Switchable K-class Hyperplanes for Noise-Robust Representation Learning

Boxiao Liu[1,2,*], Guanglu Song[3,*], Manyuan Zhang[3,4], Haihang You[1,2,†], and Yu Liu[3]

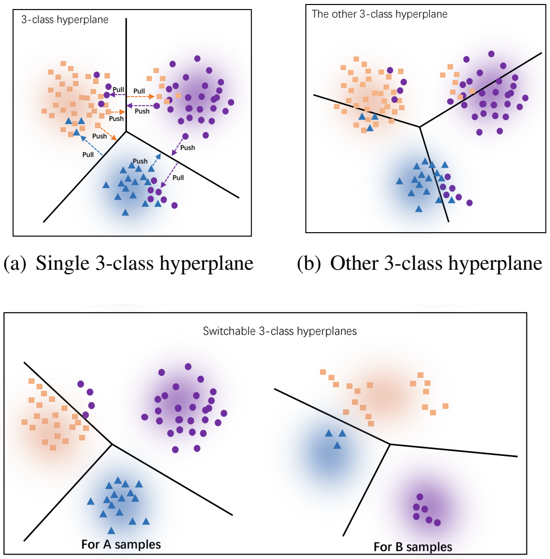
[1]State Key Laboratory of Computer Architecture, Institute of Computing Technology, CAS
[2]University of Chinese Academy of Sciences
[3]SenseTime Research
[4]CUHK - SenseTime Joint Lab, The Chinese University of Hong Kong
{liuboxiao,youhaihang}@ict.ac.cn,
{zhangmangyuan,songguanglu}@sensetime.com, liuyuisanai@gmail.com

Abstract

Optimizing the K-class hyperplanes in the latent space has become the standard paradigm for efficient representation learning. However, it's almost impossible to find an optimal K-class hyperplane to accurately describe the latent space of massive noisy data. For this potential problem, we constructively propose a new method, named Switchable K-class Hyperplanes (SKH), to sufficiently describe the latent space by the mixture of K-class hyperplanes. It can directly replace the conventional single K-class hyperplane optimization as the new paradigm for noise-robust representation learning. When collaborated with the popular ArcFace on million-level data representation learning, we found that the switchable manner in SKH can effectively eliminate the gradient conflict generated by real-world label noise on a single K-class hyperplane. Moreover, combined with the margin-based loss functions (e.g. ArcFace), we propose a simple Posterior Data Clean strategy to reduce the model optimization deviation on clean dataset caused by the reduction of valid categories in each K-class hyperplane. Extensive experiments demonstrate that the proposed SKH easily achieves new state-of-the-art on IJB-B and IJB-C by encouraging noise-robust representation learning. Our code will be available at `https://github.com/liubx07/SKH.git`.

1. Introduction

Optimizing the K-class hyperplane in the latent space to encourage intra-class compactness and inter-class discrepancy has become the standard paradigm for efficient representation learning. Benefit from it, million-level face recognition has achieved remarkable improvement in recent years [21, 17, 26, 6, 25]. From the early CASIA-WebFace [31] to more recent MegaFace [15], MSCeleb-1M [11] and Celeb500K [3], the growing scale of training dataset introduces more complex data distribution and also inevitably introduces real-world noise. This leads to that it's almost impossible to find an optimal K-class hyperplane to accurately describe the latent space with massive noisy data. We explore this potential problem by a latent space with



(a) Single 3-class hyperplane (b) Other 3-class hyperplane

(c) *Switchable 3-class hyperplanes*

Figure 1. Illustration of the 3-class hyperplane in latent space of noisy data. (a) Conventional single 3-class hyperplane. (b) Try to find another 3-class hyperplane. (C) The proposed *Switchable 3-class hyperplanes*.

[*] Equally-contributed.
[†] Corresponding author.

3 categories as shown in Fig. 1(a). Optimizing a single 3-class hyperplane in this latent space is extremely difficult and the gradient conflict generated by the local training samples collapses the hyperplane optimization. Fig. 1(b) attempts to optimize some local outliers but biases other samples. For this potential problem, we constructively propose a new method, named *Switchable K-class Hyperplanes* (SKH), to sufficiently describe the latent space with massive noisy data. We introduce the mixture of K-class hyperplanes and optimize them by a switchable manner as shown in Fig. 1(c). To sufficiently describe the latent space in Fig. 1(c), we can adopt a simple greedy mechanism based on loss value to perform K-class hyperplane switch.

According to the aforementioned analysis, given a random latent space, we can always describe it better than traditional training paradigm by assigning multiple K-class hyperplanes. We found that directly replacing the conventional K-class hyperplane by SKH can be effective for noise-robust representation learning. There are two types of conflict caused by label noise, *intra-class conflict* (different identities with same label ID) and *inter-class conflict* (same identity with different label IDs). In latent space, the intra-class compactness and inter-class discrepancy is significant to improve the robustness of representation. We take the conventional single K-class hyperplane to demonstrate the influence of label noise on latent space optimization. For different identities with same ID in *intra-class conflict*, we optimize the latent space by maximizing the inner product of its feature and a single class center. It's hard to provide the efficient hyperplane to compact them as shown in Fig. 1(a). For same identity with different IDs in *inter-class conflict*, we will minimize the inner product of features in one ID and class center of the other ID. However, they represent a same identity and forcing them apart makes the hyperplane unreliable.

Although the presence of label noise complicates latent space, the proposed *Switchable K-class Hyperplanes* can efficiently alleviate this issue. It encourages different identifies with the same ID to select different K-class hyperplanes to avoid *intra-class conflict*. Also, the *inter-class conflict* can be eliminated by assigning samples of different IDs (actually the same identity) to optimize different hyperplanes. This makes them independent of each other and avoids generating gradient conflict on the same hyperplane as shown in Fig. 1(c). Extensive experimental results demonstrate that the proposed SKH provides a new state-of-the-art for noise-robust representation learning. Moreover, for noise-robust representation learning, we can further clean the training data by dropping intra high-confident noise samples and merging inter samples with high-similar centers in different K-class hyperplanes. After this, we can effectively improve intra-class compactness and inter-class discrepancy, and achieve comparable performance compared to the model trained on the manually cleaned dataset.

To sum up, the contribution of this papers is threefold:

(1) *A novel Switchable K-class Hyperplanes* - We introduce a mixture of K-class hyperplanes with a switchable manner to better describe the latent space with complex noisy data. It can effectively improve the robustness of noisy data training.

(2) *A posterior data clean strategy* - We can further clean the training data jointly considering intra-class noise and inter-class noise. This process is easy to perform and effectively improves performance.

(3) *Superior performance on noise-robust representation learning in face recognition* - We apply SKH to different types of label noise and conduct extensive experiments to thoroughly evaluate its superiority to other methods. It provides a new state-of-the-art for noise-robust representation learning.

2. Related Works

Datasets for Face recognition. The training of the modern face recognition model heavily relies on large-scale datasets. Many existing studies [24, 11, 3, 1, 31, 4] confirm that the performance of face recognition model improves with the growth of the training dataset scale. MS1M [11] is the first million-scale public face recognition dataset. It consists of 10M faces of 100K celebrities. Celeb-500K [3] is another large-scale face dataset, containing over 500K identities. However, most of those face recognition datasets are scratched from the Internet by a pre-collected celebrity list, and they tend to be noisy [24, 6]. For example, the noise-rate of MS1M is around 50%. Many works [30, 24, 1] explore recursive training and cleaning processes to build noise-controlled datasets. However recursive process is time-consuming and it is proven that there still exists noise [6, 5]. While some works rely on human labors to clean datasets [24], however, it is not realistic for datasets with tens of millions of samples.

Loss functions for Face recognition. The core problem for building an excellent face recognition model is to make the model generate discriminative features, which means intra-class compactness and inter-class separation. Many popular works [26, 6, 17, 21, 23, 17] tend to utilize well-designed loss functions to help the model to form discriminate feature. Contrastive loss [23] and triplet loss [21] are proposed to increase Euclidean margin for more discriminative feature embeddings. Specifically, they force the Euclidean distance between instances from different classes larger than those from the same class by a large margin. However, the number of tuple and triplet grows exponentially as the dataset becomes larger.

Liu *et al.* [17] brings about a new perspective to this problem. They pioneer A-softmax which considers the last fully connected layer inner product as a space project operation and the weight of the linear transformation matrix could represent the corresponding class anchor. Then a multiplicative angular margin penalty is proposed to encourage intra-class compactness. Deng *et al.* [6] directly adds an additive margin penalty in the angular space, which has a clear geometric interpretation and easy to be implemented. Even though these margin-based loss functions have achieved remarkable success, they all rely on clean training datasets, which is unpractical in reality. Designing more effective loss functions to handle datasets with noise has recently drawn much attention in face recognition.

Face recognition with noise. Learning from imperfect annotations [2, 20, 16, 28, 9, 10] has become an important research area when facing massive data. As for face recognition, there are several works focus on alleviating the negative impact of noisy data [32, 27, 5, 13, 8]. Zhong *et al.* [32] design a noise-resistant loss function that combines original labels and predicted labels of current model to form a hypothetical training labels. However, the quality of hypothetical labels is heavily limited by the performance of the initial model. Wang *et al.* [27] introduces a co-mining strategy which trains two networks simultaneously and uses the loss value to re-weight the training examples. The twin-network design avoids error accumulation. However, the co-mining is unpractical for training with huge networks on large-scale datasets since it doubles the training cost. Deng *et al.* [5] introduces multiple centers for each class to alleviate the intra-class noise conflict, while the inter-class conflict still limits the model performance. In this paper, multiple K-class hyperplane with a greedy switching mechanism and a practical data clean strategy are introduced to handle the intra-class and inter-class noise simultaneously.

3. Preliminary Understanding

3.1. A Unified Formulation of Loss Functions

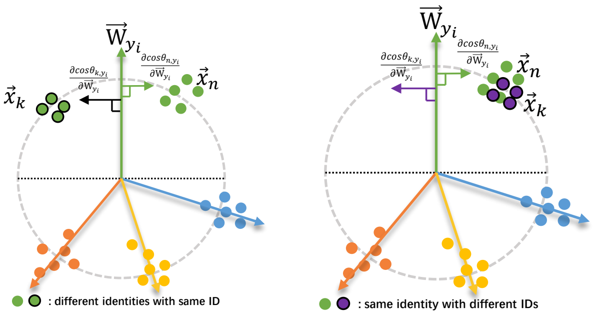
In face recognition tasks, we can formulate a unified equation to represent variant loss functions as:

$$\mathcal{L}(\vec{x}_i) = -logP_{i,y_i} = -log\frac{e^{\hat{f}_{i,y_i}}}{e^{\hat{f}_{i,y_i}} + \sum_{j\neq y_i} e^{f_{i,j}}}, \quad (1)$$

where i is the index of samples in the current batch data and y_i represents the label ID of sample I_i. Assuming that the vector $\vec{x}_i$ denotes the feature representation of a face image I_i and $\vec{W}_j$ indicates the j-th class center. The logit $\hat{f}_{i,y_i}$ and $f_{i,j}$ can be formulated as:

$$\hat{f}_{i,y_i} = s \cdot [m_1 \cdot \cos(\theta_{i,y_i} + m_2) - m_3], \quad (2)$$

$$f_{i,j} = s \cdot \cos(\theta_{i,j}), \quad (3)$$


$\vec{\mathrm{W}}_{y_i}$
$\frac{\partial cos\theta_{k,y_i}}{\partial \vec{\mathrm{W}}_{y_i}}$
$\frac{\partial cos\theta_{n,y_i}}{\partial \vec{\mathrm{W}}_{y_i}}$
$\vec{x}_k$
$\vec{x}_n$
● ○ : different identities with same ID
$\vec{\mathrm{W}}_{y_i}$
$\frac{\partial cos\theta_{k,y_i}}{\partial \vec{\mathrm{W}}_{y_i}}$
$\frac{\partial cos\theta_{n,y_i}}{\partial \vec{\mathrm{W}}_{y_i}}$
$\vec{x}_n$
$\vec{x}_k$
● ● : same identity with different IDs

(a) Intra-class conflict

(b) Inter-class conflict

Figure 2. Illustration of the label noise conflict in latent space. $\frac{\partial \cos\theta_{*,y_i}}{\partial \vec{W}_{y_i}}$ is the gradient in the direction of the tangent line. The intra-class conflict and inter-class conflict disrupt the optimization of the model.

$$\cos\theta_{i,j} = \frac{\left\langle \vec{x}_i, \vec{W}_j \right\rangle}{\|\vec{x}_i\|_2 \left\|\vec{W}_j\right\|_2}, \quad (4)$$

where s, m_1, m_2, m_3 are hyperparameters. ArcFace can be obtained by assigned $m_1 = 1, m_3 = 0$ and $m_2 = 0.5$. In order to improve the intra-class compactness and inter-class discrepancy, the model needs to enlarge P_{i,y_i}, i.e., enlarge $\hat{f}_{i,y_i}$ and reduce $f_{i,j}$.

3.2. Analysis of Intra-class and Inter-class Conflict

To make the analysis clear, we take ArcFace as a basic loss function to explore the influence of label noise on model optimization.

At the backward propagation stage, the gradient of $\vec{W}_{y_i}$ in ArcFace is calculated as

$$\frac{\partial\mathcal{L}(\vec{x}_i)}{\partial\vec{W}_{y_i}} = (P_{i,y_i} - 1)\nabla_{\cos\theta_{i,y_i}}\hat{f}_{i,y_i} \cdot \frac{\partial\cos\theta_{i,y_i}}{\partial\vec{W}_{y_i}}. \quad (5)$$

As $\nabla_{\cos\theta_{i,y_i}}\hat{f}_{i,y_i}$ and $(P_{i,y_i} - 1)$ are always scalars, the gradient of $\vec{W}_{y_i}$ follows the same direction of $\frac{\partial\cos\theta_{i,y_i}}{\partial\vec{W}_{y_i}}$, and

$$\frac{\partial\cos\theta_{i,y_i}}{\partial\vec{W}_{y_i}} = \frac{1}{\left\|\vec{W}_{y_i}\right\|_2}(\mathbf{x}_i - \cos\theta_{i,y_i} \cdot \mathbf{W}_{y_i}), \quad (6)$$

where $\mathbf{x}_i$ and $\mathbf{W}_{y_i}$ means normalized vectors. Thus, the direction of gradient is perpendicular to the direction of W_{y_i}. Similarly, we have that the gradient $\frac{\partial\mathcal{L}(\vec{x}_i)}{\partial\vec{W}_j}, j \neq y_i$, is perpendicular to W_j.

According to the aforementioned derivation, we can easily understand the effect of *intra-class conflict* and *inter-class conflict* on model optimization as shown in Fig. 2. For *intra-class noise*, we consider two samples $\vec{x}_k$ and $\vec{x}_n$ belonging to different identities but with the same ID. The gradients of class center W_{y_i} generated by samples $\vec{x}_k$ and $\vec{x}_n$

are contradictory in the direction of the tangent line (perpendicular to the direction of W_{y_i}), and so are for *inter-class noise*. This makes the training process unstable and easy to collapse the K-class hyperplane in latent space. In order to eliminate the effects of label noise, we need to jointly eliminate the both intra-class conflict and inter-class conflict to conduct a noise-robust training method.

4. The Proposed Approach

In this paper, we are committed to eliminating the label noise conflict via a simple and effective manner that can be directly plugged into any loss functions. The switchable manner ensures stability and will significantly improve their robustness to label noise.

Inspired by the observation stated in the previous section, we propose a novel learning mechanism *Switchable K-class Hyperplanes* (SKH) to wisely avoid intra-class conflict and inter-class conflict. We will first introduce our proposed *Switchable K-class Hyperplanes* and then give a deep analysis to better understand its effectiveness and robustness. Finally, we conduct extensive experiments to evaluate its performance on real-world noise and variant synthetic noise.

4.1. Switchable K-class Hyperplanes

Let the K-class hyperplane represents the final classification layer. At the training stage, we construct M K-class hyperplanes $\mathcal{F} = \{f^m | m \in \{1, \ldots, M\}\}$ at the end of the backbone. We initialize the weight matrix $W^m \in \mathbb{R}^{D \times C}$ for each anchor f^m independently. D, C indicates the feature dimension and identity number. In one training step, let $\vec{x}_i$ denotes the feature vector of sample I_i, and $\mathcal{L}^m(\vec{x}_i)$ indicates the loss of sample I_i calculated with anchor f^m. Note that f^m is dynamically selected for each training sample and we employ the greedy manner to control the selection. The loss L_{CE} of sample I_i can be calculated as:

$$\mathcal{L}_{CE}(\vec{x}_i) = \mathcal{L}^{\arg\min_m \mathcal{L}^m(\vec{x}_i)}(\vec{x}_i). \tag{7}$$

This indicates that after the forward and backward propagation, only one f^m will be updated and provide gradient to the backbone for each sample. To be clear, we demonstrate the pipeline of SKH in Fig. 3. In this manner, SKH is able to capture the complex distribution of the whole training data with potential label noise, no matter for intra-class noise or inter-class noise. The optimization target is to minimize the loss $\mathcal{L}_{CE}$ over the whole training data. To achieve this, it tends to switch the samples of different identities with the same ID or the same identity with different IDs to different f^m. Therefore, the gradients $\frac{\partial \cos\theta_{n,y_i}}{\partial \vec{W}_{y_i}}$ and $\frac{\partial \cos\theta_{k,y_i}}{\partial \vec{W}_{y_i}}$ in Fig. 2 are substituted to $\frac{\partial \cos\theta_{n,y_i}}{\partial \vec{W}_{y_i}^{m1}}$ and $\frac{\partial \cos\theta_{k,y_i}}{\partial \vec{W}_{y_i}^{m2}}$, where $m_1, m_2 \in \{1, ldots, M\}$ and $m_1 \neq m_2$. $\vec{W}_{y_i}^{m_1}$ and $\vec{W}_{y_i}^{m_2}$ denote the class centers in f^{m_1} and f^{m_2} for ID y_i, respectively. This potential switchable manner effectively eliminate the *intra-class conflict* and *inter-class conflict*.

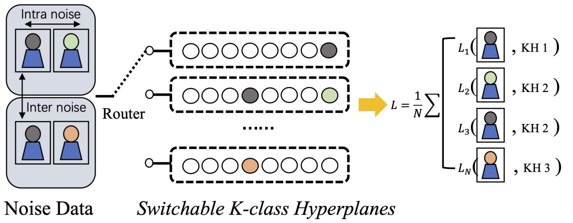


Figure 3. The pipeline of *Switchable K-class Hyperplanes*. Only one K-class hyperplane in SKH will be activated by the specific training sampling to alleviate the *intra-class conflict* and *inter-class conflict*.

4.2. Posterior Data Clean Strategy

When applied to the practical optimization, we observed that the clean samples, hard samples and noise samples will automatically converge to their respective weight centers. Even though *Switchable K-class Hyperplanes* can effectively improve the robustness under label noise, the intra-class compactness and inter-class discrepancy still suffer an inevitable decline. The clean samples and hard samples are separated into different weight centers in different anchors. The clean samples and noise samples still do. This may become more serious when the dataset only contains few noise images.

Based on this observation, we propose a straightforward strategy to recapture intra-class compactness and inter-class discrepancy after the standard training process. For each class, we only preserve the weight center with the largest number of samples between M anchors. Let $\vec{W}_i$ and $\vec{W}_j$ denote the preserved weight centers of i-th and j-th ID, respectively. The similarity of $\vec{W}_i$ and $\vec{W}_j$ is calculated by:

$$sim(\vec{W}_i, \vec{W}_j) = \frac{\left\langle \vec{W}_i, \vec{W}_j \right\rangle}{\left\| \vec{W}_i \right\|_2 \left\| \vec{W}_j \right\|_2}. \tag{8}$$

We introduce the constant angle threshold $\mathcal{T}_1, \mathcal{T}_2$. We drop the samples when $sim(\vec{W}_{y_i}, \vec{x}_i) \leq \mathcal{T}_1$ and merge samples belonging to $\vec{W}_i$ and $\vec{W}_j$ when $sim(\vec{W}_i, \vec{W}_j) \geq \mathcal{T}_2$. After that, we can retrain the model based on the cleaned dataset.

4.3. Robustness and Effectiveness Analysis

We give a more intuitive illustration to understand the training process based on noise samples. The first figure in Fig. 4 describes the initial state of the model with 3 hyperplanes. Each sample will get similar loss value with different f^m and the greedy selection strategy now is more like random selection. After a few steps of training, the

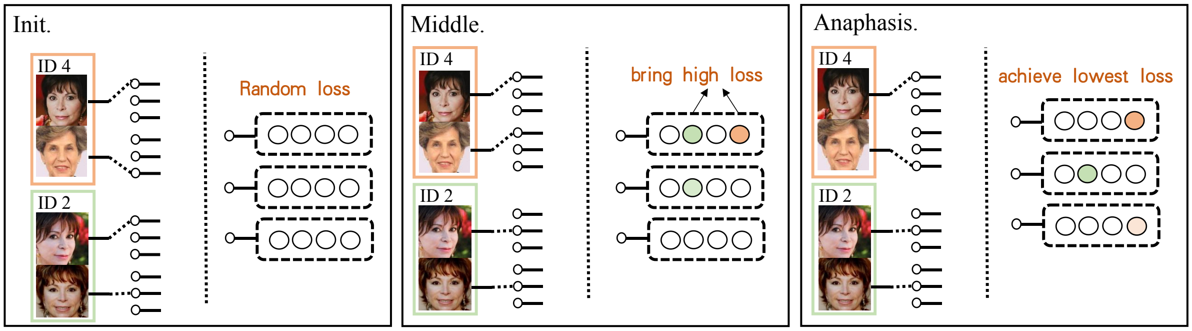


Figure 4. An intuitive illustration to understand the training process based on noise samples. The second sample in ID 4 is intra-class noise and the inter-class noise samples in ID 2 have the same identity as ID 4. These three figures indicate different training stages.

inter-class noise samples will obtain high loss when activating the same hyperplanes. Therefore, some of them will be switched to others with the lowest loss as shown in the second figure in Fig. 4. Finally, the optimal state can be achieved by separating the noise samples to different anchors to eliminate the gradient conflict on the class center.

Except for the switchable manner with the lowest loss, we can also consider other different strategies. (1) Adopt the largest loss manner to switch the sample to different anchors. This forces noise samples to disturb the class centers dominated by clean samples and collapse the model optimization. (2) Adopt max pooling on the class-wise cosine similarity as done in sub-center ArcFace [5]. This can effectively dispose the intra-class conflict but intensify the inter-class conflict. It will force to generate the gradient conflict as shown in Fig. 2(b). On the contrary, the proposed *Switchable K-class Hyperplanes* can jointly alleviate the influence of intra-class and inter-class noise, and can be combined with variant loss functions.

5. Experiments

5.1. Experimental Settings

Datasets The datasets for training in our experiments include MS1MV0 (about 10M images of 100K identities) , MS1MV2 (about 5.8M faces of 85K identities), and MS1MV3 (about 5.1M faces of 91K identities). MS1MV0 is the original version of dataset proposed in [11], and it is shown that the proportion of noise is about 50%. MS1MV2 and MS1MV3 are both semi-automatically cleaned from MS1MV0, introduced in [6] and [7] respectively. MS1MV2, also known as “emore”, is popular in face recognition research, and MS1MV3 is a strengthened version of MS1MV2. For evaluation, we mainly employ IJB-B [29] and IJB-C [18] as the testing datasets and follow the standard test protocol. Moreover, the results of our final result on LFW [14], CFP-FP [22], and AgdDB-30 [19] are reported.

Implementation Details We follow ArcFace [6] to get the aligned face crops and resize them into (112×112). Then, a ResNet-like [12] network R50 is used to extract representation and returns a 512-D embedding for each image. For all the experiments in this paper, we set the learning rate as 0.1 at the start of training and downscale it by 0.1 at 100K, 160K, and 220K iterations. The training process ends after 240K iterations. We also utilize a warm-up of learning rate from 0.01 in the first 12K iterations. The weight decay is set to 5e-4 and the momentum of the SGD optimizer is 0.9. We train the network on 8 NVIDIA V100 GPUs, with a total batch size of 512. The experiments are implemented by PyTorch.

5.2. Comparison with the State-of-the-art

We conduct experiments with ResNet-50 to investigate the proposed *Switchable K-class Hyperplanes* based on the noisy MS1MV0. In Tab. 1, the performance of vanilla ArcFace is decreased by a large margin compared to training on the clean data of MS1MV3. As analyzed in Fig. 2, the serious potential intra-class conflict and inter-class conflict in MS1MV0 greatly hinder the optimization of the network. With the proposed *Switchable K-class Hyperplanes*, the performance of basic ArcFace is remarkably improved under massive noise.

We further compare our method with the state-of-the-art methods, including co-mining [27] and sub-center ArcFace. As shown in Tab. 1, the proposed SKH outperforms existing works with a significant margin, which demonstrates the effectiveness of our method on noisy data. To be specific, when training on the noisy MS1MV0, SKH outperforms co-mining for 1.43% and sub-center for 1.53% under TAR@FAR=1e-4 on IJB-C testset. The IJB-B testset has shown a similar trend. The superiority of our method is mainly due to the delicate switchable strategy with multiple K-class hyperplanes, which can handle intra-class noise and inter-class noise simultaneously.

Method	Dataset	IJB-B			IJB-C		
		1e-5	1e-4	1e-3	1e-5	1e-4	1e-3
ArcFace	MS1MV0	78.88	89.24	93.86	85.49	91.82	95.15
sub-center ArcFace $M = 3$	MS1MV0	85.62	91.70	94.88	90.59	93.72	95.98
co-mining [27]	MS1MV0	85.57	91.80	94.99	90.71	93.82	95.95
NT [13]	MS1MV0	85.56	91.57	94.79	90.48	93.65	95.86
NR [32]	MS1MV0	85.53	91.58	94.77	90.41	93.60	95.88
SKH + ArcFace $M = 3$	MS1MV0	**89.34**	**93.50**	**95.89**	**93.00**	**95.25**	**96.85**
SKH + ArcFace $M = 2$	MS1MV0	88.12	93.04	95.44	91.92	94.63	96.47
SKH + ArcFace $M = 3$	MS1MV0	**89.34**	**93.50**	**95.89**	**93.00**	**95.25**	**96.85**
SKH + ArcFace $M = 4$	MS1MV0	89.04	93.47	95.77	92.99	95.11	96.72
SKH + ArcFace $M = 5$	MS1MV0	88.88	93.50	95.74	92.71	95.04	96.70
sub-center ArcFace $M = 3 \downarrow 1$	MS1MV0	89.40	94.56	96.49	94.03	95.92	97.40
SKH + ArcFace $M = 3 \downarrow 1$	MS1MV0	**90.71**	**94.82**	**96.55**	**94.18**	**96.26**	**97.53**
ArcFace	MS1MV2	88.64	94.41	96.16	93.63	95.81	97.23
ArcFace	MS1MV3	90.74	95.04	96.66	94.66	96.44	97.64

Table 1. Ablation experiments of different settings on MS1MV0 and MS1MV3. The 1:1 verification accuracy (TAR@FAR) is used as the evaluation metric on IJB-B and IJB-C. $M = 3 \downarrow 1$ indicates we perform the posterior data clean strategy.

Method	$\mathcal{T}_1$	$\mathcal{T}_2$	IJB-B	IJB-C
SKH $M = 3 \downarrow 1$	0.70	-	94.72	96.12
SKH $M = 3 \downarrow 1$	0.75	-	94.80	96.22
SKH $M = 3 \downarrow 1$	0.75	0.8	**94.82**	**96.26**
SKH $M = 3 \downarrow 1$	0.75	0.9	94.80	96.15

Table 2. Experiments with different choices of $\mathcal{T}_1$ and $\mathcal{T}_2$. We adopt the 1:1 verification TAR (@FAR=1e-4) on the IJB-B and IJB-C dataset as the evaluation metric.

Method	Dataset	IJB-B	IJB-C
ArcFace	IA-MS1MV3	93.84	95.31
SKH + ArcFace	IA-MS1MV3	**94.67**	**96.06**
ArcFace	IT-MS1MV3	86.96	90.69
SKH + ArcFace	IT-MS1MV3	**94.81**	**95.04**

Table 3. Experiments on variant label noises. We adopt the 1:1 verification TAR (@FAR=1e-4) on the IJB-B and IJB-C dataset as the evaluation metric.

5.3. Ablation Study

Exploration on Hyperparameters. The *Switchable K-class Hyperplanes* is effective but simple to implement, with only three hyper-parameters: the anchor number M, the constant threshold $\mathcal{T}_1$ and $\mathcal{T}_2$. We first conduct several experiments to explore the sensitivity of performance to the value of M. The number of switchable hyperplanes M plays an important role of balance between hard samples and noise samples. If M is too small, the hyperplanes are not capable of isolating all noise samples. If M is too large, the intra-class and inter-class variance in each hyperplane is reduced, leading to degenerated discriminability. On MS1MV0 dataset, we conduct ablation experiments on M, shown in Tab. 1, and find $M = 3$ achieve the best performance. More hyperplanes do not lead to better performance.

We further examine the influence of different $\mathcal{T}_1$ and $\mathcal{T}_2$ and show the results in Tab. 2. We find that the combination of $\mathcal{T}_1 = 0.75$ and $\mathcal{T}_2 = 0.8$ is the best, while the performance is actually insensitive to these hyperparameters. In the following experiments, we use $M = 3$, $\mathcal{T}_1 = 0.75$ and $\mathcal{T}_2 = 0.8$ as the default setting.

Effectiveness on Synthetic Noise. *Switchable K-class Hyperplanes* is designed for noise-robust representation learning. We elaborately construct synthetic noisy datasets to explore the robustness of SKH on different types of label noise. We employ the MS1MV3, cleaned multiple times by a semi-automatic manner, to establish *intra-class noise*, *inter-class noise* and *mixture of noise*.

1. For *intra-class noise*, we merge the images from one identity into another. Specifically, we change the labels of images belonging to identity y_i into $y_i - 1$ if y_i is a even number. After the merging process, we get a noisy dataset with ~45K different IDs.

2. For *inter-class noise*, we split the images of one identity into two different IDs. Specifically, we randomly select half of the images belonging to y_i, and change the labels into $y_i + n$, where n is the number of identities of original MS1MV3. Then we get a noisy dataset with ~180K different IDs.

3. For *mixture of noise*, we further generate both of intra-class noise and inter-class noise to simulate the real world situation. Firstly, we keep half of images in each

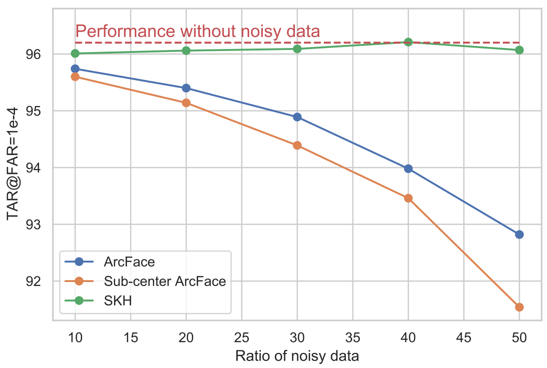


Figure 5. Evaluation of SKH with different ratios of noisy data in Mix-MS1MV3.

ID to form a clean subset. Then, the rest images for each identity will be randomly split into two groups, which are given two random labels respectively. It is noteworthy that we can control the number of images inserted into the clean subset to achieve various noise ratio.

The results on intra-class noise and inter-class noise are shown in Tab.3. The original ArcFace on both of them results in a significant performance drop. By contrast, the proposed SKH effectively enhances the robustness towards variant label noises.

For mixture of noise, as shown in Fig. 5, SKH surpasses ArcFace and sub-center by a large margin, especially when the noise ratio getting bigger. When the noise ratio increases, the performance of SKH almost keeps the same, and can even increase a little. However, the performance of ArcFace and sub-center drops by an increasing margin. This indicates that the proposed SKH is more robust for noise-tolerant face recognition training.

Posterior Data Clean. Based on the discriminative power of the converged SKH, we can further clean the data to filter out some noise samples by an offline manner. The related experiments with "$M = 3 \downarrow 1$" in Tab. 1 demonstrate the results. The offline data cleaning can further improve the performance. By performing only once, the cleaned dataset leads to a superior performance than MS1MV2, which using elaborate semi-automatic cleaning.

Note that, the proposed SKH has a slight performance gap compared with ArcFace when training on MS1MV3. However, with the proposed posterior data clean strategy, we can catch up and even achieve better performance when compared with MS1MV3. The results on several popular testsets are shown in Table 4. Furthermore, we carefully look over the data cleaning process on MS1MV3. Surprisingly, we find that there are about 2.4K identities are overlap with each other, and we show several examples in Fig. 6. This explains why the cleaned dataset based on MS1MV3 can result in better performance than the original one, and demonstrates the effectiveness of our methods further more.

Method	IJB-B	IJB-C	LFW	CFP-FP	AgeDB-30
ArcFace	95.04	96.44	99.83	98.57	98.12
SKH m=3	93.50	95.25	99.78	98.59	98.23
SKH m=3$\downarrow$1	94.98	**96.48**	99.77	**98.70**	**98.25**

Table 4. Experiments on MS1MV3 cleaned by the proposed posterior data clean strategy, showing superior performance than the original one. We adopt the 1:1 verification TAR (@FAR=1e-4) on the IJB-B and IJB-C dataset as the evaluation metric.

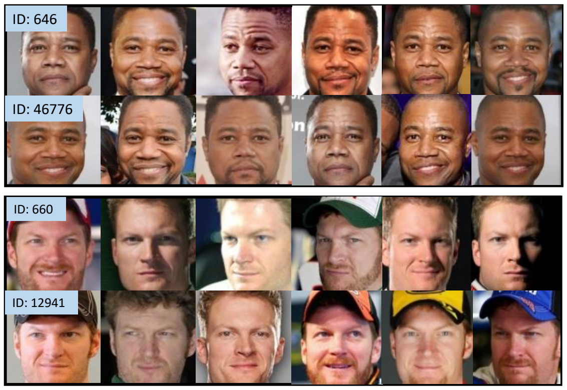


Figure 6. An illustration of the inter-class noise in the semi-automatically refined MS1MV3. The ID at the left-top of the images indicates the given label in MS1MV3.

Method	CosFace	sub-center	SKH
IJB-C	92.16	94.02	**94.92**
IJB-B	89.25	92.21	**93.02**

Table 5. Experiments with CosFace. We adopt the 1:1 verification TAR (@FAR=1e-4) on the IJB-B and IJB-C dataset as the evaluation metric.

Generalization on Other Loss Function. We further examine the generalization ability of SKH on CosFace [26] loss, another popular loss other than ArcFace. Comparison of the results in Table 5 demonstrates that our method can still surpass ArcFace and sub-center by a significant margin.

Discussion. To understanding the reason behind the effectiveness of our method, we analyze the training process in detail. We first give an intuitive illustration to show the evolution of feature space. Specifically, we visualize the changes of features of samples and anchors in different training stages to demonstrate how SKH separates the intra-class noise samples and inter-class noise samples to different *anchors*. As in Fig. 7, at the initial stage of training, both

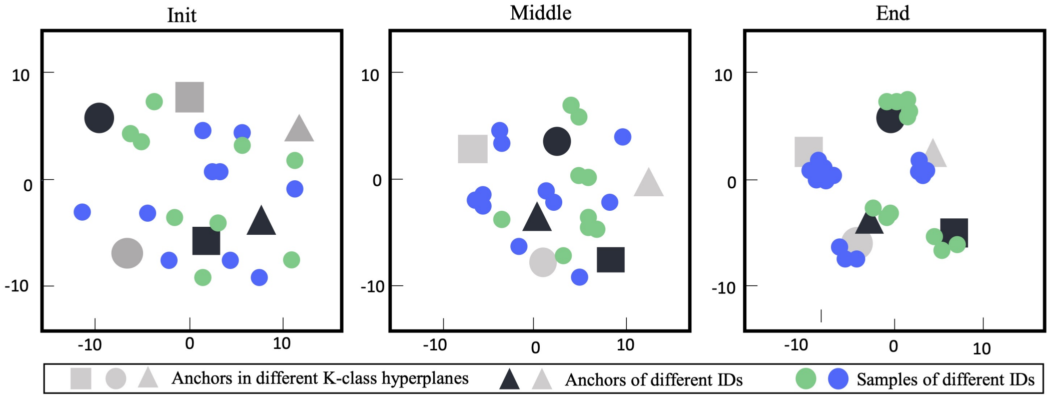


Figure 7. Visualization of the evolution of feature space in SKH at different training stages. SKH can separate the intra-class noise samples and inter-class noise samples to different anchors.

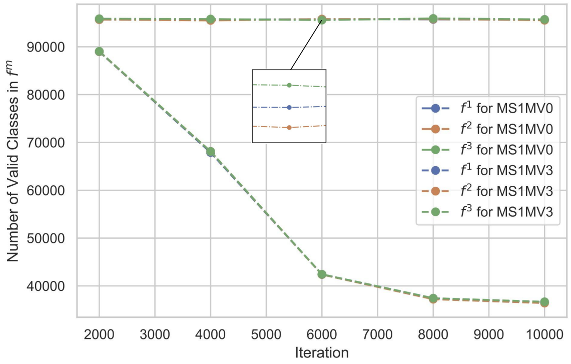


Figure 8. The number of valid classes along the training process for MS1MV0 and MS1MV3 datasets.

training samples and anchors are scattered randomly over the feature space. Along with the training process, the samples and anchors begin to converge under the greedy selection strategy. When the training completes, the intra-class noise samples and inter-class noise samples are switched to different *anchors*.

Furthermore, we count the number of valid classes, i.e., the number of anchors that have been selected as the class center by at least one sample, and plot the results on MS1MV0 and MS1MV3 at different training epochs. We can see that, for MS1MV0, the number of valid classes is close to 99743, the number of all classes, along the training process. This indicates that the proposed SKH learns to effectively utilize most of anchors in 3 K-class hyperplanes to handle the intra-class and inter-class noise. On the contrary, the number of valid classes on MS1MV3 decreases significantly as training to about one third of the of identitie number. This shows that our SKH learns to gather most of the intra-class samples together randomly with one of the three class centers. The decrease of valid classes helps to explain the performance gap between SKH and ArcFace on MS1MV3. We verify this explanation by training SKH with greater margin in ArcFace, and find that the performance increases as the margin equals 0.6, which results in inferior performance in standard ArcFace on MS1MV3.

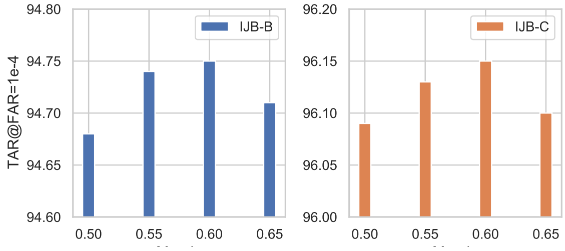


Figure 9. The result of SKH trained on MS1MV3 with different values of margin.

6. Conclusion

In this work, we present *Switchable K-class Hyperplanes* to better describe the latent space under massive noisy data. It can directly replace the conventional single K-class hyperplane as the new paradigm for noise-robust representation learning. When collaborated with the popular ArcFace on million-level data representation learning, we found that the switchable manner in SKH can effectively eliminate the gradient conflict generated by real-world label noise. A posterior data clean strategy is further introduced to refine the noisy dataset. Extensive experiments on noisy data training demonstrate the effectiveness of SKH and it provides a new state-of-the-art for noise-robust representation learning.

Acknowledgments

This research is partially sponsored by the grant from the National Key Research and Development Program (2018YFC1406200) and Natural Science Foundation of China Grant 41930110.

References

[1] Xiang An, Xuhan Zhu, Yang Xiao, Lan Wu, Ming Zhang, Yuan Gao, Bin Qin, Debing Zhang, and Ying Fu. Partial fc: Training 10 million identities on a single machine. *arXiv preprint arXiv:2010.05222*, 2020.

[2] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016.

[3] Jiajiong Cao, Yingming Li, and Zhongfei Zhang. Celeb-500k: A large training dataset for face recognition. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2406–2410. IEEE, 2018.

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[5] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *European Conference on Computer Vision*, pages 741–757. Springer, 2020.

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[7] J. Deng, J. Guo, D. Zhang, Y. Deng, X. Lu, and S. Shi. Lightweight face recognition challenge. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2638–2646, 2019.

[8] Yuanyuan Ding, Yongbo Cheng, Xiaoliu Cheng, Baoqing Li, Xing You, and Xiaobing Yuan. Noise-resistant network: a deep-learning method for face recognition under noise. *EURASIP Journal on Image and Video Processing*, 2017(1):1–14, 2017.

[9] Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, and Dongrui Fan. C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9834–9843, 2019.

[10] Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, and Dongrui Fan. Utilizing the instability in weakly supervised object detection. *arXiv preprint arXiv:1906.06023*, 2019.

[11] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016.

[12] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[13] Wei Hu, Yangyu Huang, Fan Zhang, and Ruirui Li. Noise-tolerant paradigm for training face recognition cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11887–11896, 2019.

[14] G. Huang, M. Mattar, Tamara L. Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. 2008.

[15] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4873–4882, 2016.

[16] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5447–5456, 2018.

[17] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[18] Brianna Maze, J. Adams, J. A. Duncan, Nathan D. Kalka, T. Miller, Charles Otto, Anil K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother. Iarpa janus benchmark - c: Face dataset and protocol. *2018 International Conference on Biometrics (ICB)*, pages 158–165, 2018.

[19] Stylianos Moschoglou, A. Papaioannou, Christos Sagonas, Jiankang Deng, I. Kotsia, and S. Zafeiriou. Agedb: The first manually collected, in-the-wild age database. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1997–2005, 2017.

[20] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952, 2017.

[21] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[22] S. Sengupta, Jun-Cheng Chen, Carlos D. Castillo, V. Patel, R. Chellappa, and D. Jacobs. Frontal to profile face verification in the wild. *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016.

[23] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1891–1898, 2014.

[24] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 765–780, 2018.

[25] Feng Wang, Jian Cheng, Weiyang Liu, and H. Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25:926–930, 2018.

[26] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface:

Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.

[27] Xiaobo Wang, Shuo Wang, Jun Wang, Hailin Shi, and Tao Mei. Co-mining: Deep face recognition with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9358–9367, 2019.

[28] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1568–1576, 2017.

[29] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, J. Adams, T. Miller, Nathan D. Kalka, Anil K. Jain, J. A. Duncan, Kristen Allen, J. Cheney, and P. Grother. Iarpa janus benchmark-b face dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 592–600, 2017.

[30] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.

[31] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[32] Yaoyao Zhong, Weihong Deng, Mei Wang, Jiani Hu, Jianteng Peng, Xunqiang Tao, and Yaohai Huang. Unequal-training for deep face recognition with long-tailed noisy data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7812–7821, 2019.