcFace[2] can efficiently handle million-level identities on a single server with 8 GPUs (11GB 1080ti). This straightforward solution has changed the ingrained belief that large-scale global comparison with all classes is usually not attainable due to the bottleneck of GPU memory [26,28].

3.2 Sub-center ArcFace

Even though ArcFace [5] has shown its power in efficient and effective face feature embedding, this method assumes that training data are clean [5,30]. However, this is not true especially when the dataset is in large scale. How to enable ArcFace to be robust to noise is one of the main challenges that impeding the development of face representation and recognition [30]. In this paper, we address this problem by proposing the idea of using sub-classes for each identity, which can be directly adopted by ArcFace and will significantly increase its robustness.

Foster Sub-classes. As illustrated in Fig. 2, we set a sufficiently large K for each identity. Based on a ℓ_2 normalisation step on both embedding feature $\mathbf{x}_i \in \mathbb{R}^{512\times 1}$ and all sub-centers $W \in \mathbb{R}^{N\times K\times 512}$, we get the subclass-wise similarity scores $\mathcal{S} \in \mathbb{R}^{N\times K}$ by a matrix multiplication $W^T\mathbf{x}_i$. Then, we employ a max pooling step on the subclass-wise similarity score $\mathcal{S} \in \mathbb{R}^{N\times K}$ to get the class-wise similarity score $\mathcal{S}' \in \mathbb{R}^{N\times 1}$. The

[2] https://github.com/deepinsight/insightface/tree/master/recognition

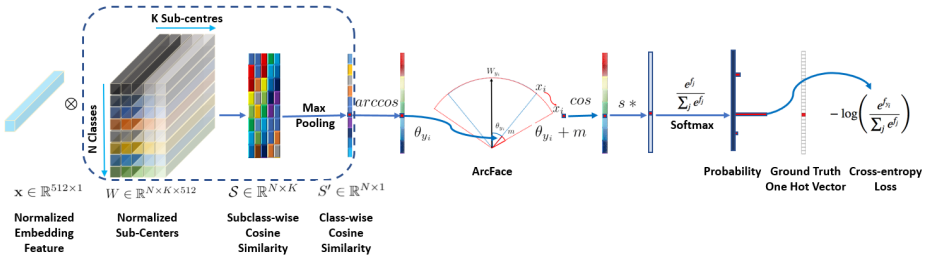
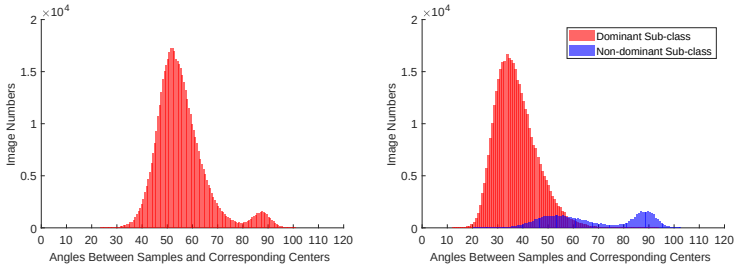


Fig. 2. Training the deep face recognition model by minimizing the proposed sub-center ArcFace loss. The main contribution in this paper is highlighted by the blue dashed box. Based on a ℓ_2 normalisation step on both embedding feature $\mathbf{x}_i \in \mathbb{R}^{512\times 1}$ and all sub-centers $W \in \mathbb{R}^{N\times K\times 512}$, we get the subclass-wise similarity score $\mathcal{S} \in \mathbb{R}^{N\times K}$ by a matrix multiplication $W^T\mathbf{x}_i$. After a max pooling step, we can easily get the class-wise similarity score $\mathcal{S}' \in \mathbb{R}^{N\times 1}$. The following steps are same as ArcFace [5].

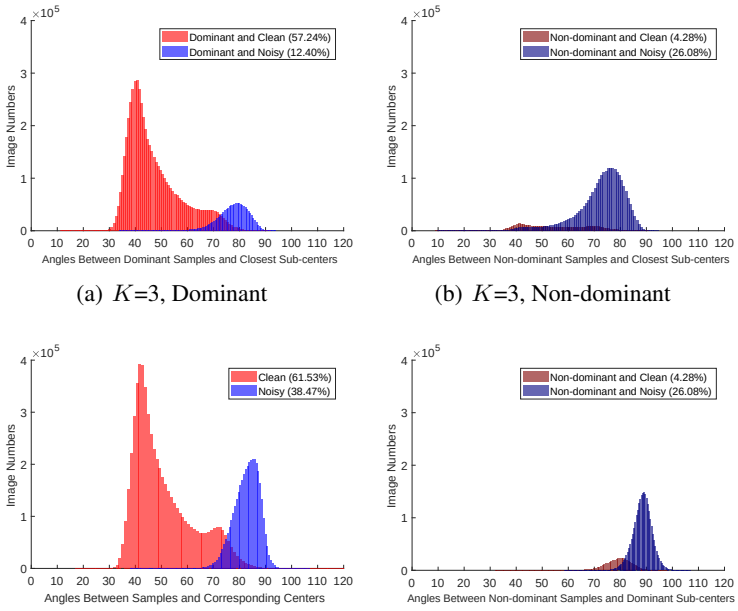
Table 1. The strictness and robustness analysis of different comparison strategies. In the angular space, "Min" is closest and "Max" is farest. "intra" refers to comparison between the training sample and the positive sub-centers (K). "inter" refers to comparison between the training sample and all negative sub-centers ($(N-1)\times K$). "outlier" denotes the open-set noise and "label flip" denotes the close-set noise.

Constraints	Sub-center?	Strictness?	Robustness to outlier?	Robustness to label flip?
(1) Min(inter) - Min(intra) $\geq$ m	$\surd$	+++	++	+
(2) Max(inter) - Min(intra) $\geq$ m	$\surd$	+	++	++
(3) Min(inter) - Max(intra) $\geq$ m		++++		
(4) Max(inter) - Max(intra) $\geq$ m		++		+



(a) Angle Distribution (b) Clean Data Isolation

Fig. 3. (a) Angle distribution of samples to their corresponding centers predicted by the pre-trained ArcFace model [5]. Noise exists in the CASIA dataset [40,30]. (b) Angle distribution of samples from the dominant and non-dominant sub-classes. Clean data are automatically isolated by sub-center ArcFace (K=10).



(a) K=3, Dominant (b) K=3, Non-dominant

(c) K=1, All (d) $K = 3 \downarrow 1$, Non-dominant

Fig. 4. Data distribution of ArcFace (K=1) and the proposed sub-center ArcFace (K=3) before and after dropping non-dominant sub-centers. MS1MV0 [9] is used here. $K = 3 \downarrow 1$ denotes sub-center ArcFace with non-dominant sub-centers dropping.

proposed sub-center ArcFace loss can be formulated as:

$$\ell_{\text{ArcFace}_{\text{subcenter}}} = -\log \frac{e^{s\cos(\theta_{i,y_i}+m)}}{e^{s\cos(\theta_{i,y_i}+m)} + \sum_{j=1,j\neq y_i}^{N} e^{s\cos\theta_{i,j}}}, \tag{2}$$

where $\theta_{i,j} = arccos\left(\max_k \left(W_{j_k}^T \mathbf{x}_i\right)\right), k \in \{1, \cdots, K\}$.

Robustness and Strictness Analysis. Given a large K, sub-classes are able to capture the complex distribution of the whole training data. Except for applying max pooling on the subclass-wise cosine similarity score, we can also consider other different comparison strategies. In Tab. 1, we give the strictness and robustness analysis of four comparison strategies. (1) adds angular margin between the closest inter-class sub-center and the closest intra-class sub-center. For intra-class comparison, choosing the closest positive sub-center can relax the intra-class constraint and improve the robustness under noise. For inter-class comparison, choosing the closest negative sub-center will enhance the inter-class constraint as sub-centers can better capture the complex geometric distributions of the whole data set compared to a single center for each class. However, the enhanced inter-class comparison is less robust under the close-set noise. The training procedure of (2) can not converge as the initial status between inter-classes is orthogonal and relaxing both of the inter-class and intra-class comparisons will disorient the

training, as there is no loss from inter-class comparisons. (3) and (4) can not foster sub-classes as stiffening intra-class comparison will compress sub-centers into one point in the high-dimension feature space thus undermine the robustness to noise.

Dominant and Non-dominant Sub-classes. In Fig. 1(b), we have visualised the clustering results of one identity from the CASIA dataset [40] after employing the sub-center ArcFace loss ($K = 10$) for training. It is obvious that the proposed sub-center ArcFace loss can automatically cluster faces such that hard samples and noisy samples are separated away from the dominant clean samples. Note that some sub-classes are empty as $K = 10$ is too large for a particular identity. In Fig. 3(a) and Fig. 3(b), we show the angle distribution on the CASIA dataset [40]. We use the pre-trained ArcFace model [5] to predict the feature center of each identity and then calculate the angle between the sample and its corresponding feature center. As we can see from Fig. 3(a), most of the samples are close to their centers, however, there are some noisy samples which are far away from their centers. This observation on the CASIA dataset matches the noise percentage estimation ($9.3\% \sim 13.0\%$) in [30]. To automatically obtain a clean training dataset, the noisy tail is usually removed by a hard threshold (e.g. angle $\geq 77^\circ$ or cosine ≤ 0.225). Since sub-center ArcFace can automatically divide the training samples into dominant sub-classes and non-dominant sub-classes, we visualise these two different kinds of samples in Fig. 3(b). As we can see from the two histograms, sub-center ArcFace can automatically separate clean samples from hard and noisy samples. More specifically, the majority of clean faces (85.6%) go to the dominant sub-class, while the rest hard and noisy faces go to the non-dominant sub-classes.

Drop Non-dominant Sub-centers and High-confident Noises. Even though using sub-classes can improve the robustness under noise, it undermines the intra-class compactness as hard samples are also kept away as shown in Fig. 3(b). In [9], MS1MV0 (around 10M images of 100K identities) is released with the estimated noise percentage around $47.1\% \sim 54.4\%$ [30]. In [6], MS1MV0 is refined by a semi-automatic approach into a clean dataset named MS1MV3 (around 5.1M images of 93K identities). Based on these two datasets, we can get clean and noisy labels on MS1MV0. In Fig. 4(a) and Fig. 4(b), we show the angle distributions of samples to their closest sub-centers (training settings: [MS1MV0, ResNet-50, Sub-center ArcFace K=3]). In general, there are four categories of samples: (1) easy clean samples belonging to dominant sub-classes (57.24%), (2) hard noisy samples belonging to dominant sub-classes (12.40%), (3) hard clean samples belonging to non-dominant sub-classes (4.28%), and (4) easy noisy samples belonging to non-dominant sub-classes (26.08%). In Fig. 4(c), we show the angle distribution of samples to their corresponding centers from the ArcFace model (training settings: [MS1MV0, ResNet50, ArcFace K=1]). By comparing the percentages of noisy sample in Fig. 4(a) and Fig. 4(c), we find that sub-center ArcFace can significantly decrease the noise rate to around one third (from 38.47% to 12.40%) and this is the reason why sub-center ArcFace is more robust under noise. During the training of sub-center ArcFace, samples belonging to non-dominant sub-classes are pushed to be close to these non-dominant sub-centers as shown in Fig. 4(b). Since we have not set any constraint on sub-centers, the sub-centers of each identity can be quite different and even orthogonal. In Fig. 4(d), we show the angle distributions of non-dominant samples to their dominant sub-centers. By combining Fig. 4(a) and Fig. 4(d), we find

that even though the clean and noisy data have some overlaps, a constant angle threshold (between 70° and 80°) can be easily searched to drop most of high-confident noisy samples.

Based on the above observations, we propose a straightforward approach to recapture intra-class compactness. We directly drop non-dominant sub-centers after the network has enough discriminative power. Meanwhile, we introduce a constant angle threshold to drop high-confident noisy data. After that, we retrain the model from scratch on the automatically cleaned dataset.

3.3 Comparison with Softtriple and Re-weighting Methods

The proposed sub-center ArcFace is different from Softtriple [22] in the following aspects:

- Softtriple shows improvement in fine-grained retrieval by employing multi-centers. However, we have not found obvious improvement when we directly use sub-centers on the clean dataset as sub-centers can undermine the intra-class compactness which is important for the face recognition problem. Our experimental analysis indicates that sub-centers can increase robustness under noise such that sub-center ArcFace can be trained on raw web faces without any manual cleaning step.
- Softtriple employs the softmax pooling (from sub-class similarity to class similarity) considering the smoothness. By contrast, we use the built-in max pooling without any performance drop. Max pooling is much more efficient than softmax pooling, especially for large-scale classification problem.
- Softtriple adds a similarity regularization between sub-centers. However, it is not reasonable that noisy data should be similar in our case. To enhance intra-class compactness, we only keep the dominant sub-center and drop the non-dominant sub-centers after the model has enough discriminative power. To decrease the affection from noisy data, we directly drop high-confident noisy data instead of employing complicated re-weighting strategies [41,11].
- Softriple only employs a small cosine margin (0.01) to explicitly break the tie during training. On the contrary, we use a large angular margin (0.5) setting as done by ArcFace.

The main difference between the proposed sub-center ArcFace and re-weighting methods [11,41] is that sub-center ArcFace is less affected by the noisy data from the beginning of the model training. By contrast, the discriminative power of the initial model is important for both NT [11] and NR [41] methods as their adaptive weights are predicted from the model.

Our sub-center ArcFace achieves high accuracy in face recognition while keeps extreme simplicity, only adding two hyper-parameters: the sub-center number and the constant threshold to drop high-confident noisy data.

4 Experiments

4.1 Experimental Settings

Datasets. Our training datasets include MS1MV0 (∼10M images of 100K identities) [9], MS1MV3 (∼5.1M faces of 91K identities) [6], and Celeb500K [1]. MS1MV0 is a raw data with the estimated noise percentage around $47.1\% \sim 54.4\%$ [30]. MS1MV3 is cleaned from MS1MV0 by a semi-automatic approach [6]. Celeb500K [1] is collected as MS1MV0 [9], using half of the MS1M name list [9] to search identities from Google and download the top-ranked face images. Our testing datasets consist of IJB-B [35], IJB-C [17], MegaFace [13], and Face Recognition Vendor Test (FRVT). Besides, we also report our final results on widely used verification datasets (e.g. LFW [12], CFP-FP [27], and AgeDB-30 [18]).

Implementation Details. For data pre-possessing, we follow ArcFace [5] to generate the normalised face crops (112×112) by utilising five facial points predicted by RetinaFace [7]. We employ ResNet-50 and ResNet-100 [10,5] to get the 512-D face embedding feature. Following [5], the feature scale s is set to 64 and the angular margin m is set to 0.5. All experiments in this paper are implemented by MXNet [3]. We set the batch size for back-propagation as 512 and train models on 8 NVIDIA Tesla P40 (24GB) GPUs. We set momentum to 0.9 and weight decay to $5e-4$. For the training of ArcFace on MS1MV0 and MS1MV3, the learning rate starts from 0.1 and is divided by 10 at the 100K, 160K, and 220K iteration steps. We finish the training process at 240K steps. For the training of the proposed sub-center ArcFace, we also employ the same learning rate schedule to train the first round of model (K=3). Then, we drop non-dominant sub-centers ($K = 3 \downarrow 1$) and high-confident noisy data ($> 75°$) by using the first round model through an off-line way. Finally, we retrain the model from scratch using the automatically cleaned data.

4.2 Ablation Study

To facilitate comparisons, we abbreviate different settings by the experiment number (E*) in the table and only focus on the TAR@FAR=1e-4 of IJB-C, which is more objective and less affected by the noise within the test data [38].

Real-world Noise. In Tab. 2, we conduct extensive experiments to investigate the proposed Sub-center ArcFace. We train ResNet-50 networks on different datasets (MS1MV0, MS1MV3 and Celeb500K) with different settings. From Tab. 2, we have the following observations: **(a)** ArcFace has an obvious performance drop (from E14 96.50% to E1 90.27%) when the training data is changed from the clean MS1MV3 to the noisy MS1MV0. By contrast, sub-center ArcFace is more robust (E2 93.72%) under massive noise. **(b)** Too many sub-centers (too large K) can obviously undermine the intra-class compactness and decrease the accuracy (from E2 93.72% to E5 67.94%). This observation indicates that robustness and strictness should be balanced during training, thus we select K=3 in this paper. **(c)** The nearest sub-center assignment by the max pooling is slightly better than the softmax pooling [22] (E2 93.72% vs. E3 93.55%). Thus, we choose the more efficient max pooling operator in the following experiments. **(d)**

Table 2. Ablation experiments of different settings on MS1MV0, MS1MV3 and Celeb500K. The 1:1 verification accuracy (TAR@FAR) is reported on the IJB-B and IJB-C datasets. ResNet-50 is used for training.

Settings	IJB-B					IJB-C				
	$1e{-}6$	$1e{-}5$	**1e-4**	$1e{-}3$	$1e{-}2$	$1e{-}6$	$1e{-}5$	**1e-4**	$1e{-}3$	$1e{-}2$
(1) MS1MV0,K=1	34.14	74.74	87.87	93.27	96.40	67.08	81.11	90.27	94.59	97.08
(2) MS1MV0,K=3	40.89	85.62	91.70	94.88	96.93	86.18	90.59	93.72	95.98	97.60
(3) MS1MV0,K=3, softmax pooling [22]	38.4	85.49	91.53	94.76	96.83	85.43	90.40	93.55	95.87	97.36
(4) MS1MV0,K=5	39.24	85.48	91.47	94.68	96.96	85.49	90.38	93.62	95.88	97.59
(5) MS1MV0,K=10	19.81	49.03	63.84	76.09	87.73	45.98	55.74	67.94	79.44	89.29
(6) MS1MV0, $K = 3 \downarrow 1$, drop $> 70°$	47.61	90.60	94.44	96.44	97.71	90.40	94.05	95.91	97.42	98.42
(7) MS1MV0, $K = 3 \downarrow 1$, drop $> 75°$	46.78	89.40	94.56	96.49	97.83	89.17	94.03	95.92	97.40	98.41
(8) MS1MV0, $K = 3 \downarrow 1$, drop $> 80°$	38.05	88.26	94.04	96.19	97.64	86.16	93.09	95.74	97.19	98.33
(9) MS1MV0, $K = 3 \downarrow 1$, drop $> 85°$	42.89	87.06	93.33	96.05	97.59	81.53	92.01	95.10	97.01	98.24
(10) MS1MV0, K=3, regularizer [22]	39.92	85.51	91.53	94.77	96.92	85.44	90.41	93.64	95.85	97.40
(11) MS1MV0,Co-mining [33]	40.96	85.57	91.80	94.99	97.10	86.31	90.71	93.82	95.95	97.63
(12) MS1MV0,NT [11]	40.84	85.56	91.57	94.79	96.83	86.14	90.48	93.65	95.86	97.54
(13) MS1MV0,NR [41]	40.86	85.53	91.58	94.77	96.80	86.07	90.41	93.60	95.88	97.44
(14) MS1MV3, K=1	35.86	91.52	95.13	96.61	97.65	90.16	94.75	96.50	97.61	98.40
(15) MS1MV3, K=3	40.16	91.30	94.84	96.66	97.74	90.64	94.68	96.35	97.66	98.48
(16) MS1MV3, $K = 3 \downarrow 1$	40.18	91.32	94.87	96.70	97.81	90.67	94.74	96.43	97.66	98.47
(17) Celeb500K, K=1	42.42	88.18	90.96	92.19	93.00	88.18	90.87	92.15	95.47	97.64
(18) Celeb500K, K=3	43.84	90.91	93.76	95.12	96.00	90.92	93.66	94.90	96.21	98.02
(19) Celeb500K, $K = 3 \downarrow 1$	44.64	92.71	95.65	96.94	97.89	92.73	95.52	96.91	97.87	98.42

Dropping non-dominant sub-centers and high-confident noisy samples can achieve better performance than adding regularization [22] to enforce compactness between sub-centers (E7 95.92% vs. E10 93.64%). Besides, The performance of our method is not very sensitive to the constant threshold (E6 95.91%, E7 95.92% and E8 95.74%), and we select $75°$ as the threshold for dropping high-confident noisy samples in the following experiments. **(e)** Co-mining [33] and re-weighting methods [11,41] can also improve the robustness under massive noise, but sub-center ArcFace can do better through automatic clean and noisy data isolation during training (E7 95.92% vs. E11 93.82%, E12 93.65% and E13 93.60%). **(f)** On the clean dataset (MS1MV3), sub-center ArcFace achieves similar performance as ArcFace (E16 96.43% vs. E14 96.50%). **(g)** The proposed sub-center ArcFace trained on noisy MS1MV0 can achieve comparable performance compared to ArcFace trained on manually cleaned MS1MV3 (E7 95.92% vs. E14 96.50%). **(h)** By enlarging the training data, sub-center ArcFace can easily achieve better performance even though it is trained from noisy web faces (E19 96.91% vs. E13 96.50%).

Synthetic Noise. In Tab. 3, we investigate the robustness of the proposed sub-center ArcFace under synthetic open-set and close-set noise. We train ResNet-50 networks on MS1MV3 with different noise types and levels. To simulate the training data with controlled open-set noise, we randomly select 75% and 50% identities from MS1MV3 [6] and the face images of the rest identities are assigned with random labels of selected identities. To simulate the training data with controlled close-set noise, we use all iden-

Table 3. Ablation experiments of different settings under synthetic open-set and close-set noise. The 1:1 verification accuracy (TAR@FAR) is reported on the IJB-B and IJB-C datasets. ResNet-50 is used for training.

Settings	IJB-B					IJB-C				
	$1e{-}6$	$1e{-}5$	**1e-4**	$1e{-}3$	$1e{-}2$	$1e{-}6$	$1e{-}5$	**1e-4**	$1e{-}3$	$1e{-}2$
	Synthetic Open-set Noise									
(1) 75%CleanID,K=1	37.49	90.02	94.48	96.48	97.72	90.10	94.18	96.00	97.45	98.38
(2) 75%CleanID+25%NoisyID,K=1	37.80	86.68	92.96	94.72	95.80	86.19	92.03	94.52	95.89	97.29
(3) 75%CleanID+25%NoisyID,K=3	38.31	87.87	94.17	95.83	97.15	87.23	93.01	95.57	96.95	97.75
(4) 75%CleanID+25%NoisyID,$K=3\downarrow 1$	38.36	88.14	94.20	96.15	97.94	87.51	93.27	95.89	97.29	98.43
(5) 50%CleanID,K=1	34.43	89.36	93.97	96.26	97.63	88.35	93.49	95.65	97.28	98.35
(6) 50%CleanID+50%NoisyID,K=1	35.96	81.45	90.77	92.69	94.56	80.97	88.49	92.25	93.84	95.10
(7) 50%CleanID+50%NoisyID,K=3	34.15	85.13	92.62	94.98	96.77	84.43	91.00	94.50	95.79	97.33
(8) 50%CleanID+50%NoisyID,$K=3\downarrow 1$	34.55	86.43	93.85	96.13	97.37	85.22	91.82	95.50	96.73	98.16
	Synthetic Close-set Noise									
(9) 75%CleanIM,K=1	38.44	89.41	94.76	96.42	97.71	89.31	94.19	96.19	97.39	98.43
(10) 75%CleanIM+25%NoisyIM,K=1	36.16	83.46	92.29	94.85	95.61	82.20	91.24	94.28	95.58	97.58
(11) 75%CleanIM+25%NoisyIM,K=3	36.09	83.16	91.45	94.33	95.23	81.28	90.02	93.57	94.96	96.32
(12) 75%CleanIM+25%NoisyIM,$K=3\downarrow 1$	37.79	85.50	94.03	95.53	97.42	84.09	93.17	95.13	96.85	97.61
(13) 50%CleanIM,K=1	36.85	90.50	94.59	96.49	97.65	90.46	94.32	96.08	97.44	98.33
(14) 50%CleanIM+50%NoisyIM,K=1	17.54	43.10	71.76	82.08	93.38	28.40	55.46	75.80	88.22	94.68
(15) 50%CleanIM+50%NoisyIM,K=3	17.47	41.63	66.42	78.70	91.37	26.03	54.23	72.04	86.36	94.19
(16) 50%CleanIM+50%NoisyIM,$K=3\downarrow 1$	22.19	68.11	85.86	88.13	95.08	44.34	69.25	78.12	90.51	96.16

tities ($\sim 100K$) from MS1MV3 [6] but randomly select 25% and 50% face images of each identity and assign random labels to these face images.

From Tab. 3, we have the following observations: **(a)** Performance drops as the ratio of synthetic noise increases, especially for the close-set noise (E2 94.52% vs. E6 92.25% and E10 94.28% vs. E14 75.80%). In fact, close-set noise is also found to be more harmful than open-set noise in [30]. **(b)** Under the open-set noise, the proposed sub-center can effectively enhance the robustness of ArcFace (E3 95.57% vs. E2 94.52% and E7 94.50% vs. E6 92.25%). By dropping non-dominant sub-centers and high-confident noisy samples, the performance of sub-center arcface can even approach Arcface trained on the clean dataset (E4 95.89% vs. E1 96.00% and E8 95.50% vs. E5 95.65%). **(c)** Under the close-set noise, the performance of sub-center Arcface is worse than ArcFace (E11 93.57% vs. E10 94.28% and E15 72.04% vs. E14 75.80%), as the inter-class constraint of sub-center Arcface is more strict than ArcFace. By dropping non-dominant sub-centers and high-confident noisy samples, the performance of sub-center Arcface outperforms ArcFace (E12 95.13% vs. E10 94.28% and E16 78.12% vs. E14 75.80%) but still lags behind ArcFace trained on the clean dataset (E12 95.13% vs. E9 96.19% and E16 78.12% vs. E13 96.08%), which indicates the capacity of the network to drop noisy samples depends on its initial discriminative power. Sub-center ArcFace trained on 50% close-set noise is far from accurate (E15 72.04%) and the step of dropping noisy samples is also not accurate. Therefore, it is hard to catch up with ArcFace trained on the clean dataset. However, in the real-world data, close-set noise

Table 4. Column 2-3: 1:1 verification TAR (@FAR=1e-4) on the IJB-B and IJB-C dataset. Column 4-5: Face identification and verification evaluation on MegaFace Challenge1 using FaceScrub as the probe set. "Id" refers to the rank-1 face identification accuracy with 1M distractors, and "Ver" refers to the face verification TAR at 10^{-6} FAR. Column 6-8: The 1:1 verification accuracy on the LFW, CFP-FP and AgeDB-30 datasets. ResNet-100 is used for training.

Settings	IJB		MegaFace		Quick Verification Datasets		
	IJB-B	IJB-C	Id	Ver	LFW	CFP-FP	AgeDB-30
MS1MV0, K=1	87.91	90.42	96.52	96.75	99.75	97.17	97.26
MS1MV0, $K = 3 \downarrow 1$	94.94	96.28	98.16	98.36	99.80	98.80	98.31
MS1MV3, K=1 [5,6]	95.25	96.61	98.40	98.51	99.83	98.80	**98.45**
Celeb500K, $K = 3 \downarrow 1$	**95.75**	**96.96**	**98.78**	**98.69**	**99.86**	**99.11**	98.35

Table 5. FRVT 1:1 verification results. Sub-center ArcFace ($K = 3 \downarrow 1$) employs ResNet-100 and is trained on the Celeb500K dataset. FNMR is the proportion of mated comparisons below a threshold set to achieve the false match rate (FMR) specified. FMR is the proportion of impostor comparisons at or above that threshold.

Rank	Submissions	WILD FNMR @FMR $\leq$ 1e-5	VISA FNMR @FMR $\leq$ 1e-6	VISA FNMR @FMR $\leq$ 1e-4	MUGSHOT FNMR @FMR $\leq$ 1e-5	MUGSHOT FNMR @FMR $\leq$ 1e-5 DT=14 YRS	VISABORDER FNMR @FMR $\leq$ 1e-6
1	deepglint-002	0.0301	0.0027	0.0004	0.0032	0.0041	0.0043
2	everai-paravision-003	0.0302	0.0050	0.0011	0.0036	0.0053	0.0092
3	Sub-center ArcFace	0.0303	0.0081	0.0027	0.0055	0.0087	0.0083
4	dahua-004	0.0304	0.0058	0.0019	0.0036	0.0051	0.0051
5	xforwardai-000	0.0305	0.0072	0.0018	0.0036	0.0051	0.0074
6	visionlabs-008	0.0308	0.0036	0.0007	0.0031	0.0044	0.0045
7	didiglobalface-001	0.0308	0.0092	0.0016	0.0030	0.0048	0.0088
8	vocord-008	0.0310	0.0038	0.0008	0.0042	0.0054	0.0045
9	paravision-004	0.0311	0.0046	0.0012	0.0030	0.0041	0.0091
10	ntechlab-008	0.0312	0.0061	0.0011	0.0056	0.0106	0.0042
11	tevian-005	0.0325	0.0062	0.0020	0.0057	0.0081	0.0070
12	sensetime-003	0.0355	0.0027	0.0005	0.0027	0.0033	0.0051
13	yitu-003	0.0360	0.0026	0.0003	0.0066	0.0083	0.0064

is not dominant, much less than 50% (e.g. only a small part of celebrities frequently appear in others' album).

4.3 Benchmark Results

Results on IJB-B [35] and IJB-C [35]. We employ the face detection scores and the feature norms to re-weigh faces within templates [24,16]. In Tab. 4, we compare the TAR (@FAR=1e-4) of ArcFace and the proposed sub-center ArcFace trained on noisy data (e.g. MS1MV0 and Celeb500K). The performance of ArcFace significantly drops from 96.61% to 90.42% on the IJB-C dataset when the training data is changed from the manually cleaned data (MS1MV3) to the raw noisy data (MS1MV0). By contrast, the proposed sub-center ArcFace is robust to massive noise and can achieve similar results compared with ArcFace trained on the clean data (96.28% vs. 96.61%). When we apply sub-center ArcFace on large-scale training data (Celeb500K), we further improve the TAR (@FAR=1e-4) to 95.75% and 96.96% on IJB-B and IJB-C, respectively.

Results on MegaFace [13]. We adopt the refined version of MegaFace [5] to give a fair evaluation. As shown in Tab. 4, the identification accuracy of ArcFace obviously drops from 98.40% to 96.52% when the training data is changed from MS1MV3 to MS1MV0, while the proposed sub-center ArcFace is more robust under massive noise within MS1MV0, achieving the identification accuracy of 98.16%. ArcFace trained on MS1MV3 only slightly outperforms our method trained on MS1MV0 under both verification and identification protocols. Finally, the sub-center ArcFace model trained on the large-scale Celeb500K dataset achieves state-of-the-art identification accuracy of 98.78% on the MegaFace dataset.

Results on LFW [12], CFP-FP [27], and AgeDB-30 [18]. We follow the *unrestricted with labelled outside data* protocol to report the verification performance. As reported in Tab. 4, sub-center ArcFace trained on noisy MS1MV0 achieves comparable performance compared to ArcFace trained on clean MS1MV3. Moreover, our method trained on the noisy Celeb500K outperforms ArcFace [5], achieving the verification accuracy of 99.86%, 99.11%, 98.35% on LFW, CFP-FP and AgeDB-30, respectively.

Results on FRVT. The Face Recognition Vendor Test (FRVT) is the most strict industry-level face recognition test, and the participants need to submit the whole face recognition system (e.g. face detection, alignment and feature embedding) to the organiser. No test image has been released for hyper-parameter searching and the submission interval is no less than three months. Besides, the submitted face recognition system should complete face detection and face feature embedding within 1000ms on Intel Xeon CPU (E5-2630 v4 @ 2.20GHz processors) by using the single-thread inference. We build our face recognition system by RetinaFace (ResNet-50) [7] and sub-center ArcFace (ResNet-100), and accelerate the inference by the openVINO toolkit. In Tab. 5, we show the top-performing 1:1 algorithms measured on false non-match rate (FNMR) across several different tracks. As we can see from the results, the proposed sub-center ArcFace trained on the Celeb500K dataset achieves state-of-the-art performance on the wild track (0.0303, rank 3rd). Considering several hundred of industry submissions to FRVT, the overall performance of our single model is very impressive.

5 Conclusion

In this paper, we have proposed sub-center ArcFace which first enforces sub-classes by nearest sub-center selection and then only keeps the dominant sub-center to achieve intra-class compactness. As we relax the intra-class compactness from beginning, the proposed sub-center ArcFace is robust under massive label noise and can easily train face recognition models from raw downloaded data. Extensive experimental results show that our method consistently outperforms ArcFace on real-world noisy datasets and achieve comparable performance compared to using manually refined data.

Acknowledgements. Jiankang Deng acknowledges the Imperial President's PhD Scholarship. Tongliang Liu acknowledges support from the Australian Research Council Project DE-190101473. Stefanos Zafeiriou acknowledges support from the Google Faculty Fellowship, EPSRC DEFORM (EP/S010203/1) and FACER2VM (EP/N007743/1). We are thankful to Nvidia for the GPU donations.

References

1. Cao, J., Li, Y., Zhang, Z.: Celeb-500k: A large training dataset for face recognition. In: ICIP (2018)
2. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: FG (2018)
3. Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C., Zhang, Z.: Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. arXiv:1512.01274 (2015)
4. Cheng, J., Liu, T., Ramamohanarao, K., Tao, D.: Learning with bounded instance-and label-dependent label noise. ICML (2020)
5. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR (2019)
6. Deng, J., Guo, J., Zhang, D., Deng, Y., Lu, X., Shi, S.: Lightweight face recognition challenge. In: ICCV Workshops (2019)
7. Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., Zafeiriou, S.: Retinaface: Single-stage dense face localisation in the wild. arXiv:1905.00641 (2019)
8. Deng, J., Zhou, Y., Zafeiriou, S.: Marginal loss for deep face recognition. In: CVPR Workshops (2017)
9. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: ECCV (2016)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
11. Hu, W., Huang, Y., Zhang, F., Li, R.: Noise-tolerant paradigm for training face recognition cnns. In: CVPR (2019)
12. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep. (2007)
13. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: CVPR (2016)
14. Liu, T., Tao, D.: Classification with noisy labels by importance reweighting. TPAMI (2015)
15. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: CVPR (2017)
16. Masi, I., Tran, A.T., Hassner, T., Sahin, G., Medioni, G.: Face-specific data augmentation for unconstrained face recognition. IJCV (2019)
17. Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J.: Iarpa janus benchmark–c: Face dataset and protocol. In: ICB (2018)
18. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: The first manually collected in-the-wild age database. In: CVPR Workshops (2017)
19. Movshovitz-Attias, Y., Toshev, A., Leung, T.K., Ioffe, S., Singh, S.: No fuss distance metric learning using proxies. In: ICCV (2017)
20. Müller, R., Kornblith, S., Hinton, G.: Subclass distillation. arXiv:2002.03936 (2020)
21. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: CVPR (2016)
22. Qian, Q., Shang, L., Sun, B., Hu, J., Li, H., Jin, R.: Softtriple loss: Deep metric learning without triplet sampling. In: ICCV (2019)
23. Qian, Q., Tang, J., Li, H., Zhu, S., Jin, R.: Large-scale distance metric learning with uncertainty. In: CVPR (2018)
24. Ranjan, R., Bansal, A., Xu, H., Sankaranarayanan, S., Chen, J.C., Castillo, C.D., Chellappa, R.: Crystal loss and quality pooling for unconstrained face verification and recognition. arXiv:1804.01159 (2018)

25. Rippel, O., Paluri, M., Dollar, P., Bourdev, L.: Metric learning with adaptive density discrimination. In: ICLR (2016)
26. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR (2015)
27. Sengupta, S., Chen, J.C., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: WACV (2016)
28. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. In: NeurIPS (2016)
29. Wan, H., Wang, H., Guo, G., Wei, X.: Separability-oriented subclass discriminant analysis. TPAMI (2017)
30. Wang, F., Chen, L., Li, C., Huang, S., Chen, Y., Qian, C., Loy, C.C.: The devil of face recognition is in the noise. In: ECCV (2018)
31. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. SPL (2018)
32. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: CVPR (2018)
33. Wang, X., Wang, S., Wang, J., Shi, H., Mei, T.: Co-mining: Deep face recognition with noisy labels. In: ICCV (2019)
34. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: ECCV (2016)
35. Whitelam, C., Taborsky, E., Blanton, A., Maze, B., Adams, J.C., Miller, T., Kalka, N.D., Jain, A.K., Duncan, J.A., Allen, K.: Iarpa janus benchmark-b face dataset. In: CVPR Workshops (2017)
36. Wu, C.Y., Manmatha, R., Smola, A.J., Krahenbuhl, P.: Sampling matters in deep embedding learning. In: ICCV (2017)
37. Wu, X., He, R., Sun, Z., Tan, T.: A light cnn for deep face representation with noisy labels. TIFS (2018)
38. Xie, W., Li, S., Zisserman, A.: Comparator networks. In: ECCV (2018)
39. Yang, J., Bulat, A., Tzimiropoulos, G.: Fan-face: a simple orthogonal improvement to deep face recognition. In: AAAI (2020)
40. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv:1411.7923 (2014)
41. Zhong, Y., Deng, W., Wang, M., Hu, J., Peng, J., Tao, X., Huang, Y.: Unequal-training for deep face recognition with long-tailed noisy data. In: CVPR (2019)
42. Zhu, M., Martínez, A.M.: Optimal subclass discovery for discriminant analysis. In: CVPR Workshops (2004)
43. Zhu, M., Martinez, A.M.: Subclass discriminant analysis. TPAMI (2006)