Know your Face Recognition data

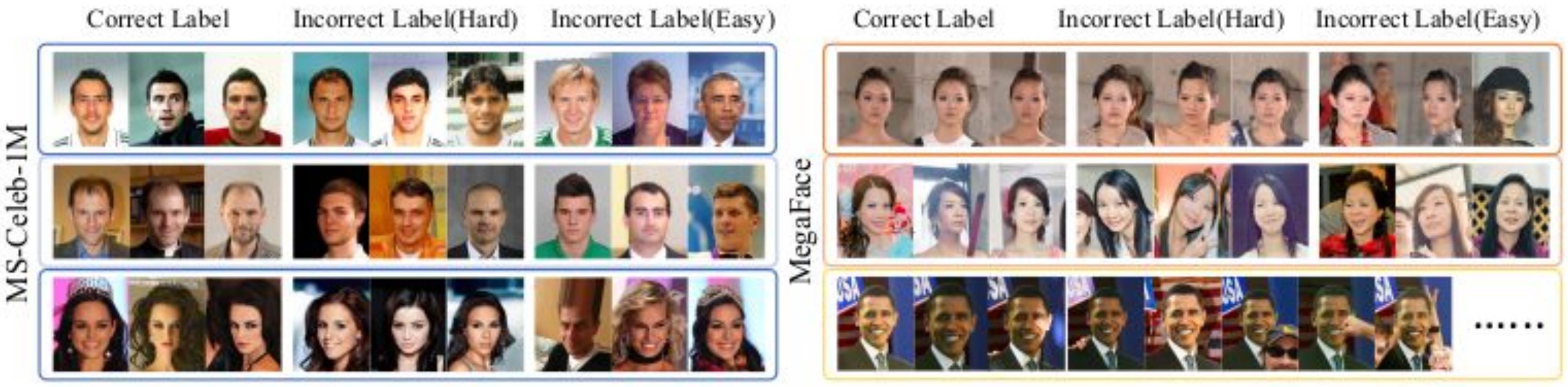
The Devil of Face Recognition is in the Noise

Fei Wang ⋆1[0000−0002−1024−5867], Liren Chen ⋆2[0000−0003−0113−5233], Cheng Li1[0000−0002−0892−4705], Shiyao Huang1[0000−0002−5198−2492], Yanjie Chen1[0000−0003−1918−6776], Chen Qian1[0000−0002−8761−5563], and Chen Change Loy3[0000−0001−5345−1591]

[1] SenseTime Research
[2] University of California San Diego
[3] Nanyang Technological University
{wangfei, chengli, huangshiyao, chenyanjie, qianchen}@sensetime.com, lic002@eng.ucsd.edu, ccloy@ieee.org

Problem

- Large Face Recognition datasets (MegaFace, MS-Celeb-1M) required for training strong convolutional network, are creating using automatic/semi-automatic methods, thus contain large amount of labe noises.
- labels flip: sample has been given wrong label of another class within the dataset.
- outliers: sample does not belong to any of the classes within dataset, but mistakenly has one of their labels.



Contribution

- Analyze the effect of noisy labels on Face Recognition networks.
- Contribute relatively large manually cleaned dataset - IMDB-Face.
- Analyze of labeling methods efficiency.

Datasets Overview

LFW - 13K images : 1.6K ID, collected from Yahoo News, running Viola-jones detector. Limited by detector most of faces are frontal. Considered sufficiently clean.
CASIA-WebFace - 500K images : 10K ID, collected from IMDB, semi-automatically cleaned via tag-constrained similarity clustering.
MS-Celeb-1M - scrapping from public search engines, approximately 100 images per ID. the data is deliberately left uncleaned.
MegaFace - based on YFCC100M dataset collected from Flickr, semi-automatic cleaned.

Dataset	#Identities	#Images	Source	Cleaned?	Availablity
LFW [7]	5K	13K	Search Engine	Automatic Detection	Public
CelebFaces [19,20]	10K	202K	Search Engine	Manually Cleaned	Public
VGG-Face [15]	2.6K	2.5M	Search Engine	Semi-automated Clean	Public
CASIA-WebFace [25]	10k	0.5M	IMDb	Automatic Clean	Public
MS-Celeb-1M(v1) [5]	100k	10M	Search Engine	None	Public
MegaFace [13]	670K	4.7M	Flickr	Automatic Cleaned	Public
Facebook [21]	4k	4.4M	–	–	Private
Google [18]	8M	200M	–	–	Private
IMDb-Face	**59K**	**1.7M**	**IMDb**	**Manually Cleaned**	**Public**

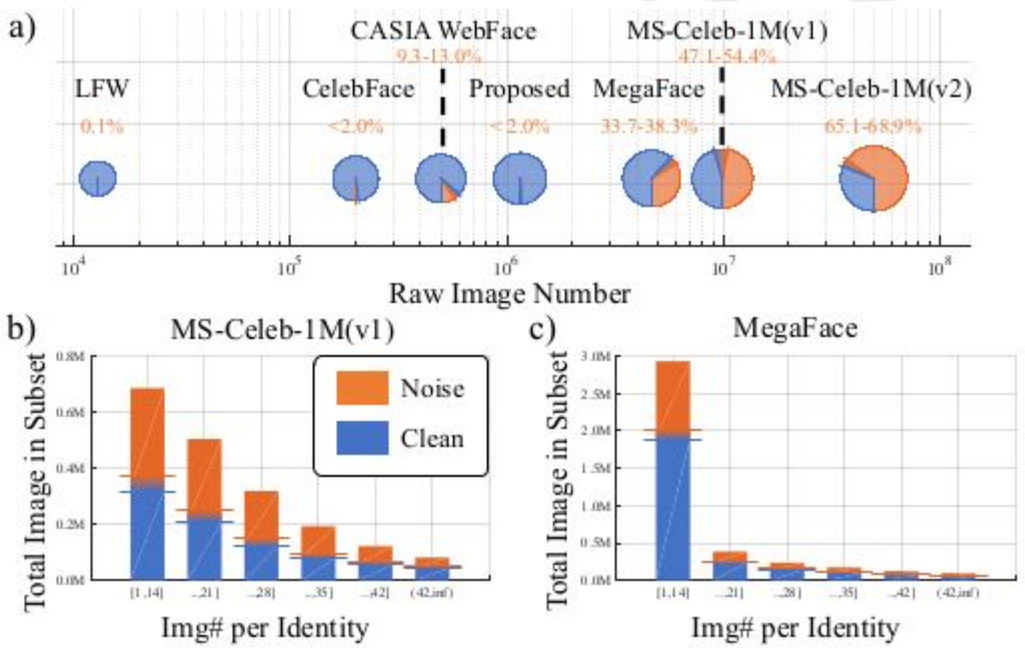
Signal to Noise Ratio

Manually cleaned subsets:

- 2.7M from MegaFace.
- 3.7M from MS-Celeb-1M
- Casia / CelebFaces 30 Id's.

Face recognition datasets with more than million samples have a noise ratio higher than 30%.

Imdb-Face: manually cleaned by workers, with approximated noise level under 2%.



Experiment

- Attention-56, batch-size of 256, 256D feature.
- Imdb-Face dataset.

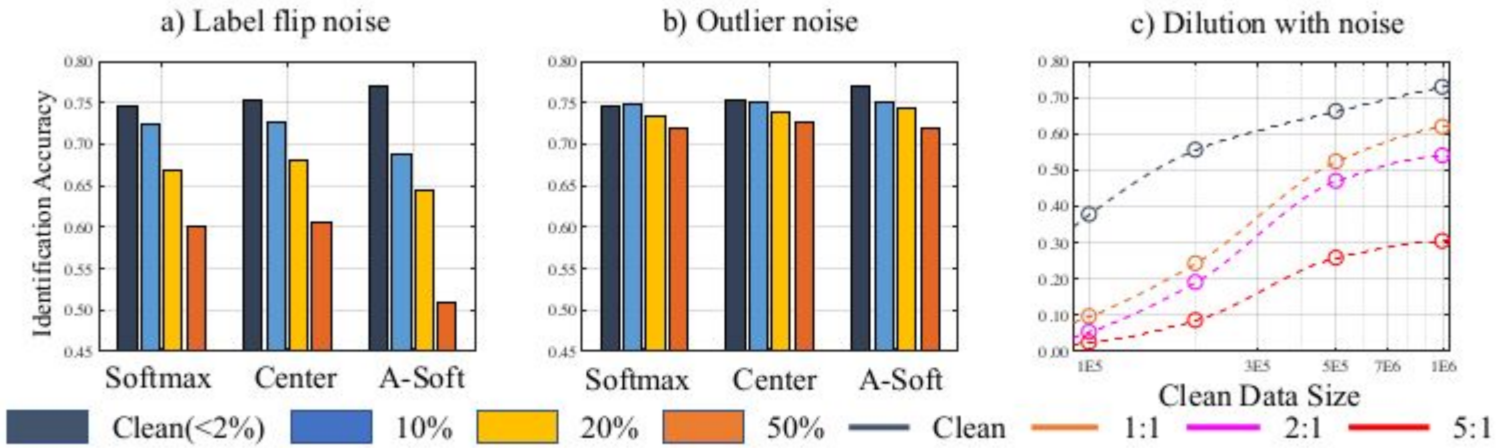


Fig. 7. 1:1M rank-1 identification results on MegaFace benchmark: (a) introducing label flips to IMDb-Face, (b) introducing outliers to IMDb-Face, and (c) fixing the size of clean data and dilute it with different ratios of label flips.

Table 2. Noisy data vs. Clean data. The results are obtained from rank-1 identification test on the MegaFace benchmark [8]. Abbreviation MSV1 = MS-Celeb-1M(v1).

Dataset	#Iden.	#Imgs.	MegaFace Rank-1(%)		
			Softmax	Center	A-softmax
MSV1-raw	96k	8.6M	71.70	73.82	73.99
-sampled	46k	3.7M	66.15	69.81	70.56
-clean	46k	1.76M	70.66	73.15	73.53
MegaFace-raw	670k	4.7M	64.32	64.71	66.95
-sampled	270k	2.7M	59.68	62.55	63.12
-clean	270k	1.5M	62.86	67.64	68.88

- average improvement of accuracy between clean and sampled is 4.14%.
- close and in some cases better results than the larger raw dataset.

Training with different datasets:

Dataset	#Iden.	#Imgs.	Rank-1 (%)		
			Softmax	Center Loss	A-Softmax
CelebFaces	10k	0.20M	36.15	42.54	43.72
CASIA-WebFace	10.5k	0.49M	65.17	68.09	70.89
MS-Celeb-1M(V1)	96k	8.6M	71.70	73.82	73.99
MegaFace	670k	4.7M	64.32	64.71	66.95
IMDbFace	59k	1.7M	**74.75**	**79.41**	**84.06**

Comparison to SOTA?? :

Method, Dataset	LFW	Mega(Ident.)	YTF
Vocord-deep V3†, Private	-	**91.76**	-
YouTu Lab†, Private	-	83.29	-
DeepSense V2†, Private	-	81.23	-
Marginal Loss♯ [4] MS-Celeb-1M	99.48	80.278	95.98
SphereFace [12],CASIA-WebFace	99.42	75.77	95.00
Center Loss [23],CASIA-WebFace	99.28	65.24	94.90
A-Softmax♯, MS-Celeb-1M	99.58	73.99	97.45
A-Softmax♯, IMDb-Face	**99.79**	**84.06**	**97.67**

† Commercial, have not been published

♯ Single Model

ArcFace: Additive Angular Margin Loss for Deep Face Recognition

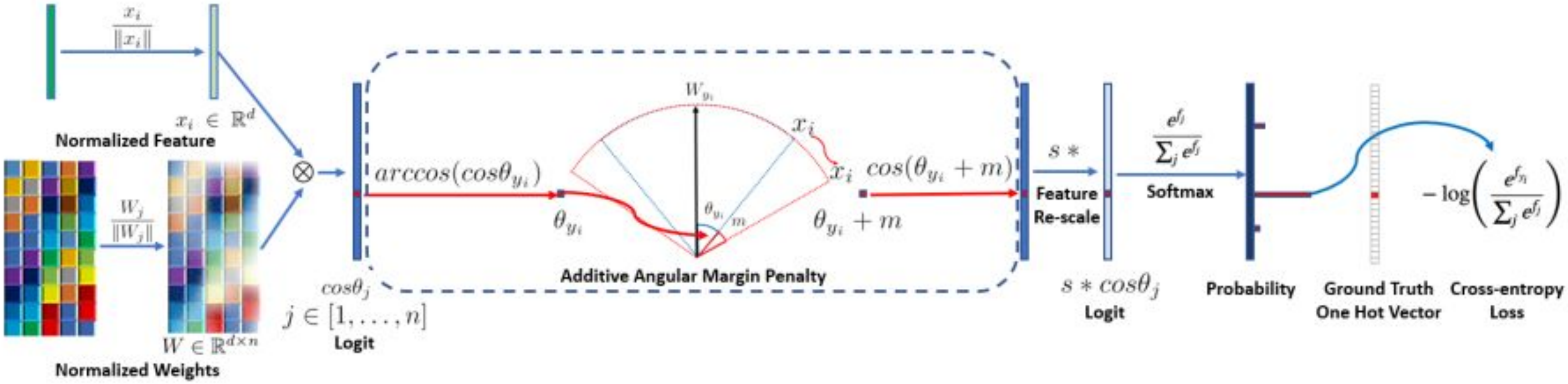
Jiankang Deng *
Imperial College London
j.deng16@imperial.ac.uk

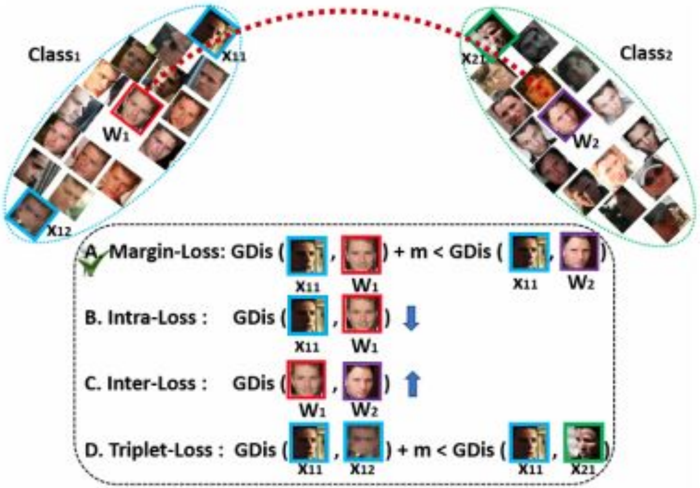
Jia Guo *
InsightFace
guojia@gmail.com

Niannan Xue
Imperial College London
n.xue15@imperial.ac.uk

Stefanos Zafeiriou
Imperial College London
s.zafeiriou@imperial.ac.uk

Arc Loss





$$L_3 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))}+\sum_{j=1,j\neq y_i}^{n}e^{s\cos\theta_j}}$$

s - hypersphere radius
n - number of classes (identities)
m - additive margin parameter
x - normlized features embeding
W - norrmalized weight class
Ø - angle between x and W
N - batch size

Arc Loss

Method	#Image	LFW	YTF
DeepID [32]	0.2M	99.47	93.20
Deep Face [33]	4.4M	97.35	91.4
VGG Face [24]	2.6M	98.95	97.30
FaceNet [29]	200M	99.63	95.10
Baidu [16]	1.3M	99.13	-
Center Loss [38]	0.7M	99.28	94.9
Range Loss [46]	5M	99.52	93.70
Marginal Loss [9]	3.8M	99.48	95.98
SphereFace [18]	0.5M	99.42	95.0
SphereFace+ [17]	0.5M	99.47	-
CosFace [37]	5M	99.73	97.6
MS1MV2, R100, ArcFace	5.8M	**99.83**	**98.02**

Table 4. Verification performance (%) of different methods on LFW and YTF.

s = 64, m= 0.5

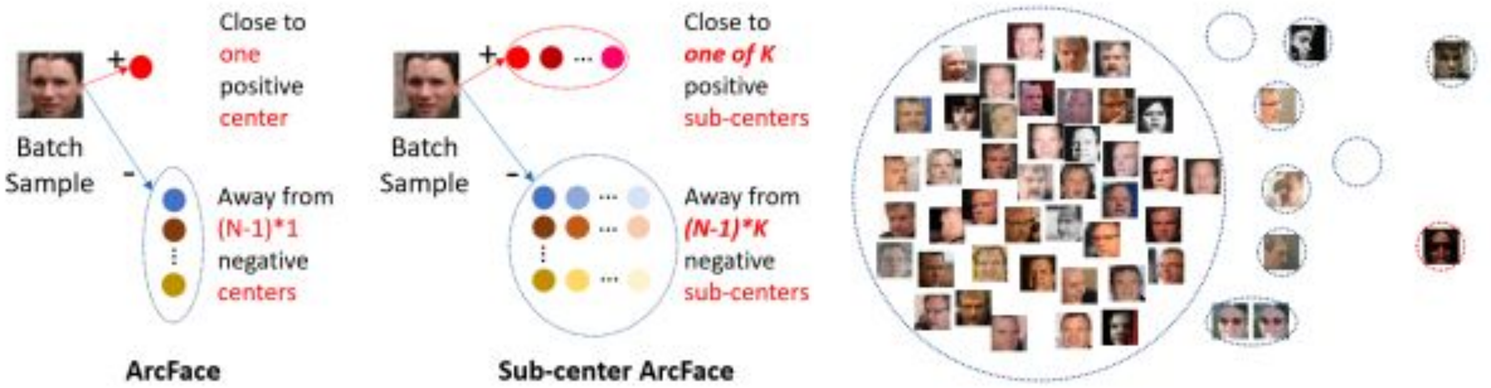
MS1MV2 - Semi Automatic refined version of MS-Celeb-1M dataset

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

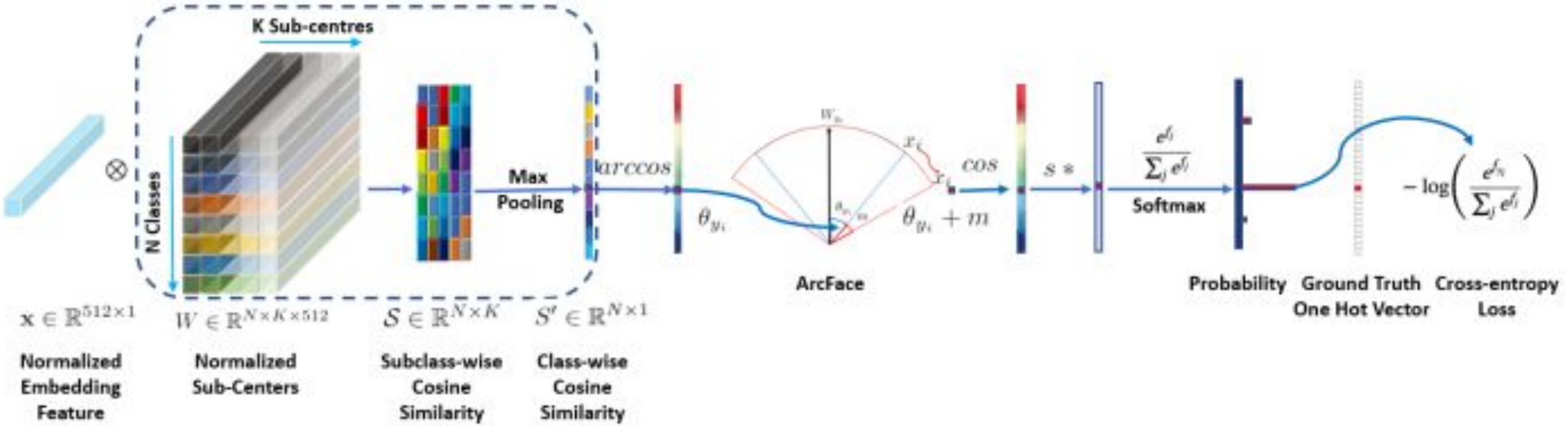
Jiankang Deng * [1] Jia Guo * [2] Tongliang Liu[3]
Mingming Gong [4] Stefanos Zafeiriou[1]

[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

- Even though ArcFace is efficient, this method assume that training data is clean.
- Reduce the intra-class constraint, and improve the robustness to label noise.
- Design of K-sub-centers for each class.
- Most of clean faces will be close to a dominant sub-class, and non-dominant sub classes will include noisy faces.



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



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

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

Casia-Webface distribution

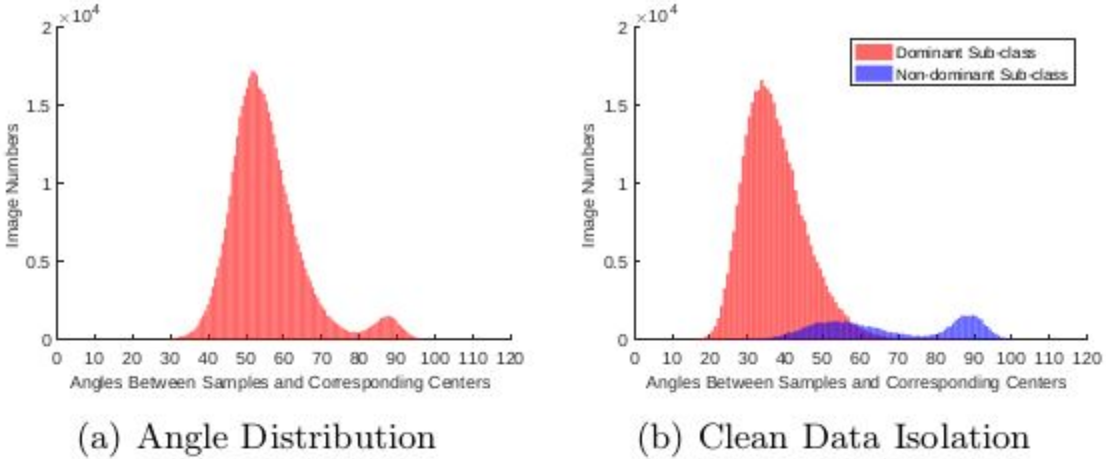
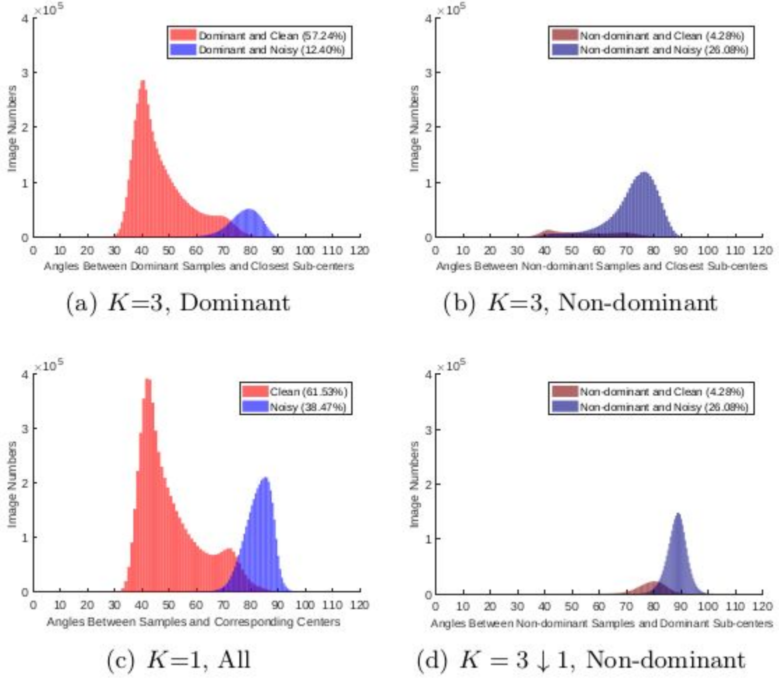


Fig. 3. (a) Angle distribution of samples to their corresponding centers predicted by the pre-trained ArcFace model [5]. Noise exists in the CASIA dataset [40,30]. (b) Angle distribution of samples from the dominant and non-dominant sub-classes. Clean data are automatically isolated by sub-center ArcFace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

MS1MV0 (raw) distribution

- estimated noise: 47.1% ~ 54.4%
- Training ResNet-50, MS1MV0, ArcFace
- “clean” and “noisy” are defined by MS1MV3 semi-automatic cleaned dataset.
- Sub-center ArcFace reduce noise from 38.47% to 12.40%.
- Angle threshold between 70 and 80 can be easily searched to drop most noisy samples.



(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 training flow

- Training Network using Sub-center ArcFace with K sub-classes, K>1.
- After enough discriminative power, we can clean the dataset by droping all no-dominant sub-classes, and using angle-threshold Øt.
- Retrain Network from scratch with the cleaned dataset, with no sub-classes, for example denoted as K = 3 ↓ 1

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

Adding synthtic noise and training ResNet-50 on clean MS1MV3.

open-set: 75% of Id’s remain clean, other are assigned with random labels. (outliers)

close-set: randomly select 25% images of each Id and assign random labels. (label-flips)

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

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

Ethnicity Bias ??

When combining all identities from MS1MV2 and Asian celebrities from DeepGlint, Arc-Face achieves the best identification performance 84.840%.

Method	Id (@FPR=1e-3)	Ver(@FPR=1e-9)
CASIA	26.643	21.452
MS1MV2	80.968	78.600
DeepGlint-Face	80.331	78.586
MS1MV2+Asian	**84.840** (1st)	80.540
CIGIT_IRSEC	84.234 (2nd)	**81.558** (1st)

Table 8. Identification and verification results (%) on the Trillion-Pairs dataset. ([Dataset*, ResNet100, ArcFace])

Future improvements

- Train with Imdb-Face dataset, large mannual cleaned.
- Train with larger datasets such as, CelebFace500K or other, using Sub-Center ArcFace to relax intra-classes noise.
- Scraping more data on search engines and imply ethnicity bias, regarding the mission.
- Adding our labeled data, to imply ethnicity bias.

Appendix - MS1MV3 (our dataset training)

Lightweight Face Recognition Challenge

Jiankang Deng [1] Jia Guo [1]
Debing Zhang [2] Yafeng Deng[2] Xiangju Lu [3] Song Shi [3]
[1]InsightFace [2]DeepGlint [3]IQIYI

- Cleaned from raw MS-Celeb-1M
- face images are pre-processed to the size of 112 × 112 by the five facial landmarks predicted by RetinaFace.
- Semi-automatic refinement by employing the pre-trained ArcFace model and ethnicity-specific annotators.
- Named also MS1M-RetinaFace, and contains 5.1M images of 93K identities.