Rethinking Robust Representation Learning Under Fine-grained Noisy Faces

Bingqi Ma[1*], Guanglu Song[1*], Boxiao Liu[1,2], and Yu Liu[1†]

[1] Sensetime Research
[2] SKLP, Institute of Computing Technology, CAS
{mabingqi,songguanglu}@sensetime.com, liuboxia@ict.ac.cn,
liuyuisanai@gmail.com

1 Terms and notations

We provide a holistic description of terms and notations in Table 1.

Table 1. Detailed description of terms and notations in our paper.

Terms/Notations	Meaning
Identity	Faces sharing *identity* means these images come from the same person.
Class	Faces annotated with the same label construct a *class*. There may be annotation errors in the class.
Cluster	If there are no less than two faces for an identity, these images build a meaningful *cluster*.
$\mu_{j,m_j}, \sigma_{j,m_j}$	The mean and standard deviation of the cosine similarity between the sub-center W_{j,m_j} and features.
λ_1	λ_1 aims to reduce the inter-class conflict.
λ_2	λ_2 controls the decision boundary of *producing* operation, which aims to reduce the intra-class conflict.
λ_3	λ_3 controls the decision boundary of *dropping* operation, which aims to reduce the intra-class conflict.
λ_4	λ_4 controls the decision boundary of *merging* operation, which aims to reduce the inter-class conflict.
M_j	Initial number of sub-centers for class j.

2 Synthetic Noisy Datasets

In our paper, we reformulate the noise type in each class with a more fine-grained manner as `N-identities`|$\mathtt{K}^C$`-clusters`. Different types of the noisy face can be generated by adjusting the values of N, K, and C. We will give a detailed description on the process of constructing synthetic noisy dataset.

– **Intra-class Conflict.** Intra-class conflicts arises from faces with multiple identities sharing the same label. We design a ingenious method to construct fine-grained intra-class noisy data. Specifically, we firstly sample a subset of IDs from MS1MV3 as clean data according to the noise ratio. For the rest data, we randomly group them into a maximum of 5 IDs, and then merge them into a random ID in the clean subset. Furthermore, we introduce up to 10 shallow faces into each ID to increase the grainedness of noisy data. Based on above strategy, we can obtain intra-class conflict dataset with $N \geq K \geq 1$, and $C = 0$ in each class.

*Equal contributions.
†Corresponding author.

Table 2. Experiments of different settings on synthetic noisy dataset comparing with state-of-the-art methods. The 1:1 verification accuracy (TAR@FAR) on IJB-C dataset is reported to evaluate the performance.

Method	Dataset	IJB-C		
		$1e{-}3$	**1e−4**	$1e{-}5$
ArcFace	Mixture of Noise	94.99	90.03	82.40
Sub-center ArcFace M=2	Mixture of Noise	93.92	89.17	81.24
Sub-center ArcFace M=3	Mixture of Noise	94.20	89.32	81.43
Sub-center ArcFace M=4	Mixture of Noise	93.63	88.19	81.10
Sub-center ArcFace M=5	Mixture of Noise	93.45	87.91	81.06
Sub-center ArcFace M=6	Mixture of Noise	93.58	88.02	80.72
SKH + ArcFace M=2	Mixture of Noise	96.23	94.93	92.46
SKH + ArcFace M=3	Mixture of Noise	96.74	95.05	92.59
SKH + ArcFace M=4	Mixture of Noise	96.89	95.16	92.71
SKH + ArcFace M=5	Mixture of Noise	96.34	94.87	92.43
SKH + ArcFace M=6	Mixture of Noise	95.96	94.52	92.35
ESL + ArcFace	Mixture of Noise	**97.62**	**96.22**	**93.60**

- **Inter-class Conflict.** Inter-class Conflict arise from faces with the same identity but assigned with multiple labels. We randomly sample a subset of images from each identity according to the noise ratio, which construct the clean subset. For the rest images, we split images with the same label to a maximum of 5 groups, and images in each group will be assigned with a new label, which does not conflict with current labels in clean subset. Then we can obtain inter-class conflict dataset with $N = K = 1$, and $C > 0$ in each class.
- **Mixture of Conflict.** Mixture of conflict contains both intra-class conflict and inter-class conflict. We randomly sample a subset of images from each identity according to the noise ratio, which construct the clean subset. For the rest images, we split images with the same label to a maximum of 5 groups, and images in each group will be assigned with a random label in the clean subset. Furthermore, we also introduce random shallow faces in each class. In this manner, we can obtain fine-grained mixture of conflict dataset with $N \geq K \geq 1$, and $C \geq 0$ in each class.

3 Compare with State-of-the-art Methods

For Sub-center ArcFace [1] and SKH [2], the sub-center number M of each class is an important hyperparameter, which has a significant effect on the performance. In Table. 2, we make a grid search on M for fair comparison with our proposed ESL on the mixture conflict dataset with 50% noise. Sub-center ArcFace [1] achieves best performance when M is set as 3, while M in SKH [2] should be set as 4. ESL can easily outperform them by a large margin, which further verify the robustness of ESL on dealing with fine-grained noisy faces.

Table 3. Experiments on different noise distributions. *GPU hours* represents the cost of re-searching hyper-parameters.

Dataset	Hyper-parameters	GPU hours	IJB-B	IJB-C	LFW	CFP-FP	AgeDB-30
50% Intra-conflict	Inherit from mixture noise dataset	-	94.71	96.37	99.78	98.46	98.12
50% Intra-conflict	Re-search on intra-conflict dataset	1040	**94.80**	**96.44**	**99.79**	**98.48**	**98.15**
50% Inter-conflict	Inherit from mixture noise dataset	-	94.43	96.11	99.59	**98.11**	97.98
50% Inter-conflict	Re-search on inter-conflict dataset	1040	**94.57**	**96.22**	**99.64**	98.10	**98.01**

Table 4. Experiments on WebFace dataset. *GPU hours* represents the cost of re-searching hyper-parameters. We use a total of 104 Nvidia A100 GPUs to search on WebFace dataset.

Method	Hyper-parameters	GPU hours	IJB-B	IJB-C	LFW	CFP-FP	AgeDB-30
ArcFace	-	-	89.01	91.43	99.65	97.52	97.53
Sub-center ArcFace M=3	-	-	92.18	94.35	99.75	97.82	97.94
SKH + ArcFace M=3	-	-	94.06	95.82	99.78	99.07	98.13
ESL + ArcFace	Inherit from MS1M	-	95.27	96.89	99.80	98.71	**98.34**
ESL + ArcFace	Re-search on WebFace	5200	**95.34**	**97.01**	**99.81**	**98.74**	98.32

4 Generalization of hyper-parameters

Intuitively, re-searching all of the hyper-parameters for each given training data will lead to optimal accuracy. However, this introduces a too heavy computational cost, especially for large-scale datasets. We argue that the hyper-parameters in ESL can directly generalize to different noise distributions and training datasets. This allows ESL to work flexibly with different training data without re-searching the hyper-parameters.

4.1 Generalization across noise distributions

As shown in Table 3, re-searching hyper-parameters will bring huge time cost but little performance improvement.

4.2 Generalization across training datasets

We conduct experiments on a sample subset of WebFace260M with 40M images as shown in Table 4. **ESL can also achieve outstanding performance on large scale dataset**.

5 Statistics value for MS1MV0

MS1MV0 contains 99743 classes and 9607705 images. After training with ESL, we keep 94837 sub-centers and 5296374 images in MS1MV0, which is similar to 93431 and 5179510 in MS1MV3 from iterative training and cleaning.

6 Experiments on query-based retrieval task

In Table 5, we further conduct experiments on Google Landmarks Dataset v2 (GLDv2) [4], which is collected from the web and contains lots of noise. We adopt a ResNet-50 as the backbone and follow the setting in [3]. Hyper-parameters in ESL follow the setting on MS1M.

Table 5. The peformance is evaluated on ROxford test set.

Method	ArcFace	Sub-center	SKH	ESL
mAP	53.64	54.21	54.98	**56.14**

References

1. Deng, J., Guo, J., Liu, T., Gong, M., Zafeiriou, S.: Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In: European Conference on Computer Vision. pp. 741–757. Springer (2020)
2. Liu, B., Song, G., Zhang, M., You, H., Liu, Y.: Switchable k-class hyperplanes for noise-robust representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3019–3028 (2021)
3. Mei, K., et al.: 3rd place solution to" google landmark retrieval 2020". arXiv preprint arXiv:2008.10480 (2020)
4. Weyand, T., et al.: Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In: CVPR (2020)