DDLN: Detecting and Discarding Label Noise for Noise-Robust Face Recognition

Fanglong Wu[1], Youqiang Gui[1], and Peng Cheng[1,2]*

[1] National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu, China
[2] School of Aeronautics and Astronautics, Sichuan University, Chengdu, China
fanglongwu@foxmail.com, chengpeng@scu.edu.cn

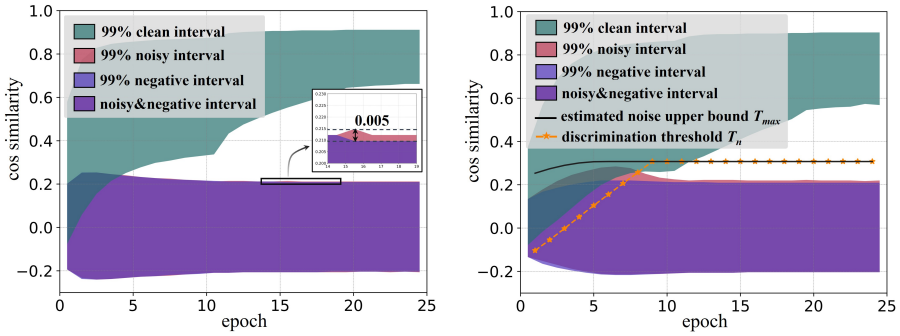
Abstract. Label noise is a major challenge in large-scale supervised face recognition, where weak or automatic annotations often introduce errors that mislead training. To address this, we propose a noise robust framework that performs noise detection and sample selection directly in cosine-similarity space. We observe that the non-target cosine similarities of clean samples share a highly consistent distribution profile with the target similarities of unfitted mislabeled samples. This phenomenon can be explained from a backpropagation perspective and provides a cue for monitoring label noise during training. Based on this observation, we formulate noise detection as boundary estimation in similarity space. Specifically, we track the upper bound of high-confidence clean non-target similarities to determine the filtering threshold, without requiring prior knowledge of the noise rate or auxiliary networks. We further introduce a progressive rule during early training, where the threshold gradually increases from the estimated noise lower bound to the upper bound. This process discards unreliable samples while retaining hard but clean samples. Extensive experiments on eight synthetic and three real-world noisy datasets demonstrate that our method achieves superior noise detection and state-of-the-art recognition accuracy, with only about 0.3% average measured overhead. The filtered dataset produced by our method is also beneficial for subsequent training. Source code is available at https://github.com/wfl95/DDLN.

Keywords: Face Recognition · Robust Face Recognition · Label Noise Detection · Sample Selection · Noise Robust Learning

1 Introduction

Large-scale face recognition systems increasingly rely on data collected from the open web [13,59,68]. While web-crawling drastically reduces annotation costs, it inevitably introduces label noise. Such erroneous supervision corrupts training dynamics, propagates uncertainty, and severely degrades model generalization. Therefore, developing scalable and lightweight models robust to label noise is crucial for real-world deployments.

* Corresponding author.



(a) simulated discarding of all noisy pairs **(b)** noisy pairs discarded by our method

Fig. 1: Left: Distributions of clean $\cos\theta_y$ (clean interval), noisy $\cos\theta_y$ (noisy interval), and clean $\cos\theta_{j,j\neq y}$ (negative interval) under a simulated setting where all noisy pairs are discarded. The maximum difference between the main distribution ranges of clean $\cos\theta_{j,j\neq y}$ and noisy $\cos\theta_y$ is only 0.005, indicating strong distributional consistency during training. We further quantify this consistency using KL divergence and Wasserstein distance in Supplementary Material. **Right:** Training distributions obtained when applying our method. Pairs with $\cos\theta_y$ below the threshold T_n (orange dashed line) are treated as noisy and excluded from backpropagation. In early training stages, T_n gradually moves from the lower to the upper bound of the clean $\cos\theta_{j,j\neq y}$ distribution, allowing sufficient learning of hard clean pairs while progressively discarding noise.

A dominant strategy to mitigate this issue is sample selection, largely motivated by the memorization dynamics of deep networks [1,37]: models tend to fit clean, generalizable patterns early in training before memorizing noisy instances. However, existing selection methods often introduce significant overhead or rely on brittle assumptions. For instance, sub-center [7,25] and co-mining [14,45] approaches introduce substantial computational and memory footprints, making them less suitable for large-scale systems. Other dynamic thresholding methods, such as the OTSU-based estimation in RVFace [44], heavily depend on the assumption of a clear bimodal loss distribution. Unfortunately, this bimodal structure is often absent during early training stages or collapses entirely under severe noise rates, rendering the estimated boundaries unstable. This highlights a critical open challenge: *how to robustly and efficiently identify mislabeled samples across diverse datasets without relying on prior noise-rate knowledge or heavy auxiliary modules.*

To address this, we shift our focus to the intrinsic training dynamics of softmax-based losses. We observe that sample–incorrect-label pairs behave similarly to sample–non-target-class pairs, both reflecting a mismatch between the sample and the label. From a backpropagation perspective, both similarities operate in a weak-gradient regime, leading to close alignment between the non-target similarity distribution of clean samples and the target-similarity distribution of mislabeled samples. This insight provides an interpretable, data-driven cue: the observable distribution of high-confidence clean non-target similarities

can serve as a stable boundary for noise detection without a noise-rate prior or auxiliary estimation network.

Building on this property, we propose a lightweight, progressive sample selection framework to Detect and Discard Label Noise (DDLN). As illustrated in Fig. 1, rather than applying a rigid threshold that might incorrectly discard hard-but-informative clean samples, DDLN adopts a dynamic scheduling strategy. During early training epochs, the discrimination threshold gradually tightens from a conservative lower bound toward the estimated statistical upper bound. This mechanism allows the network to sufficiently learn hard but clean samples in early training, while progressively filtering label noise as the similarity distributions stabilize. Once excluded, these noisy pairs rarely cross the estimated noise upper bound again. The main contributions of this work are summarized as follows:

1. We theoretically and empirically reveal a consistent distributional alignment between clean non-target similarities and unfitted noisy target similarities. Leveraging this, we formulate a robust, data-driven noise boundary estimation that operates without prior knowledge of the noise rate.
2. We propose DDLN, a progressive sample selection strategy that gradually increases the filtering threshold from the lower bound to the upper bound of the estimated noise distribution, progressively discarding noisy samples while preserving hard but clean samples.
3. Extensive experiments on eight synthetic and three real-world noisy face benchmarks show that DDLN improves both noise detection and recognition performance. The method adds about 0.2% measured overhead in our ResNet50–CosFace setting and produces a cleaned training subset beneficial for downstream training.

The remainder of this article is organized as follows. Section 2 reviews related work. Section 3 introduces background concepts and analyzes the relationship between noisy and negative pairs from a backpropagation perspective. Section 4 details the proposed DDLN framework. Section 5 presents experimental evaluations on various noisy benchmarks. Finally, Section 6 concludes the paper.

2 Related Work

2.1 Face Recognition

Deep neural networks have significantly advanced face recognition over the past decade [12, 43]. These improvements are supported by large-scale datasets [3, 4, 9, 23, 24, 56, 68], powerful backbone architectures [11, 15, 17], and discriminative loss functions [8, 27, 28, 40, 42].

Modern systems predominantly adopt normalized softmax-based losses, which enforce intra-class compactness and inter-class separability in the embedding space [12,43]. Margin-based formulations such as SphereFace [26,27], CosFace [40, 42], and ArcFace [8] further enhance discrimination by introducing angular margins. Subsequent works extend these designs with adaptive margins conditioned

on image quality [22,58], pose [29,54], age [16,60], or global consistency [61,67]. However, when trained on large web-collected datasets, these models remain vulnerable to label noise.

2.2 Learning from Noisy Data

Learning under noisy supervision has been studied through robust architectures [14,49], loss functions [18,65], and training strategies. For example, LightCNN [49] suppresses noisy activations using Max-Feature-Map operations, while SFace [65] employs gradient rescaling on hypersphere manifolds. Other approaches estimate clean probabilities for sample reweighting [18] or jointly address label noise and long-tail distributions [66]. Although effective, these methods mainly rely on soft weighting, which may accumulate errors during training [39].

Another line of work focuses on sample selection, which explicitly removes suspected noisy samples. Co-mining [45] identifies reliable samples via cross-network validation but requires multiple networks. Sub-Center ArcFace [7] models intra-class variations with multiple sub-centers and removes noisy samples using fixed thresholds. Later works further explore geometric modeling in latent space [25,30]. Other methods estimate noise boundaries dynamically using training statistics [18,44,64]. Noise-Tolerant Face Recognition [18] estimates sample reliability from target-angle statistics and applies soft weights, while RVFace [44] estimates a threshold from the target-class cosine distribution using an OTSU-based bimodality assumption. Pleiss et al. [34] introduce auxiliary noise classes for boundary estimation. These approaches, however, often depend on fixed or heuristic thresholds, distributional assumptions, or additional model complexity.

In contrast, our approach estimates the noise boundary from the non-target similarity distribution of clean samples, rather than relying on target-similarity thresholds. This enables progressive noise filtering without a noise-rate prior, a clean validation set, or auxiliary networks.

3 Background and Analysis

This section provides the background and analysis required to motivate our noise detection strategy. We briefly review softmax-based face recognition (Sec. 3.1) and robust learning under noisy supervision (Sec. 3.2), then present a backpropagation-based explanation of the observed distributional consistency (Sec. 3.3). Detailed derivations are deferred to the supplementary material.

3.1 Softmax-based Face Recognition

Modern face recognition systems commonly adopt ℓ_2-normalized softmax losses [12, 35,38,41]. With normalized features and class weights, the objective becomes a cosine-similarity softmax on a hypersphere:

$$\mathcal{L}(x_i, y_i) = -\log \frac{e^{sf(\theta_{i,y_i})}}{e^{sf(\theta_{i,y_i})} + \sum_{j \neq y_i}^{C} e^{sf(\theta_{i,j})}}, \tag{1}$$

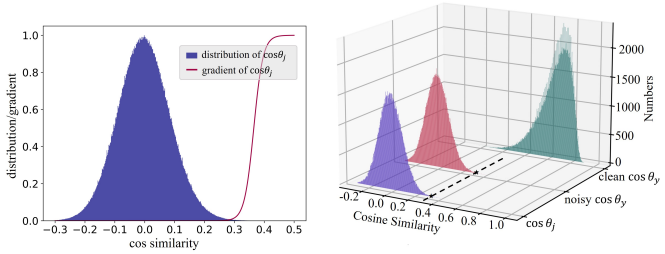


Fig. 2: Left: Non-target similarities $\cos\theta_{i,j}$ $(j \neq y_i)$ and their normalized gradient magnitude $P_{i,j}$ during training. $P_{i,j}$ is near-zero over the main density region. **Right:** Cosine-similarity distributions at epoch 5. Discarded noisy target similarities (red) closely align with clean non-target similarities (blue).

where $\theta_{i,j}$ is the angle between x_i and class prototype W_j, s is a scaling factor, and $f(\theta) = \cos\theta$. Margin-based variants (SphereFace/CosFace/ArcFace) introduce angular or additive margins on the target logit and serve as our baseline objectives [8, 27, 42].

3.2 Robust Face Recognition

Large-scale web data inevitably contains label noise [39, 66], which injects erroneous supervision into the target class and degrades recognition performance. A common formulation assigns a confidence weight $w_i \in [0, 1]$ to each training pair:

$$\mathcal{L}_{robust}(x_i, y_i) = -w_i \log \frac{e^{sf(\theta_{i,y_i})}}{e^{sf(\theta_{i,y_i})} + \sum_{j \neq y_i}^{C} e^{sf(\theta_{i,j})}}. \tag{2}$$

Soft reweighting may accumulate errors when confidence estimation is inaccurate. Binarizing $w_i \in \{0, 1\}$ yields *sample selection*, which excludes suspected noisy pairs from backpropagation [44, 45]. Our method follows this paradigm and focuses on a stable, data-driven criterion to determine the selection mask.

3.3 Backpropagation Analysis

We explain why the cosine-similarity distribution of discarded noisy targets $\cos\theta_{i,y_i}$ closely matches that of clean non-target pairs $\cos\theta_{i,j}$ $(j \neq y_i)$. Although arising from different pairs, both measure the similarity between a sample and an incorrect identity, and both operate under weak gradient coupling.

For Eq. (1), the gradient w.r.t. a non-target similarity is

$$\frac{\partial \mathcal{L}}{\partial \cos\theta_{i,j}} = sP_{i,j} = \frac{s\, e^{sf(\theta_{i,j})}}{e^{sf(\theta_{i,y_i})} + \sum_{k \neq y_i} e^{sf(\theta_{i,k})}}, \quad j \neq y_i, \tag{3}$$

$$\left| \frac{\partial \mathcal{L}}{\partial \cos\theta_{i,(k)}} \right| = sP_{i,(k)} \leq \min\left(\frac{s}{k},\, se^{-s\Delta_{1,k}}\right), \tag{4}$$

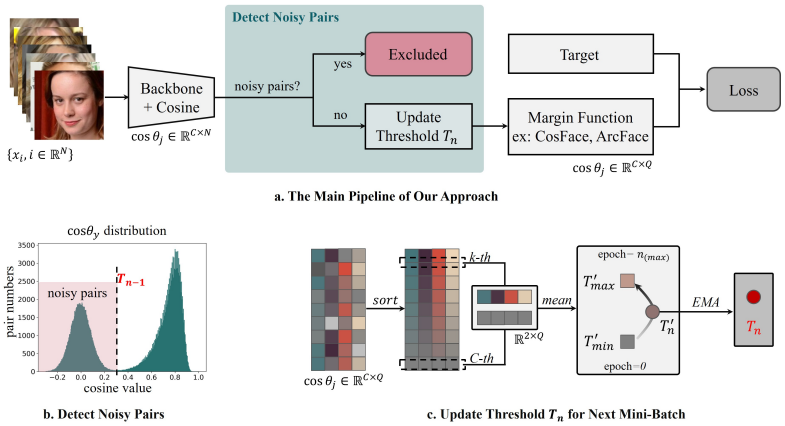


Fig. 3: Overview of Detecting and Discarding Label Noise (DDLN). **(a)** A detector excludes suspected noisy pairs from loss/backpropagation via a dynamic threshold T_{n-1} in the $\cos\theta_y$ domain. **(b)** The binary mask q_i is determined by comparing $\cos\theta_{i,y_i}$ with T_{n-1}. **(c)** For each mini-batch, a local threshold T'_n is estimated from retained pairs' non-target similarities and smoothed into T_n by EMA.

where $P_{i,j}$ is the softmax probability. As shown in Fig. 2 (left), for most non-target classes this probability is close to zero, implying negligible gradients. Formally, $P_{i,(k)}$ admits simple rank- and gap-based upper bounds (see Supplementary Material), indicating that meaningful gradients concentrate on a small set of hardest negatives, while the majority of non-target similarities are effectively saturated.

For discarded noisy pairs, once the incorrect target class is excluded from backpropagation by our criterion, the gradient-driven attraction between x_i and the wrong prototype vanishes. Therefore, both (i) clean non-target similarities and (ii) discarded noisy target similarities are governed by weak coupling, leading to closely matched similarity statistics (Fig. 2 (right)). This provides the motivation for using the clean non-target similarity distribution to estimate noise boundaries in our method.

4 Proposed Approach

Building on the distributional insights from Section 3, we propose a robust framework to Detect and Discard Label Noise (DDLN). As illustrated in Fig. 3, DDLN introduces a dynamic discrimination threshold T_n in the cosine-similarity space to progressively filter unreliable samples during backpropagation.

Overall Objective. DDLN augments standard margin-based softmax losses with a binary mask q_i. For a sample x_i with the assigned label y_i, the robust

objective is formulated as:

$$\mathcal{L}_{\text{DDLN}} = -q_i \log \frac{e^{sf(\theta_{i,y_i})}}{e^{sf(\theta_{i,y_i})} + \sum_{j \neq y_i} e^{sf(\theta_{i,j})}}. \tag{5}$$

The binary mask q_i determines whether the sample contributes to the gradient update, which is governed by the dynamic threshold T_{n-1} from the previous step:

$$q_i = \begin{cases} 1, & \cos\theta_{i,y_i} \geq T_{n-1}, \\ 0, & \cos\theta_{i,y_i} < T_{n-1}. \end{cases} \tag{6}$$

Dynamic Threshold Estimation.

To operate without a noise-rate prior, we anchor the filtering threshold to observable clean non-target similarities. Let $\mathcal{Q}_n = \{i \mid q_i = 1\}$ be the retained set in the current mini-batch. For each $i \in \mathcal{Q}_n$, we sort its non-target similarities $\{\cos\theta_{i,j}\}_{j \neq y_i}$ in descending order, denoting the k-th largest value as $\cos\theta_{i,(k)}$. Thus, top-k is computed per sample along the non-target class dimension; the mini-batch only averages these per-sample statistics. We establish the batch-wise bounds as follows:

$$T'_{\max} = \frac{1}{|\mathcal{Q}_n|} \sum_{i \in \mathcal{Q}_n} \cos\theta_{i,(k)} + \alpha, \quad T'_{\min} = \frac{1}{|\mathcal{Q}_n|} \sum_{i \in \mathcal{Q}_n} \min_{j \neq y_i} \cos\theta_{i,j}. \tag{7}$$

Here, $T'_{\max}$ estimates the average upper tail of the clean non-target distribution. We use the shared defaults k=10, α=0.02, and λ=0.01 without dataset-specific retuning. The moderate rank k=10 is less sensitive to extreme outliers than the hardest negative, while the slack α reduces filtering of borderline hard-clean samples.

Progressive Scheduling. Deep networks tend to fit clean samples faster than noisy ones [1]. To properly leverage this progressive separability, DDLN does not apply the strict $T'_{\max}$ immediately. Instead, we implement a linear interpolation between the bounds:

$$T'_n = T'_{\min} + \beta_n \left(T'_{\max} - T'_{\min}\right), \quad \beta_n = \frac{\min(n, n_{\max})}{n_{\max}}, \tag{8}$$

where n is the current training step. We set $n_{\max}$ to finish before the first learning-rate decay. This linear schedule starts conservatively near $T'_{\min}$, retaining hard-but-informative clean pairs, and progressively tightens to $T'_{\max}$ as the similarity distributions stabilize.

Finally, to stabilize the decision boundary across mini-batches, we smooth the local threshold T'_n into the global threshold T_n via Exponential Moving Average (EMA) [20]:

$$T_n = (1 - \lambda) T_{n-1} + \lambda T'_n, \tag{9}$$

where $\lambda = 0.01$. Once noisy pairs drop below this tightening boundary, they lose direct gradient attraction toward their incorrect targets and are unlikely to recover after exclusion, which stabilizes the purification process. Detailed analyses of recovery behavior and hyperparameter sensitivity are provided in the Supplementary Material.

5 Experiments

We evaluate DDLN from three perspectives: (i) the accuracy of its label noise detection, (ii) its impact on downstream face recognition under noisy supervision, and (iii) its computational overhead.

Datasets and Metrics. For simulated noise experiments, we use the manually cleaned CASIA-WebFace [56] as a clean source and inject mixed synthetic noise (label flips and outliers) at rates from 10% to 80%. For real-world evaluation, we use datasets with naturally occurring noise at varying severity: CASIA-WebFace (low noise, ∼10%), MS-Celeb-1M [13] (moderate noise, ∼50%), and a 2-million subset of WebFace260M [68] named WebFace2M-Noise (extreme noise, ∼80%). We evaluate downstream recognition performance across nine standard benchmarks, including the Avg-5 metric (LFW [19], AgeDB30 [32], CFP-FP [36], CALFW [63], CPLFW [62]), IJB-B/C (TAR@FAR) [31, 47], SLLFW [10], BLUFR, and the low-quality TinyFace [6] dataset. To assess noise detection quality on simulated data, we report Precision, Recall, F1-score, False Positive Rate (FPR), and False Negative Rate (FNR).

Implementation Details. We integrate DDLN into three representative margin-based baselines: ArcFace [8], CosFace [42], and SphereFace [27]. We adopt SEResNet50-IR for simulated experiments and standard ResNet-18/50/100 architectures for large-scale real-world datasets. All models are trained from scratch using standard SGD optimization. Comprehensive details regarding dataset synthesis protocols, metric definitions, and exact hyperparameter configurations are provided in Supplementary Material.

5.1 Simulated Noisy Datasets Comparison

We compare DDLN against three reference conditions: (1) None (training directly on noisy labels), (2) Clean (an oracle setting where all noisy pairs are perfectly removed), and (3) RVFace [44] (a representative OTSU-based dynamic thresholding method). All methods are retrained under identical backbone and optimization settings.

Noise Detection Evaluation. As shown in Table 1, DDLN achieves consistently high F1-scores across different noise rates and corruption types. In seven out of nine settings, the average F1-score exceeds 99.1%. On the representative #40-20-20 setting, DDLN reaches 99.42% F1, substantially reducing detection errors compared to RVFace. We also evaluate robustness under identity-wise noise imbalance (Table 2). When per-identity noise ratios vary between 0% and 80% while keeping the overall flip rate at 40%, DDLN maintains nearly identical detection quality (F1: 99.11 vs. 99.22) and comparable recognition accuracy, indicating strong stability under uneven noise distributions.

To better understand this detection quality, Table 3 compares the False Positive Rate (FPR) and False Negative Rate (FNR) of clean samples against the OTSU-based RVFace. Across the 10% to 60% noise range, DDLN reduces the FNR by 70% to 84% relative to the baseline. At 60% noise, DDLN reduces the FNR from 7.26% to 1.24%, below RVFace's FNR at 20% noise (1.29%). Under

Table 1: Validation and noise-detection results (%) on simulated noisy datasets using SEResNet50-IR. Configurations follow `total-outliers-flips` (%). BLUFR reports TAR at FAR=10^{-3}, 10^{-4}, and 10^{-5}; Avg averages SLLFW and the three BLUFR scores. Precision, Recall, and F1-score treat noisy samples as positive. *Clean* is an oracle reference. Bold numbers indicate the better result between OTSU [44] and DDLN.

Setting	Methods	SLLFW	BLUFR (TAR@FAR)			Avg	Prec.	Rec.	F1
			10^{-3}	10^{-4}	10^{-5}				
	None	96.86	97.87	94.40	81.33	92.61	–	–	–
10-5-5	Clean	97.20	98.24	95.48	83.20	93.53	–	–	–
(CosFace)	OTSU	96.92	98.14	95.21	**85.37**	**93.91**	90.51	96.41	93.36
	DDLN	**97.15**	**98.33**	**95.59**	82.76	93.46	**97.04**	**99.64**	**98.32**
	None	95.62	97.22	92.72	81.82	91.84	–	–	–
20-10-10	Clean	96.80	98.12	95.20	86.85	94.24	–	–	–
(CosFace)	OTSU	96.97	98.09	94.82	84.10	93.50	94.98	97.84	96.39
	DDLN	**97.32**	**98.23**	**95.53**	**84.85**	**93.98**	**98.85**	**99.34**	**99.10**
	None	92.62	92.75	82.81	65.49	83.42	–	–	–
40-20-20	Clean	96.43	97.83	94.05	81.02	92.33	–	–	–
(CosFace)	OTSU	95.70	97.31	92.71	80.06	91.44	94.74	98.58	96.62
	DDLN	**96.45**	**98.02**	**95.01**	**84.04**	**93.38**	**99.14**	**99.70**	**99.42**
	None	83.10	68.27	49.98	33.68	58.76	–	–	–
60-30-30	Clean	95.33	97.04	92.49	82.37	91.81	–	–	–
(CosFace)	OTSU	93.35	94.09	85.42	67.62	85.12	95.05	98.00	96.50
	DDLN	**95.15**	**96.88**	**92.14**	**80.29**	**91.12**	**99.17**	**99.14**	**99.16**
	None	70.40	27.24	15.46	8.68	30.45	–	–	–
80-40-40	Clean	92.75	94.60	86.63	70.59	86.14	–	–	–
(CosFace)	OTSU	83.92	69.72	51.47	35.15	60.06	92.65	**98.30**	95.39
	DDLN	**91.70**	**88.69**	**75.42**	**56.63**	**78.11**	**95.16**	98.13	**96.63**
	None	90.40	88.33	74.83	55.71	77.32	–	–	–
40-0-40	Clean	96.83	98.18	95.61	87.58	94.55	–	–	–
(CosFace)	OTSU	96.15	97.21	92.88	80.72	91.74	94.30	96.93	95.59
	DDLN	**96.67**	**98.19**	**95.71**	**86.97**	**94.38**	**99.11**	**99.34**	**99.22**
	None	93.65	94.38	85.17	66.36	84.89	–	–	–
40-40-0	Clean	95.65	97.49	93.22	80.48	91.71	–	–	–
(CosFace)	OTSU	95.48	96.79	91.35	78.76	90.59	94.79	98.34	96.53
	DDLN	**95.80**	**97.53**	**93.15**	**80.56**	**91.76**	**99.59**	**98.71**	**99.15**
	None	92.38	92.76	82.75	65.81	83.42	–	–	–
40-20-20	Clean	96.38	97.59	93.37	80.36	91.93	–	–	–
(**ArcFace**)	OTSU	95.48	97.15	92.45	79.04	91.03	96.06	98.64	97.33
	DDLN	**96.50**	**98.10**	**95.16**	**85.40**	**93.79**	**99.20**	**99.47**	**99.34**
	None	92.27	91.82	80.58	62.77	81.89	–	–	–
40-20-20	Clean	96.60	97.76	94.14	82.25	92.69	–	–	–
(**Sphere**)	OTSU	95.47	96.48	91.13	77.84	90.23	94.88	95.45	95.17
	DDLN	**96.43**	**97.61**	**93.75**	**82.96**	**92.69**	**99.35**	**98.95**	**99.15**

the extreme 80% noise setting, DDLN still discards 19.96% of clean samples, although this is lower than RVFace's 31.19% (an absolute reduction of 11.23%).

Table 2: Performance (%) under identity-wise noise imbalance with DDLN (CosFace). `ID-normal` applies a uniform 40% flip rate across all identities (#40-0-40). `ID-imbalanced` randomly samples per-identity flip rates from 0% to 80% while keeping the overall dataset flip rate at 40%.

Datasets	SLLFW	BLUFR			Avg	Precision	Recall	F1-score
		10^{-3}	10^{-4}	10^{-5}				
ID-normal	96.67	**98.19**	**95.71**	**86.97**	**94.38**	**99.11**	99.34	**99.22**
ID-imbalanced	**96.70**	97.92	94.62	85.08	93.58	98.65	**99.53**	99.11

Table 3: False Positive Rates (FPR, %) and False Negative Rates (FNR, %) of clean samples under the *clean-as-positive* convention across different noise rates. The total noise is split evenly between outliers and label flips (1:1). Lower values indicate better performance (↓).

Noise rate (%)	FPR ↓		FNR ↓	
	OTSU [44]	**DDLN (Ours)**	OTSU [44]	**DDLN (Ours)**
10	3.59	**0.36**	1.12	**0.34**
20	2.16	**0.66**	1.29	**0.29**
40	1.42	**0.30**	3.65	**0.58**
60	2.00	**0.86**	7.26	**1.24**
80	**1.70**	1.87	31.19	**19.96**

This residual error is partly associated with the extremely limited clean supervision, as fewer than 13 clean samples per identity remain on average. Additional training dynamics (Precision/Recall/FPR/FNR curves) are provided in Supplementary Material.

Face Recognition Evaluation. Table 1 reports the final recognition performance using the Avg metric. Across 10% to 40% noise levels, DDLN consistently narrows the gap between noisy training and the oracle Clean setting. Under #40-20-20, DDLN outperforms RVFace by 1.94%, 2.76%, and 2.46% with CosFace, ArcFace, and SphereFace, respectively. At 60% noise (#60-30-30), DDLN nearly closes the gap to the Clean baseline. Under the extreme #80-40-40 setting, it improves Avg from 30.45% (None) to 78.11%, exceeding RVFace by 18.05%. This improvement is consistent with the lower clean-sample FNR, indicating that DDLN better preserves hard but informative clean samples.

Finally, extended comparisons in Supplementary Material show that DDLN achieves recognition accuracy comparable to heavy co-training [14, 51] and sub-center [7] paradigms under 80% extreme noise, but crucially operates within a highly efficient single-network pipeline without requiring exact prior noise rates.

Table 4: Performance (%) of noise-robust models and loss functions on CASIA-WebFace [56], evaluated with ResNet50 on IJB-C (TAR@FAR=10^{-3}, 10^{-4}, 10^{-5}) [31]. **Left:** advanced margin-based loss optimization methods. **Right:** representative noise-robust training methods.

Methods	10^{-3}	10^{-4}	10^{-5}
CosFace [42]	92.60	85.77	67.97
ElasticFace [2]	92.75	85.81	64.23
AdaFace [22]	91.89	80.30	43.29
UniFace [67]	–	88.96	78.40
GFace [61]	–	89.57	80.79

Methods	10^{-3}	10^{-4}	10^{-5}
NR [66]	90.65	82.76	70.20
Co-mining [45]	90.88	82.44	71.17
Sub-center [7]	92.47	87.03	79.16
DTDD [64]	92.87	87.69	80.39
DDLN	**94.07**	**89.22**	**82.29**

Table 5: Performance (%) of various noise-robust models on Webface2M-noise [68], evaluated with ResNet-50. R1, R5, R10 denote Rank-1, Rank-5, Rank-10 respectively.

Methods	LFW	CALFW	CPLFW	AgeDB	CFP	TinyFace			Avg
						R1	R5	R10	
CosFace [42]	98.58	90.50	82.37	85.97	86.07	50.40	58.37	63.89	77.02
-Sub [7]	97.95	89.63	80.50	84.96	82.89	50.99	57.51	62.42	75.86
-SubRe [7]	99.05	92.52	85.07	90.43	89.21	54.29	60.46	65.56	79.57
-RV [44]	**99.46**	93.83	88.32	93.90	93.49	61.72	66.87	70.57	83.52
-DDLN	99.43	**94.46**	**89.25**	**94.47**	**94.33**	**63.52**	**68.13**	**71.83**	**84.43**
ArcFace [8]	98.52	90.65	82.60	86.13	86.61	50.86	57.64	63.17	77.02
-Sub [7]	98.50	90.88	81.30	86.80	83.57	49.17	55.98	60.60	75.85
-SubRe [7]	99.20	93.27	85.08	91.18	89.11	53.43	59.98	64.97	79.53
-RV [44]	99.32	93.98	88.88	93.13	93.34	61.86	67.19	70.57	83.53
-RobustFace [53]	98.42	90.72	82.32	85.67	85.11	53.14	60.06	64.94	77.55
-DDLN	**99.53**	**94.57**	**90.18**	**95.10**	**95.10**	**63.25**	**68.05**	**71.86**	**84.70**
SphereFace [26]	98.08	89.58	81.33	84.07	84.96	51.50	58.88	63.49	76.49
-Sub [7]	97.88	88.80	79.73	83.70	82.34	49.41	56.52	61.96	75.04
-SubRe [7]	98.88	91.95	84.13	90.12	88.57	53.67	60.27	64.81	79.05
-RV [44]	99.07	92.78	87.48	91.72	92.60	60.72	66.58	70.31	82.66
-DDLN	**99.52**	**94.08**	**88.68**	**93.67**	**94.50**	**63.06**	**68.67**	**72.32**	**84.31**

5.2 Real Noisy Datasets Comparison

We evaluate DDLN on CASIA-WebFace [56], MS-Celeb-1M [13], and WebFace2M-Noise [68] to verify whether our purification logic generalizes to real-world, web-scale noise distributions.

CASIA-WebFace Table 4 reports the IJB-C verification performance (TAR at FAR). Although CASIA-WebFace [56] inherently contains a relatively low noise rate, integrating DDLN into the CosFace baseline still yields significant TAR improvements of 1.47%, 3.45%, and 14.32% at FARs of 10^{-3}, 10^{-4}, and 10^{-5}, respectively. Notably, equipping CosFace with our DDLN achieves highly competitive performance against the state-of-the-art GFace (e.g., 82.29% vs. 80.79% at FAR=10^{-5} at 89.22% vs. 89.57% at FAR=10^{-4}). This indicates that

Table 6: Performance (%) of various noise-robust models on MS-Celeb-1M [13], evaluated with ResNet-50 and ResNet-100 backbones (R1: Rank-1; 10^{-4}: TAR@FAR=10^{-4})

.

Method	LFW	CALFW	CPLFW	AgeDB	CFP	Avg	IJB-B		IJB-C	
							R1	10^{-4}	R1	10^{-4}
ArcFace [8]	99.77	95.92	90.93	97.35	96.91	96.18	92.44	89.55	93.56	92.25
CosFace [42]	99.75	**96.22**	91.97	97.35	97.31	96.52	91.81	88.52	92.98	91.54
NT [18]	-	-	-	-	-	-	-	91.57	-	93.65
NR [66]	-	-	-	-	-	-	-	91.58	-	93.60
Co-Mining [45]	-	93.28	85.70	95.80	93.32	-	-	91.80	-	93.82
Sub-Arc [7]	99.72	95.68	90.25	97.28	96.59	95.90	-	91.70	-	93.72
RVFace [44]	99.80	96.17	93.15	97.63	98.31	97.01	93.28	91.69	94.57	93.89
BoundaryFace [48]	99.72	95.59	91.39	96.96	97.41	96.21	-	84.80	-	88.13
RobustFace [53]	99.76	95.87	93.33	97.55	98.41	96.98	-	91.30	-	94.02
Arc+DDLN (Ours)	**99.82**	96.15	93.25	**98.43**	**98.73**	**97.28**	**95.29**	**94.98**	**96.63**	**96.44**
Cos+DDLN (Ours)	99.78	96.05	**93.38**	98.32	98.66	97.24	95.19	94.87	96.48	96.33
Sub-Arc (R100) [7]	99.80	96.13	92.68	98.12	98.40	97.03	95.26	93.43	96.34	95.25
DTDD (R100) [64]	**99.82**	**96.18**	92.90	-	-	-	-	-	95.39	95.47
Arc+DDLN (R100)	**99.82**	95.97	93.93	98.45	**99.03**	97.44	**95.57**	**95.68**	**96.98**	**96.96**
Cos+DDLN (R100)	**99.82**	96.03	**93.95**	**98.62**	98.94	**97.47**	95.44	95.47	96.83	96.87

even modest amounts of real-world noise severely bottleneck low-FAR performance if left unhandled.

WebFace2M-Noise Table 5 evaluates the highly corrupted WebFace2M-Noise dataset. As a 2-million-image subset sampled from the roughly 80%-noise WebFace260M [68], it presents a notoriously difficult denoising challenge due to its restricted data scale combined with an extreme noise ratio. Under this severe setting, DDLN delivers massive improvements across all three baseline architectures. Averaged across eight benchmarks, DDLN boosts CosFace, ArcFace, and SphereFace baselines by 7.41%, 7.68%, and 7.82% (reaching 84.43%, 84.70%, and 84.31%, respectively). On the challenging low-resolution TinyFace benchmark, our method outperforms RVFace by 1.77% in Rank-1 accuracy. When compared to the multi-stage SubRe baseline (which requires retraining on a cleaned subset), our *in-training* purification establishes a striking 5.10% lead. Finally, DDLN outperforms the recent RobustFace by a 7.15% margin on the 8-benchmark average, confirming its robust applicability under extreme real-world label corruption.

MS-Celeb-1M Table 6 reports results on MS-Celeb-1M [13], a widely-used benchmark for evaluating raw, uncurated web noise. We assess the 5-dataset average (Avg-5 on LFW, CALFW, CPLFW, AgeDB, CFP) alongside the surveillance-oriented IJB-B/C metrics (Rank-1 and TAR@FAR=10^{-4}). DDLN pushes the Avg-5 performance to near-saturation, achieving 97.28% with ResNet-50 and 97.47% with ResNet-100 (a 0.3% absolute gain over RobustFace). More importantly, on the demanding IJB-B and IJB-C benchmarks that closely mimic real-world open-set recognition, DDLN outperforms RobustFace [53] by significant margins of 3.68% and 2.55% at FAR=10^{-4}. These results strongly support

Table 7: MegaFace [21] evaluation (%) with MS-Celeb-1M [13] and ResNet50.

Metric	RV [44]	Boundary [48]	RobustFace [53]	**DDLN-Arc**	**DDLN-Cos**
Rank-1	96.33	96.67	97.75	98.59	**98.60**
TAR@FAR=10^{-6}	96.42	96.67	97.81	**98.81**	98.71

Table 8: Hyperparameter ablation results (%) of DDLN on simulated dataset #40–20–20 with ResNet50 [15] and CosFace [42]. Row 1 is a fixed-threshold baseline that directly uses $T'_{\max}(k = 10)$ as a constant filtering threshold (no progressive scheduling). $n_{\max}$ denotes the epoch when T_n reaches its maximum, set as $n_{\max} = e_{\text{decay}} - g$ to leave a gap of g epochs before the first learning-rate decay (e_{decay}=11); we ablate $g \in \{1, 2, 3\}$, yielding $n_{\max} \in \{10, 9, 8\}$.

max T_n	$n_{\max}$	α	λ	Precision	Recall	F1-score
$T'_{\max}(k = \mathbf{10})$	-	-	-	67.72	**99.86**	80.70
0.2				99.82	93.87	96.75
0.45				96.19	99.96	98.04
$T'_{\max}(k = \mathbf{1})$	9	0	0.01	97.97	99.91	98.93
$T'_{\max}(k = \mathbf{10})$				99.38	98.90	**99.14**
$T'_{\max}(k = \mathbf{50})$				99.65	97.50	98.56
	10			99.15	99.62	99.39
$T'_{\max}(k = 10)$	**9**	0.02	0.01	99.14	99.70	**99.42**
	8			99.05	99.72	99.38
			0.1	98.94	99.64	99.29
$T'_{\max}(k = 10)$	9	0.02	**0.05**	98.94	99.67	99.30
			0.005	99.11	99.54	99.32

Table 9: Stability of default hyperparameters. Results are reported as Avg (%) using the same evaluation metrics as the corresponding main experiments.

Setting	Default	k=1	k=50	λ=0.05	α=0
CASIA-WebFace	94.78	94.49	94.88	94.70	94.65
#80–40–40	78.11	71.97	70.68	76.04	75.26

DDLN's core motivation: it successfully discards harmful web noise while retaining the hard but informative samples crucial for strict face verification.

We further evaluate DDLN on MegaFace [21], a large-scale face-recognition benchmark. As shown in Table 7, DDLN achieves the best results among the compared robust face-recognition methods with ResNet-50. DDLN-Cos obtains 98.60% Rank-1 accuracy, and DDLN-Arc reaches 98.81% TAR@FAR=10^{-6}, exceeding RobustFace by 0.85% and 1.00%, respectively. This confirms the effectiveness of DDLN under large-scale identification and strict verification settings.

The Supplementary Material provides additional large-scale results on MS-Celeb-1M, using MobileFaceNet [5] as the backbone and comparing DDLN with representative noise-robust image classification methods [14, 33, 46, 50, 52, 55, 57].

5.3 Ablation Study

We analyze the main components that affect the sample-selection behavior of DDLN, including the threshold schedule, the rank statistic k, the slack term α, the schedule length $n_{\max}$, and the EMA factor λ. All ablations in Table 8 are conducted on the simulated #40–20–20 dataset with ResNet50 [15] and CosFace [42].

The fixed-threshold baseline obtains very high recall but much lower precision. This shows that a fixed upper threshold can remove most noisy samples, but it also incorrectly filters many hard clean samples. In contrast, the progressive threshold schedule provides a better precision-recall trade-off by gradually tightening the filtering boundary during early training.

The choice of k controls how the upper tail of the non-target similarity distribution is estimated. A small value, such as k=1, is more sensitive to rare outliers, while a large value, such as k=50, over-smooths the upper tail and lowers recall. The setting k=10 gives the best F1-score among the tested choices. Adding a small slack term (α=0.02) further improves the result by reducing unstable decisions near the boundary. The method is also not sensitive to the schedule length, with similar performance for $n_{\max} \in 8, 9, 10$. The EMA factor λ remains stable over a broad range, and all tested values produce close F1-scores.

We further examine whether the default hyperparameters generalize beyond the #40–20–20 setting. As shown in Table 9, the default setting k=10, α=0.02, λ=0.01 remains stable on both real-noise CASIA-WebFace and the more challenging #80–40–40 synthetic setting. Although a few variants obtain comparable performance on CASIA-WebFace, they degrade more clearly under heavy synthetic noise. These results support using the same default hyperparameters in all non-ablation experiments.

5.4 Extended Analysis in Supplementary Material

Due to space constraints, several detailed analyses are provided in the Supplementary Material. These include: (i) an *accumulated error analysis* showing that DDLN mitigates early confirmation bias without requiring an offline cleaning phase; (ii) a *dataset quality evaluation* comparing the automatically filtered subset with the CAST-cleaned subset [68], without multi-round retraining; (iii) a *complexity analysis* reporting the measured overhead of DDLN (about 0.2%, i.e., 1.07 ms per batch); (iv) *qualitative failure case visualizations* illustrating typical error modes under extreme long-tail variations and annotation noise; and (v) a comparison with representative noisy-label methods from image classification using MobileFaceNet [5] on MS-Celeb-1M [13].

6 Conclusion

We presented DDLN, a noise-robust face recognition framework based on progressive sample selection. Our key finding is that discarded noisy target similarities ($\cos\theta_{i,y_i}$) closely align with clean non-target similarities ($\cos\theta_{i,j}, j \neq y_i$) during training. Leveraging this distributional consistency, DDLN estimates noise boundaries directly from the observable clean non-target distribution and applies a progressively tightened threshold to filter unreliable pairs while preserving hard but clean samples in early training. DDLN is lightweight, requires no noise-rate prior or auxiliary networks, and can be integrated into standard margin-based face recognition pipelines with low measured overhead.

Potential Social Impact. We acknowledge the potential risks associated with the misuse of face recognition data. Experimental results on MS-Celeb-1M (withdrawn in 2019) and IJB datasets (withdrawn in 2023) are reported solely for reproducibility and fair comparison with prior work. For example, several recent face recognition studies also report results on these benchmarks, such as Extended$T$ (TPAMI 2023) [50], GFace (TIP 2025) [61], and RobustFace (TIP 2025) [53]. These datasets are no longer recommended for future research, and we do not advocate their continued use beyond controlled reproduction and comparison. To support more responsible evaluation, we additionally report results on WebFace2M, a recent subset derived from the publicly released WebFace260M dataset.

Acknowledgements

This work was supported in part by the Sichuan Provincial Science and Technology Program (Grant Nos.2024NSFTD0040, 2024ZDZX0046), and the Open Project Program of National Key Laboratory of Space-Born Intelligent Information Processing (No. YJ-03-25-04).

References

1. Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A closer look at memorization in deep networks. In: Proc. ICML. pp. 233–242 (2017)
2. Boutros, F., Damer, N., Kirchbuchner, F., Kuijper, A.: Elasticface: Elastic margin loss for deep face recognition. In: Proc. CVPR. pp. 1578–1587 (2022)
3. Boutros, F., Grebe, J.H., Kuijper, A., Damer, N.: Idiff-face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model. In: Proc. ICCV. pp. 19650–19661 (October 2023)
4. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: Proc. FG. pp. 67–74 (2018)
5. Chen, S., Liu, Y., Gao, X., Han, Z.: Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In: Chinese conference on biometric recognition. pp. 428–438. Springer (2018)

6. Cheng, Z., Zhu, X., Gong, S.: Low-resolution face recognition. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14. pp. 605–621. Springer (2019)
7. Deng, J., Guo, J., Liu, T., Gong, M., Zafeiriou, S.: Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In: Proc. ECCV. pp. 741–757 (2020)
8. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proc. CVPR. pp. 4690–4699 (2019)
9. Deng, J., Zhou, Y., Zafeiriou, S.: Marginal loss for deep face recognition. In: Proc. CVPRW. pp. 60–68 (2017)
10. Deng, W., Hu, J., Zhang, N., Chen, B., Guo, J.: Fine-grained face verification: Fglfw database, baselines, and human-dcmn partnership. Pattern Recognition **66**, 63–73 (2017)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Int. Conf. Learn. Represent. (2021)
12. Du, H., Shi, H., Zeng, D., Zhang, X.P., Mei, T.: The elements of end-to-end deep face recognition: A survey of recent advances. ACM Comput. Surv. **54**(10s), 1–42 (2022)
13. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: Proc. ECCV. pp. 87–102 (2016)
14. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I.W., Sugiyama, M.: Co-teaching: robust training of deep neural networks with extremely noisy labels. In: Proc. NeurIPS. pp. 8536–8546 (2018)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. CVPR. pp. 770–778 (2016)
16. Hou, X., Li, Y., Wang, S.: Disentangled representation for age-invariant face recognition: A mutual information minimization perspective. In: Proc. ICCV. pp. 3692–3701 (2021)
17. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
18. Hu, W., Huang, Y., Zhang, F., Li, R.: Noise-tolerant paradigm for training face recognition cnns. In: Proc. CVPR. pp. 11887–11896 (2019)
19. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition (2008)
20. Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Li, J., Huang, F.: Curricularface: adaptive curriculum learning loss for deep face recognition. In: Proc. CVPR. pp. 5901–5910 (2020)
21. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4873–4882 (2016)
22. Kim, M., Jain, A.K., Liu, X.: Adaface: Quality adaptive margin for face recognition. In: Proc. CVPR. pp. 18750–18759 (2022)
23. Kim, M., Liu, F., Jain, A., Liu, X.: Dcface: Synthetic face generation with dual condition diffusion model. In: Proc. CVPR. pp. 12715–12725 (June 2023)
24. Kong, C., Wang, S., Li, H., et al.: Digital and physical face attacks: Reviewing and one step further. APSIPA Transactions on Signal and Information Processing **12**(1) (2022)

25. Liu, B., Song, G., Zhang, M., You, H., Liu, Y.: Switchable k-class hyperplanes for noise-robust representation learning. In: Proc. ICCV. pp. 3019–3028 (October 2021)
26. Liu, W., Wen, Y., Raj, B., Singh, R., Weller, A.: Sphereface revived: Unifying hyperspherical face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **45**(2), 2458–2474 (2022)
27. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proc. CVPR. pp. 212–220 (2017)
28. Liu, W., Wen, Y., Yu, Z., Yang, M.: Large-margin softmax loss for convolutional neural networks. In: Proc. ICML. pp. 507–516 (2016)
29. Luan, X., Ding, Z., Liu, L., Li, W., Gao, X.: A symmetrical siamese network framework with contrastive learning for pose-robust face recognition. IEEE Trans. Image Process. **32**, 5652–5663 (2023)
30. Ma, B., Song, G., Liu, B., Liu, Y.: Rethinking robust representation learning under fine-grained noisy faces. In: ECCV. pp. 614–630 (2022)
31. Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., et al.: Iarpa janus benchmark-c: Face dataset and protocol. In: Proc. ICB. pp. 158–165 (2018)
32. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: the first manually collected, in-the-wild age database. In: Proc. CVPRW. pp. 51–59 (2017)
33. Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L.: Making deep neural networks robust to label noise: A loss correction approach. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1944–1952 (2017)
34. Pleiss, G., Zhang, T., Elenberg, E., Weinberger, K.Q.: Identifying mislabeled data using the area under the margin ranking. In: Proc. NeurIPS. pp. 17044–17056 (2020)
35. Ranjan, R., Castillo, C.D., Chellappa, R.: L2-constrained softmax loss for discriminative face verification. arXiv preprint arXiv:1703.09507 (2017)
36. Sengupta, S., Chen, J.C., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: Proc. WACV. pp. 1–9 (2016)
37. Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G.: Learning from noisy labels with deep neural networks: A survey. IEEE Trans. Neural Netw. Learn. Syst. **34**(11), 8135–8153 (2023)
38. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Proc. NeurIPS. pp. 1988–1996 (2014)
39. 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: Proc. ECCV. pp. 765–780 (2018)
40. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Process. Lett. **25**(7), 926–930 (2018)
41. Wang, F., Xiang, X., Cheng, J., Yuille, A.L.: Normface: L2 hypersphere embedding for face verification. In: Proc. ACM MM. pp. 1041–1049 (2017)
42. 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: Proc. CVPR. pp. 5265–5274 (2018)
43. Wang, M., Deng, W.: Deep face recognition: A survey. Neurocomputing **429**, 215–244 (2021)
44. Wang, X., Wang, S., Liang, Y., Gu, L., Lei, Z.: Rvface: Reliable vector guided softmax loss for face recognition. IEEE Trans. Image Process. **31**, 2337–2351 (2022)
45. Wang, X., Wang, S., Wang, J., Shi, H., Mei, T.: Co-mining: Deep face recognition with noisy labels. In: Proc. ICCV. pp. 9358–9367 (2019)

46. Wei, H., Feng, L., Chen, X., An, B.: Combating noisy labels by agreement: A joint training method with co-regularization. In: Proc. CVPR. pp. 13726–13735 (2020)
47. Whitelam, C., Taborsky, E., Blanton, A., Maze, B., Adams, J., Miller, T., Kalka, N., Jain, A.K., Duncan, J.A., Allen, K., et al.: Iarpa janus benchmark-b face dataset. In: Proc. CVPRW. pp. 90–98 (2017)
48. Wu, S., Gong, X.: Boundaryface: A mining framework with noise label self-correction for face recognition. In: European Conference on Computer Vision. pp. 91–106. Springer (2022)
49. Wu, X., He, R., Sun, Z., Tan, T.: A light cnn for deep face representation with noisy labels. IEEE Trans. Inf. Forensics Secur. **13**(11), 2884–2896 (2018)
50. Xia, X., Han, B., Wang, N., Deng, J., Li, J., Mao, Y., Liu, T.: Extended TT: Learning with mixed closed-set and open-set noisy labels. IEEE Trans. Pattern Anal. Mach. Intell. **45**(3), 3047–3058 (2023)
51. Xia, X., Han, B., Zhan, Y., Yu, J., Gong, M., Gong, C., Liu, T.: Combating noisy labels with sample selection by mining high-discrepancy examples. In: Int. Conf. Comput. Vis. pp. 1833–1843 (2023)
52. Xia, X., Liu, T., Wang, N., Han, B., Gong, C., Niu, G., Sugiyama, M.: Are anchor points really indispensable in label-noise learning? In: Adv. Neural Inform. Process. Syst. pp. 6835–6846 (2019)
53. Xin, Y., Zhong, X., Zhou, Y., Jiang, J.: Robust face recognition via adaptive mining and margining of noise and hard samples. IEEE Trans. Image Process. **34**, 8114–8129 (2025)
54. Yang, J., Wang, Z., Huang, B., Xiao, J., Liang, C., Han, Z., Zou, H.: Headpose-softmax: Head pose adaptive curriculum learning loss for deep face recognition. Pattern Recognition **140**, 109552 (2023)
55. Yao, Q., Yang, H., Han, B., Niu, G., Kwok, J.T.Y.: Searching to exploit memorization effect in learning with noisy labels. In: Int. Conf. Mach. Learn. pp. 10789–10798. PMLR (2020)
56. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
57. Yu, X., Han, B., Yao, J., Niu, G., Tsang, I., Sugiyama, M.: How does disagreement help generalization against label corruption? In: Int. Conf. Mach. Learn. pp. 7164–7173. PMLR (2019)
58. Zhang, M., Liu, R., Deguchi, D., Murase, H.: Texture-guided transfer learning for low-quality face recognition. IEEE Trans. Image Process. **33**, 95–107 (2024)
59. Zhang, Y., Deng, W., Wang, M., Hu, J., Li, X., Zhao, D., Wen, D.: Global-local gcn: Large-scale label noise cleansing for face recognition. In: Proc. CVPR. pp. 7731–7740 (2020)
60. Zhao, J., Yan, S., Feng, J.: Towards age-invariant face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **44**(1), 474–487 (2020)
61. Zhao, W., Zhu, X., Shi, H., Zhang, X., Zhao, G., Lei, Z.: Global cross-entropy loss for deep face recognition. IEEE Trans. Image Process. **34**, 1672–1685 (2025)
62. Zheng, T., Deng, W.: Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. Beijing University of Posts and Telecommunications, Tech. Rep **5**(7) (2018)
63. Zheng, T., Deng, W., Hu, J.: Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. arXiv preprint arXiv:1708.08197 (2017)
64. Zhong, Y., Deng, W., Fang, H., Hu, J., Zhao, D., Li, X., Wen, D.: Dynamic training data dropout for robust deep face recognition. IEEE Trans. Multimedia **24**, 1186–1197 (2021)

65. Zhong, Y., Deng, W., Hu, J., Zhao, D., Li, X., Wen, D.: Sface: Sigmoid-constrained hypersphere loss for robust face recognition. IEEE Trans. Image Process. **30**, 2587–2598 (2021)
66. 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: Proc. CVPR. pp. 7812–7821 (2019)
67. Zhou, J., Jia, X., Li, Q., Shen, L., Duan, J.: Uniface: Unified cross-entropy loss for deep face recognition. In: Proc. ICCV. pp. 20730–20739 (2023)
68. Zhu, Z., Huang, G., Deng, J., Ye, Y., Huang, J., Chen, X., Zhu, J., Yang, T., Lu, J., Du, D., et al.: Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In: Proc. CVPR. pp. 10492–10502 (2021)