Sparsity-Inducing Divergence Losses for Biometric Verification

Dimitrios Koutsianos[1,2], Ladislav Mošner[3,4], Yannis Panagakis[2,5], and Themos Stafylakis[1,2,4]

[1] Athens University of Economics and Business, Greece
[2] Archimedes/Athena Research Center, Greece
[3] Speech@FIT, Brno University of Technology, Czechia
[4] Omilia, Greece
[5] National and Kapodistrian University of Athens, Greece
{dkoutsianos, tstafylakis}@aueb.gr

Abstract. Performance in face and speaker verification is largely driven by margin-penalty softmax losses such as CosFace and ArcFace. Recently introduced α-divergence loss functions offer a compelling alternative, particularly due to their ability to induce sparse solutions (when $\alpha > 1$). However, standard geometric margins are designed for the softmax function and do not naturally extend to this generalized probabilistic framework. In this paper we propose Q-Margin, a novel α-divergence loss that introduces a principled probabilistic margin. Unlike conventional methods that apply geometric penalties to the logits (unnormalized log-likelihoods), Q-Margin encodes the margin penalty directly into the reference measure (prior probabilities). This formulation naturally encourages discriminative embeddings while preserving the beneficial sparsity properties of the α-divergence. We demonstrate that Q-Margin achieves competitive or superior performance on the challenging IJB-B and IJB-C face verification benchmarks and similarly strong results in speaker verification on VoxCeleb. Crucially, against ArcFace and CosFace baselines trained under an identical recipe, Q-Margin consistently improves at low False Acceptance Rates (FARs), a capability critical for practical high-security applications. Finally, the extreme sparsity of the Q-Margin posteriors enables exact and memory-efficient training, offering a scalable solution for datasets with millions of identities.

1 Introduction

Margin-penalty softmax losses such as CosFace [33] and ArcFace [10] have become the de facto standard in modern face and speaker recognition. They achieve state-of-the-art accuracy by incorporating margin penalties into the logit of the ground truth identity, which explicitly maximizes inter-class separation. Such methods rely on the softargmax operator, which transforms logits into probabilities evaluated by the Cross-Entropy loss or equivalently, the Kullback–Leibler (KL) divergence with a uniform prior [28]. To date, margin penalties have been applied exclusively through geometric modifications to logits (*e.g.*, $\cos(\theta + m)$

in ArcFace, $\cos(\theta) - m$ in CosFace). While effective, these geometric margins are heuristic extensions of softmax that remain decoupled from its probabilistic formulation.

Recent work by Roulet et al. [28] introduced a principled generalization of the standard Cross-Entropy loss through Fenchel–Young losses [3] derived from f-divergences [32]. This framework extends Cross-Entropy in two key ways: (1) by replacing the KL divergence with a broader family of f-divergences, each defining a unique convex loss and corresponding probability mapping (the softargmax$_f$), and (2) by allowing non-uniform reference measures $\mathbf{q}$ to encode arbitrary prior probabilities. While Roulet et al. used non-uniform $\mathbf{q}$ to represent class frequency imbalance, no prior work has exploited this flexibility to encode margin penalties probabilistically.

Among the f-divergence family, the α-divergence is particularly useful for biometric verification due to its tunable trade-off between smooth and sparse probability mappings. It defines a continuum of generalized entropy measures: recovering the Shannon negentropy (standard Cross-Entropy) as $\alpha \to 1$, the Gini negentropy (SparseMax [20]) at $\alpha = 2$, and exhibiting intermediate behaviors for values such as $\alpha = 1.5$. For $\alpha > 1$, the α-softargmax suppresses small probabilities, producing sparse posteriors that emphasize confident predictions. In practice, this encourages more discriminative embeddings and sharper decision boundaries in face and speaker recognition.

In this paper, we propose **Q-Margin**, a method that introduces a *probabilistic margin* by encoding the penalty directly in the reference measure $\mathbf{q}$. Instead of modifying the logits geometrically, we modify the prior for the ground truth class. By down-weighting the reference measure q_y of the target class, we create a stricter learning objective that naturally enforces a margin. This formulation preserves the sparsity benefits of α-divergence, allowing efficient computation of the normalizer of the posterior probabilities (which requires an iterative algorithm for $\alpha \neq 1$). In the limiting case $\alpha \to 1$, Q-Margin reduces to CosFace.

By reinterpreting margin learning through the lens of α-divergence, we unify geometric and probabilistic perspectives within a single framework that is theoretically grounded. This bridges a fundamental gap between information-theoretic loss design and the empirical success of margin-penalty softmax.

Our contributions are as follows:

- We propose α-divergence losses with margin penalties as an inherently-sparse alternative to standard logistic loss, for training highly discriminative identity embeddings.
- We derive Q-Margin, which introduces a principled probabilistic margin through non-uniform reference measures. We show that this formulation recovers the popular CosFace loss as $\alpha \to 1$, providing a theoretical link between probabilistic and geometric margins.
- We exploit the sparsity of the posteriors for $\alpha > 1$ and propose a simple modification of the iterative bisection algorithm (used to estimate the normalizer of posterior probabilities), which substantially accelerates training in large-scale datasets.

- We comprehensively evaluate on WebFace42M, IJB-B, IJB-C, and VoxCeleb, showing that Q-Margin is competitive with state-of-the-art methods including AdaFace [15] and PartialFC [2], and is particularly strong at the low FARs that matter for high-security verification.
- We provide a detailed empirical analysis of hyperparameter sensitivity, specifically investigating the interplay between α, scale (s) and margin (m) in different biometric modalities.

2 Related Work

2.1 Margin-Penalty Losses and Scalable Classifiers for Biometric Verification

Early deep face recognition models utilized the standard Softmax loss [5], but this approach was found to be insufficient for learning features with enough discriminative power for open-set recognition tasks. This limitation spurred the development of loss functions that explicitly engineer the feature space to increase inter-class variance while minimizing intra-class variance.

L-Softmax [19] was a pioneering effort that introduced an angular margin into the loss function, forcing decision boundaries to be more compact. SphereFace [18] further refined this by normalizing the final layer's weights and constraining features to a hypersphere with a multiplicative angular margin, though both methods proved challenging to train. The field subsequently shifted towards additive margins, which proved more stable and effective. CosFace [33] proposed adding a margin directly in cosine space ($\cos(\theta) - m$), while ArcFace [10] added an angular margin to the angle itself ($\cos(\theta + m)$). ArcFace in particular has become a widely used baseline in both face and speaker verification [38], with consistent adoption across VoxSRC challenges [14] and extensions such as large margin fine-tuning [31].

A common assumption in these fixed-margin methods is that all identities and samples present the same degree of learning difficulty. Recent adaptive margin strategies relax this assumption. AdaFace [15] dynamically adjusts the margin per sample using the feature norm as a quality proxy. KappaFace [25] assigns class-level margins via von Mises-Fisher concentration estimates. ElasticFace [4] replaces the fixed margin with a stochastic one drawn from a Gaussian distribution. While these methods improve flexibility, they remain coupled to the standard softmax Cross-Entropy and introduce additional components (quality estimators, distribution parameters) that are orthogonal to our contribution. Our approach instead generalizes the underlying divergence, and a probabilistic margin formulation like Q-Margin could in principle be combined with such adaptive schemes in future work.

On the scalability front, PartialFC [1, 2] addresses the computational bottleneck of the final fully-connected layer by sampling a subset of negative class centers each training step, reducing cost while maintaining accuracy. Our work offers a complementary perspective: the extreme posterior sparsity induced by α-divergences ($\alpha > 1$) naturally concentrates computation on a small active set

of classes, a similar efficiency gain that follows as a mathematical consequence of the loss rather than a sampling heuristic.

Despite their empirical success, all of the above methods apply margin penalties through *geometric* modifications to the logits before passing them to the standard Cross-Entropy loss. This leaves unexplored the question of whether margins can be introduced *within* a more general probabilistic loss framework, a gap our work aims to address.

2.2 Generalizations of the Cross-Entropy Loss

A separate line of research has focused on generalizing the Cross-Entropy loss from an information-theoretic standpoint. Roulet et al. [28] utilized Fenchel-Young losses [3] to derive a broad family of convex losses from f-divergences, providing a unified perspective that recovers, among others, the SparseMax transformation [20], a sparse alternative to softargmax capable of producing distributions with exact zeros. Chan and Kittler [6] combined SparseMax with ArcFace margins using MobileFaceNets [7], achieving competitive results. Peters et al. further generalized SparseMax with EntMax [26], creating a family of transformations with controllable sparsity parameterized by the α-divergence. Other information-theoretic generalizations include Rényi divergence objectives for variational inference [17, 27] and f-divergence-based training criteria [24], both demonstrating the flexibility of alternative divergences for learning.

Our work builds on the α-divergence loss from this framework, but addresses a different question. Prior uses of these generalized losses have either kept the reference measure uniform or used it to model class frequency imbalance [28]. No existing work has exploited the reference measure to encode *margin penalties*, the mechanism that has driven the empirical success of face and speaker verification systems. Q-Margin bridges these two research directions by introducing a principled probabilistic margin within the α-divergence framework, recovering CosFace as the special case $\alpha \to 1$.

3 Method

3.1 Notation

Let k be the number of identities in the training set. We denote $[k] = \{1, \dots, k\}$. We denote the probability simplex by $\Delta^{k-1} = \{\mathbf{p} \in \mathbb{R}^k_+ : \langle \mathbf{p}, \mathbf{1} \rangle = 1\}$. We also denote by $\boldsymbol{\theta} \in \mathbb{R}^k$ the output logits produced by a neural network, and $\mathbf{y} \in \{e_1, \dots, e_k\}$ denotes the one-hot identity labels.

Throughout this work, we adopted the convention proposed by Roulet et al. [28]:

$$\text{softmax}(\boldsymbol{\theta}) = \log \sum_{j'=1}^{k} \exp(\theta_{j'}) \quad (1) \qquad [\text{softargmax}(\boldsymbol{\theta})]_j = \frac{\exp(\theta_j)}{\sum_{j'=1}^{k} \exp(\theta_{j'})} \quad (2)$$

This convention distinguishes the log-partition function, $\text{softmax}(\boldsymbol{\theta})$, from its

gradient, the standard probability-generating softmax function, softargmax($\boldsymbol{\theta}$), as shown in Eq. (3). The softmax is the log-sum-exp operator and softargmax is the standard softmax function.

$$\nabla_{\boldsymbol{\theta}} \operatorname{softmax}(\boldsymbol{\theta}) = \operatorname{softargmax}(\boldsymbol{\theta}) \tag{3}$$

This equation also holds for the corresponding $\operatorname{softmax}_f$ and $\operatorname{softargmax}_f$ functions discussed below [28].

3.2 f- and α-divergence Loss Framework

Our methods build upon the generalized framework of Fenchel-Young losses [3], specifically using the f-divergence loss, l_f, as proposed by Roulet et al. [28].

This framework is derived from a fundamental optimization problem. The generalized operator, $\operatorname{softmax}_f(\boldsymbol{\theta})$ is defined as the solution to a regularized maximization problem:

$$\operatorname{softmax}_f(\boldsymbol{\theta}) = \max_{\mathbf{p} \in \Delta^{k-1}} \langle \mathbf{p}, \boldsymbol{\theta} \rangle - D_f(\mathbf{p} : \mathbf{q}) \tag{4}$$

Here, the f-divergence, $D_f(\mathbf{p} : \mathbf{q}) = \langle f(\mathbf{p}/\mathbf{q}), \mathbf{q} \rangle$, where $f(\cdot)$ is a convex function, acts as a regularizer. This generalizes the standard softmax (log-sum-exp) which is recovered when the KL divergence, $D_{KL}(\mathbf{p} : \mathbf{q})$, is used. The loss function is then defined as:

$$l_f(\boldsymbol{\theta}, \mathbf{y}; \mathbf{q}) = \operatorname{softmax}_f(\boldsymbol{\theta}) + D_f(\mathbf{y} : \mathbf{q}) - \langle \boldsymbol{\theta}, \mathbf{y} \rangle \tag{5}$$

where $\boldsymbol{\theta} \in \mathbb{R}^k$ are the logits, $\mathbf{y} \in \Delta^{k-1}$ is the ground-truth (in our case a one-hot vector with 1 at index y), $\mathbf{q} \in \mathbb{R}^k_+$ is the reference measure (class priors). Note that the term $D_f(\mathbf{y} : \mathbf{q})$ is constant with respect to the logits $\boldsymbol{\theta}$ and therefore does not influence the gradients.

The α-divergence is a subclass of f-divergence, generated by the convex function:

$$f(u) = \frac{(u^\alpha - 1) - \alpha(u - 1)}{\alpha(\alpha - 1)}, \quad \alpha \neq 1 \tag{6}$$

This framework generalizes the standard Cross-Entropy loss (recovered as $\alpha \to 1$ with $\mathbf{q} = \mathbf{1}$, yielding $f(u) = u \log u - (u - 1)$) and allows for controllable sparsity for $\alpha > 1$. We refer the reader to [28] for the full derivation.

3.3 Margin-Penalty Losses

The standard Cross-Entropy loss, while ensuring separability, proved insufficient to learn the highly discriminative features required for open-set identity recognition. Early approaches like L-Softmax [19] and SphereFace [18] introduced angular margins but faced training issues regarding both numerical stability and convergence. Subsequent research shifted towards additive margins, which proved to be more stable and effective in practice.

Two prominent examples dominated the field: CosFace [33] (additive cosine margin) and ArcFace [10] (additive angular margin). Both methods penalize the target logit to force the model to learn distinctly separated features, thereby increasing discriminative power. However, these successful approaches share a common mechanism: they directly modify the target logit geometrically before it is passed to the loss function. This geometric manipulation contrasts with Q-Margin, which achieves a similar discriminative effect probabilistically by modifying the reference measure $\mathbf{q}$ within the α-divergence framework.

3.4 Q-Margin: Modifying the Reference Measure

Building on the generalized α-divergence loss framework, we propose **Q-Margin**, which introduces the margin penalty through the reference measure $\mathbf{q}$. Instead of directly manipulating the scaled cosine similarities $\boldsymbol{\theta} = s \cdot \mathbf{c}$ (where $\mathbf{c} \in [-1, 1]^k$ is the vector of cosine similarities between a sample's embedding and the k class prototype vectors), we encode the margin penalty into $\mathbf{q}$. For a sample with ground-truth label y, the reference measure is defined as:

$$q_j = \exp\left(-s \cdot m\, \delta_{y,j}\right) \tag{7}$$

where s is the scaling factor, m is the margin hyperparameter and $\delta_{\cdot,\cdot}$ is the Kronecker delta function. By significantly down-weighting q_y, we create a stricter learning objective. Minimizing the α-divergence loss $l_f(\boldsymbol{\theta}, \mathbf{y}; \mathbf{q})$ compels the model to produce a substantially higher target logit θ_y relative to non-target logits to compensate, implicitly creating the desired decision margin through probabilistic means rather than direct geometric logit modification. The final loss is

$$L_{Q-M} = l_f(s \cdot \mathbf{c}, \mathbf{y}; \mathbf{q}). \tag{8}$$

It is easy to show that CosFace is a special case of Q-Margin as $\alpha \to 1$. The reference-measure margin in Eq. (7) is orthogonal to a geometric angular margin and the two can be combined, but in preliminary experiments on MS1MV3 and WebFace42M an added angular shift yielded no gain, so we retain the purely probabilistic formulation.

Gradients and Posterior Sparsity The gradients for the α-divergence loss, l_f, are key to understanding the model's behavior. The gradient w.r.t. the j-th logit θ_j (assuming target identity y) is $\frac{\partial l_f}{\partial \theta_j} = [\text{softargmax}_f(\boldsymbol{\theta})]_j - \mathbf{y}_j$. The j-th element of the α-softargmax, which we denote as the posterior probability p_j, is given by:

$$p_j = [\text{softargmax}_f(\boldsymbol{\theta})]_j = q_j[1 + (\alpha - 1)(\theta_j - \tau^*)]_+^{1/(\alpha-1)} \tag{9}$$

where τ^* is a scalar value, computed iteratively, that ensures the probabilities sum to one $\left(\sum_j p_j = 1\right)$. For $\alpha > 1$, the $[\cdot]_+ = \max(0, \cdot)$ operator is active, meaning that if a logit θ_j falls below a certain threshold relative to τ^*, the posterior probability p_j becomes exactly 0, resulting in sparse posterior distributions.

It is instructive to note the relationship between this formulation and the standard softargmax. As $\alpha \to 1$, the p_j term recovers the standard exponential function $\exp(x) = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n$. By setting $x = \theta_j - \tau^*$ and $n = 1/(\alpha - 1)$, we see that $\lim_{\alpha\to 1} p_j = \exp\left(\theta_j - \tau^* + \log q_j\right)$. This recovers the form of the standard softargmax probability, making the α-divergence loss a polynomial generalization of the standard Cross-Entropy loss, and the CosFace loss when $\mathbf{q}$ is as in Eq. (7).

Efficient Computation of α-softargmax The α-softargmax (Eq. (9)) is computed by first finding the scalar τ^*. This is the unique solution for τ in the root finding equation:

$$\sum_{j=1}^{k} q_j f'_*(\max\{\theta_j - \tau, f'(0)\}) = 1. \tag{10}$$

Here, f is the convex function generating the α-divergence (Eq. (6)), $f'(u) = \frac{u^{\alpha-1}-1}{\alpha-1}$, $\alpha \neq 1$, the derivative of f, $f^*(v) = \frac{1}{\alpha}([1+(\alpha-1)v]_+^{\frac{\alpha}{\alpha-1}} - 1)$, $\alpha \in (1, +\infty)$ is the convex conjugate of f, while $f'_* = (f^*)'$ is the derivative of the convex conjugate. The components q_j are from the reference measure and θ_j are the input logits.

Following Roulet et al. [28], this value τ^* is found efficiently with the iterative bisection algorithm. The search is bounded within the range $[\tau_{\min}, \tau_{\max}]$ where for $t \in \operatorname{argmax}_{j\in[k]} \theta_j$ it is $\tau_{\min} = \theta_t - f'(1/q_t)$ and $\tau_{\max} = \theta_t - f'\left(1/\sum_j q_j\right)$.

A key advantage of the α-divergence with $\alpha > 1$ is the extreme sparsity of the resulting posterior distribution. In our experiments, the vast majority of class probabilities are assigned *exactly* zero mass (see Section 4.4 for detailed statistics). We leverage this property to significantly reduce computational cost. Instead of computing the root-finding step over all k identities (which can be in the millions for datasets like WebFace42M), we sort the logits (in half-precision) to retrieve the top-5% of logits and compute τ^* solely on this active set. Since the remaining 95% of logits would fall below the sparsity threshold regardless, this optimized procedure is exact, not approximate (Proposition 1), while substantially reducing the computational overhead of the iterative bisection.

Proposition 1 (Exactness of top-K truncation). *Fix $\alpha > 1$ and let τ^* solve Eq. (10) over all k classes, with active support $S = \{j : p_j > 0\} = \{j : \theta_j > \tau^* - \frac{1}{\alpha-1}\}$. If $S \subseteq \mathcal{T}_K$, where $\mathcal{T}_K$ indexes the K largest logits, then solving Eq. (10) over $\mathcal{T}_K$ alone yields the same τ^* and the same posterior $\mathbf{p}$ as the full computation.*

Proof sketch. From Eq. (9), $p_j > 0 \iff \theta_j > \tau^* - \frac{1}{\alpha-1}$, so every class outside S contributes $p_j = 0$ and drops out of the sum in Eq. (10). If $S \subseteq \mathcal{T}_K$, the discarded logits are all inactive, hence restricting the root-finding to $\mathcal{T}_K$ leaves τ^* and $\mathbf{p}$ unchanged. □

We fix $K = 5\%$ throughout (not tuned per run). This is deliberately conservative: the active support never exceeded ∼0.4% in any of our experiments, so

the wide margin guarantees that $S \subseteq T_K$ holds comfortably rather than relying on a tight threshold. As an additional safeguard we verify $S \subseteq T_K$ at runtime by checking that the smallest retained logit is itself inactive. The bound was never violated in any of our experiments.

4 Face Recognition Experiments

4.1 Training and Testing Data

Our models were trained on WebFace42M [42], a cleaned subset of WebFace260M containing 42M images of 2M identities. Following standard face-recognition protocols, all images were aligned and resized to 112×112 pixels. During training, several standard verification benchmarks served as development sets: LFW [12], CFP-FP [30], AgeDB-30 [23], CALFW [40] and CPLFW [39], for which we report True Acceptance Rate (TAR) at a False Acceptance Rate (FAR) of 10^{-3}.

The final evaluation was performed on the more challenging IJB-B [36] and IJB-C [21] benchmarks. These datasets are designed to test recognition performance in unconstrained scenarios. On these two test sets, we report TAR at FAR levels of 10^{-4} and 10^{-5} as our primary performance metrics, which represent standard operating points for evaluating these benchmarks.

Due to the significant computational cost of training on WebFace42M, most configurations were trained a single time; we additionally repeated our reference configuration ($\alpha = 1.25$, $s = 32$, $m = 0.2$) across three random seeds (Table 3) to verify stability. We further mitigate the single-run limitation by reporting a broad hyperparameter sweep in Table 2 and by validating trends across two distinct biometric domains.

4.2 Implementation Details

The experimental framework was built upon the InsightFace toolkit[1], specifically its ArcFace implementation, with minimal modifications to integrate the custom α-divergence loss. Experiments were conducted on two equivalent setups: AWS (8×NVIDIA A100 40GB GPUs) and LUMI (4×AMD MI250X GPUs, each exposing two GCDs for 8 logical devices).

We evaluated a standard backbone architecture from the ResNet family [11], the ResNet-100 (R-100) backbone. All models were trained for 20 epochs using Stochastic Gradient Descent (SGD) as an optimizer. We employed a step-wise learning rate decay schedule, starting at 0.1 for the first 7 epochs, decreasing to 0.01 for the next 6 epochs, and concluding at 0.001 for the final 7 epochs. A momentum of 0.9 and a weight decay of 5×10^{-4} were used. The batch size was set to 128 per device (GCD), resulting in an effective total batch size of 1024.

The baseline losses (ArcFace, CosFace) use their canonical published settings for each modality. For Q-Margin we grid-searched α, s, and m; the grids differ by modality because the optimal scale is coupled to α (discussed below), so face

[1] `https://github.com/deepinsight/insightface` (accessed June 1, 2025)

configurations explore larger s than speaker ones. We report every configuration evaluated (Table 2 for face; Table 5 for speaker), perform no held-out model selection, and evaluate the final-epoch checkpoint once on IJB-B/C. Baselines received comparable search effort over their standard (s, m) ranges.

4.3 Results

We present a comprehensive evaluation of Q-Margin against the industry-standard margin-penalty softmax losses ArcFace [10] and CosFace [33], alongside recent state-of-the-art methods including AdaFace [15], PartialFC [1], UniFace [41], and TopoFR [9].

As shown in Table 1, Q-Margin is the strongest method in the controlled (re-implemented) block, which is the only setting where loss-function differences are isolated from training data and backbone. Our best configuration ($\alpha = 1.25$, $s = 35$, $m = 0.2$) reaches 93.67% on IJB-B and 96.35% on IJB-C at FAR=10^{-5}, improving over the ArcFace and CosFace baselines trained identically.

Against the reported numbers, Q-Margin is competitive rather than uniformly ahead. It essentially ties PartialFC on IJB-C (97.78% vs. 97.82% at 10^{-4}; 96.35% vs. 96.46% at 10^{-5}), while methods such as UniFace and TopoFR report higher IJB-C numbers using larger backbones (R-200) and additional architectural components. These gains are therefore not attributable to the loss alone, whereas Q-Margin obtains its results purely through a change in the objective, with no auxiliary quality network, negative sampling, or deeper backbone.

Notably, Q-Margin's near-parity with PartialFC is achieved while activating only $\sim$0.4% of classes per example, suggesting the sparsity mechanism may play a role analogous to hard-negative mining.

The Detection Error Tradeoff (DET) curves in Figure 2a provide a visual corroboration of these results. Our best Q-Margin configuration (blue line) is shown to consistently outperform the ArcFace (pink) and CosFace (red) baselines.

Beyond quantitative metrics, Figure 1 provides qualitative evidence of Q-Margin's improved robustness. The examples show our method succeeding where ArcFace fails, correctly verifying genuine pairs despite large intra-class variations (*e.g.* extreme pose, aging) and correctly rejecting impostor pairs. This suggests Q-Margin learns a more robust and discriminative embedding space.

Hyperparameter Sensitivity Standard margin-penalty softmax losses, like ArcFace, typically employ a large margin penalty ($m = 0.5$) combined with a high scaling factor ($s = 64$). Our experiments indicate that this configuration is suboptimal for Q-Margin. As shown in Table 2, fixing $\alpha = 1.25$ and $s = 32$ while increasing m from 0.2 to 0.5 results in a performance drop on both IJB-B and IJB-C. This suggests that the α-divergence loss formulation is sensitive to aggressive probabilistic-margin penalties. We find that a moderate margin ($m = 0.2$) yields the most robust results. To check that these gains reflect a stronger objective rather than seed variance, we repeated the $\alpha = 1.25$, $s = 32$, $m = 0.2$ configuration across three seeds (Table 3); the maximum standard de-

Table 1: Comparison with state-of-the-art methods on IJB-B and IJB-C. TAR (%) at FAR=10^{-4} and 10^{-5}. *Reported* numbers are taken from the original papers and differ in training corpus and/or backbone, so they serve as literature context rather than controlled comparisons. *Re-implemented* methods share an identical recipe (ResNet-100, WebFace42M), isolating loss-function differences; best in bold. EntMax and SparseMax also use an ArcFace-style angular margin. †Backbone depth not specified in the original paper.

				IJB-B		IJB-C	
Method	Train Data	Backbone	Venue	10^{-4}	10^{-5}	10^{-4}	10^{-5}
Reported							
AdaFace [15]	WF4M	R-100	CVPR22	96.03	-	97.39	-
PartialFC (0.3) [1]	WF42M	R-100	CVPR22	96.47	-	97.82	96.46
CAFace + AdaFace [16]	WF4M	R-100	NeurIPS22	95.53	92.29	97.30	95.96
F^2C [34]	WF42M	R-100	CVPR22	-	-	97.31	-
QAFace [29]	MS1MV2	ResNet†	WACV23	95.67	-	97.20	-
CurricularFace [13]	MS1MV2	R-100	CVPR20	94.80	-	96.10	-
MagFace [22]	MS1MV2	R-100	CVPR21	94.51	90.36	95.97	94.08
ElasticFace [4]	MS1MV2	R-100	CVPR22	95.43	-	96.65	-
UniFace [41]	WF42M	R-200	ICCV23	-	-	97.91	96.68
TopoFR [9]	WF42M	R-200	NeurIPS24	-	-	98.01	97.10
Re-Implemented							
ArcFace (s=64, m=0.5)	WF42M	R-100	-	96.09	92.26	97.54	95.70
CosFace (s=64, m=0.5)	WF42M	R-100	-	96.05	93.15	97.49	95.82
EntMax (α=1.25, s=64, m=0.5)	WF42M	R-100	-	96.13	93.06	97.55	96.11
SparseMax (α=2, s=1.9, m=0.2)	WF42M	R-100	-	95.50	91.84	97.12	95.28
Q-Margin (α=1.25, s=35, m=0.2) (**Ours**)	WF42M	R-100	-	**96.44**	**93.67**	**97.78**	**96.35**

viation across IJB-B/C metrics is 0.07 and the ordering over the ArcFace and CosFace baselines holds in every seed.

Furthermore, we observe a distinct coupling between the divergence parameter α and the scaling factor s. Fixing $m = 0.2$, the optimal scale depends strongly on α: for $\alpha = 1.25$ performance improves steadily as s increases, whereas for $\alpha = 1.5$ it stagnates or degrades at larger scales (Tab. 2). This coupling admits a simple mechanistic explanation: increasing α makes the softargmax$_f$ mapping peakier at a fixed scale, and raising s has the same sharpening effect, so the two trade off along a single axis and the optimal scale *decreases* as α grows rather than varying arbitrarily. Accordingly, α and s cannot be tuned independently, as increasing both at once destabilizes the optimization. Therefore we adopt $\alpha = 1.25$, $s = 35$, $m = 0.2$ as the effective operating point on WebFace42M, achieving the best results while maintaining training stability.

As a practical guideline we recommend starting with $\alpha = 1.25$ and a moderate margin ($m = 0.2$), then tuning s upward, since performance is more sensitive to margin than scale across almost all configurations we evaluated. Higher α (e.g. $\alpha = 1.5$) gave diminishing returns at this scale but proved beneficial in the smaller speaker verification setting (Section 5).

4.4 Computational Efficiency

A potential concern with the α-divergence framework is the computational overhead of the α-softargmax. Unlike the standard Softmax used in ArcFace, which

Table 2: Results for the WebFace42M training dataset. For all configurations we report the single run TAR.

	Parameters			Benchmarks					IJB-B		IJB-C	
Loss	α	s	m	LFW	CFP-FP	AgeDB-30	CALFW	CPLFW	10^{-4}	10^{-5}	10^{-4}	10^{-5}
ArcFace	-	32	0.5	**99.88**	99.36	98.35	96.15	**94.90**	95.75	91.69	97.42	95.36
ArcFace	-	64	0.5	99.85	99.36	98.43	96.18	94.73	96.09	92.26	97.54	95.70
CosFace	-	32	0.5	99.87	99.24	98.22	96.03	94.77	95.81	92.50	97.41	95.52
CosFace	-	64	0.5	99.82	99.20	98.28	96.08	94.52	96.05	93.15	97.49	95.82
Q-Margin	1.25	10	0.2	99.87	**99.39**	98.42	**96.25**	94.87	96.08	92.95	97.62	95.91
Q-Margin	1.25	10	0.5	99.87	99.36	98.33	96.15	94.80	96.03	91.79	97.55	95.72
Q-Margin	1.25	20	0.2	99.85	99.34	98.45	96.22	**94.90**	96.21	92.95	97.58	95.91
Q-Margin	1.25	20	0.5	99.83	99.36	98.30	96.17	94.80	96.02	92.45	97.57	95.92
Q-Margin	1.25	32	0.2	99.87	99.36	**98.57**	96.15	94.83	96.30	93.38	97.76	**96.35**
Q-Margin	1.25	32	0.5	99.85	99.34	98.33	96.13	94.68	96.05	92.67	97.51	96.00
Q-Margin	1.25	35	0.2	99.83	**99.39**	**98.57**	96.08	94.78	**96.44**	**93.67**	**97.78**	**96.35**
Q-Margin	1.5	10	0.2	99.85	99.33	98.28	96.22	94.65	95.86	92.26	97.48	95.64
Q-Margin	1.5	10	0.5	99.87	99.29	98.15	96.08	94.42	95.56	91.65	97.07	95.04
Q-Margin	1.5	20	0.2	99.83	99.30	98.00	96.15	94.42	95.71	91.85	97.21	95.59
Q-Margin	1.5	20	0.5	99.83	99.29	98.27	96.17	94.58	95.86	92.94	97.38	95.92
Q-Margin	1.5	32	0.2	99.85	99.29	98.30	96.10	94.67	96.04	92.15	97.49	95.93
Q-Margin	1.5	32	0.5	99.85	99.37	98.35	96.12	94.70	95.99	92.27	97.43	95.76

benefits from highly optimized closed-form GPU implementations, our formulation requires an iterative bisection algorithm to find the root τ^*. On extreme-scale datasets like WebFace42M, which contains over two million distinct identity classes, evaluating this root-finding algorithm across all logits simultaneously can introduce a noticeable computational bottleneck.

To quantify this, we benchmarked the training throughput (samples per second) on the WebFace42M dataset using the ResNet-100 backbone. As shown in Table 4, the naive implementation of Q-Margin evaluated over all 2M+ identities (Q-M 100%) yields a throughput of 1466.45 samples/s, representing a roughly 27% slowdown compared to the ArcFace baseline (2005.13 samples/s).

However, we mitigate this overhead by explicitly leveraging the extreme sparsity of the α-divergence posterior ($\alpha > 1$). In practice, across all training examples on WebFace42M, the highest number of non-zero elements in $\mathbf{p}$ was 7,976 out of over 2M classes ($\sim 0.4\%$), with an average of only 1,020 per example. We can therefore perform a partial sort to isolate the top-$K\%$ of logits and compute

Table 3: Multi-seed results for Q-Margin ($\alpha = 1.25$, $s = 32$, $m = 0.2$). Mean $\pm$ standard deviation over three runs with distinct random seeds. The maximum standard deviation across IJB-B/C metrics is 0.07, confirming the reported gains are stable across seeds.

Benchmarks					IJB-B		IJB-C	
LFW	CFP-FP	AgeDB-30	CALFW	CPLFW	10^{-4}	10^{-5}	10^{-4}	10^{-5}
99.83±0.03	99.35±0.02	98.51±0.06	96.13±0.02	94.87±0.04	96.38±0.07	93.43±0.05	97.77±0.01	96.34±0.02

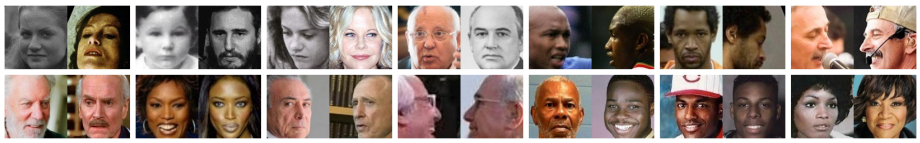


Fig. 1: Challenging pairs from AgeDB-30, CALFW, and CPLFW where our proposed method (Q-Margin, α=1.25, s=32, m=0.2) succeeds, while the ArcFace baseline fails. The top row shows genuine pairs correctly accepted by our model and falsely rejected by the baseline. The bottom row shows impostor pairs correctly rejected by our model and falsely accepted by the baseline.

τ^* solely on this active set. Since this maximum is an order of magnitude below $K = 5\%$, the active support is contained in the top-K set, so the truncation is exact by Proposition 1 rather than an approximation.

As demonstrated in Table 4, restricting the bisection to the top 5% of logits (Q-Margin 5%) yields 1900.26 samples/s, a ∼5% overhead versus ArcFace. We retain this $K = 5\%$ even though the observed active set never exceeds ∼0.4%: the generous margin guarantees $S \subseteq T_K$ without relying on a tight bound that a rare example might exceed. Pushing to the top 1% (Q-Margin 1%) recovers throughput almost entirely (1978.07 samples/s, ∼1.3% difference), so even this conservative choice scales to datasets with millions of identities.

Table 4: Average throughput on WebFace42M (> 2M identities) with ResNet-100. Q-Margin (1%) means that we sort and keep only the top 1% of the logits.

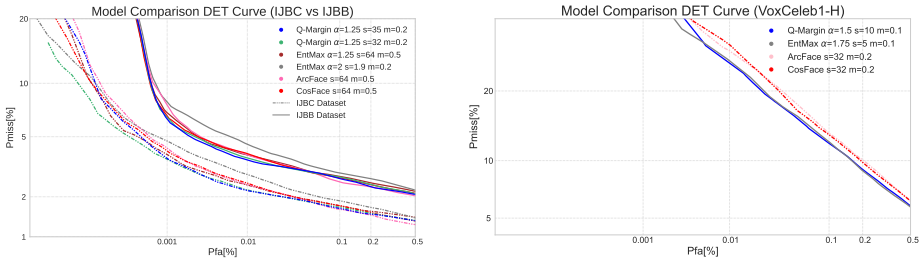
ArcFace	Q-Margin (1%)	Q-Margin (5%)	Q-Margin (100%)
2005.13	1978.07	1900.26	1466.45

5 Speaker Verification Experiments

Beyond face verification, margin-penalty softmax losses are equally prevalent in speaker verification [38], where speaker embeddings are increasingly used alongside visual features in audio-visual tasks such as speaker diarization [37]. We evaluate Q-Margin in this domain to test whether the benefits of probabilistic margins transfer across biometric modalities, particularly in a regime with far fewer training identities (5,994 vs. 2M).

5.1 Training and Testing Data

For speaker verification, we adopted the standard protocol established by the VoxCeleb benchmarks. The training corpus is the development set of VoxCeleb2 [8], comprising about 1M utterances from 5,994 speakers.



(a) DET Curves for face recognition models. **(b)** DET Curves for speaker recognition models.

Fig. 2: False acceptance vs missed detection probabilities for various losses on IJB-B and IJB-C (left) and VoxCeleb1-H (right), focused on low FARs.

For evaluation, we adopted the challenging trial list termed VoxCeleb1-H (hard). It contains 550,894 trials featuring 1,190 speakers. The difficulty of the evaluation set arises from the constraint that trials consist of utterances from speakers of the same gender and nationality. While other trial lists (VoxCeleb1-O and VoxCeleb1-E) exist, we focus on this *hard* set to maintain alignment with the challenging scenarios evaluated in the face recognition experiments. While experiments on the other trial lists yielded similar conclusions, we omit those results for brevity. Our analysis focuses on TAR(%) at FAR of 10^{-3} and 10^{-4}.

5.2 Implementation Details

Our implementation is based on the WeSpeaker toolkit[2] [35]. Training was conducted on 4 NVIDIA RTX A4000 (16GB) GPUs. A ResNet-34 (R-34) backbone was employed as the feature extraction network. The model was trained for 150 epochs using SGD with a momentum of 0.9 and a weight decay of 10^{-4}. The learning rate followed a two-phase schedule: a linear warm-up during the first six epochs, after which it decreased exponentially from 0.1 to $5 \cdot 10^{-5}$ over the remaining training period. Following common practice in speaker verification, the margin parameter m was annealed, starting from 0.0 and increasing exponentially between epochs 20 and 40 to its target value; this value was subsequently held constant. A batch size of 128 per GPU was used. To improve robustness, input data was augmented with additive noise and reverberation during training.

5.3 Results

Consistent with the face recognition evaluation, we compare our methods with the CosFace and ArcFace baselines. Table 5 reports TAR at different FAR values on VoxCeleb1-H, demonstrating the strong performance of our methods and their sensitivity to hyperparameters (α, s, m). For a given α, optimal s and m were found by empirical search over $s \in \{5, 10, 20, 32\}$ and $m \in \{0.1, 0.2, 0.3\}$.

[2] `https://github.com/wenet-e2e/wespeaker`, (accessed September 30, 2025)

Table 5: Results with a ResNet-34 (R-34) model trained on VoxCeleb2 dev. We report TAR (%) at FAR of 10^{-3} and 10^{-4}. The experiments are performed using VoxCeleb1-H as test set.

	Parameters			TAR@FAR	
Loss	α	s	m	10^{-3}	10^{-4}
ArcFace	-	32	0.2	86.62	72.08
CosFace	-	32	0.2	86.68	70.61
EntMax	2.0	5	0.1	84.71	74.37
EntMax	1.75	5	0.1	87.69	74.08
Q-Margin	1.75	5	0.1	87.02	**74.60**
Q-Margin	1.5	10	0.1	**87.91**	72.96

In contrast to the face recognition results (which favored $\alpha = 1.25$), we observed that higher values of α (1.5 and 1.75) lead to better-performing speaker verification systems. We hypothesize that this is related in part to the significantly smaller number of training identities. However, class count is not the whole story: although the speaker corpus has far fewer identities, modality-specific factors such as embedding-space geometry and intra-class variability also shape the ideal sparsity level. Such modality-dependent tuning is not unique to Q-Margin since standard margin-penalty losses exhibit the same gap, with ArcFace's optimum at $s = 64, m = 0.5$ for face versus $s = 32, m = 0.2$ for speaker. Conversely, these results reinforce observations made in Table 2, i.e., that the most consistent improvements are achieved with moderate values of scale ($s \in \{5, 10\}$) and margin ($m = 0.1$). Specifically, the same α–s interdependence identified for face holds here: higher α requires smaller s, which is consistent with the peakiness mechanism of Section 4.3 (Eq. (9)).

Both Q-Margin and EntMax yield strong results at the two operating points of interest. We observe a significant improvement from Q-Margin ($\alpha = 1.5, s = 10, m = 0.1$) over both baselines, increasing the TAR to 87.91% at 10^{-3} FAR. At the stricter 10^{-4} FAR, the Q-Margin ($\alpha = 1.75$, $s = 5$, $m = 0.1$) formulation achieves the highest overall score (74.60%).

To visually illustrate these outcomes, we present DET curves in the low-FAR region for selected systems in Figure 2b. While the curves for CosFace and ArcFace nearly overlap, the curves for models using our proposed losses are visibly separated, indicating superior performance and robustness.

6 Conclusions

Margin-penalty softmax loss functions have driven the success of embedding-based networks in face and speaker verification, while recent advances in generalized Cross-Entropy (α-divergence losses) have enabled principled control over posterior sparsity. However, combining these two paradigms is not straightfor-

ward: standard geometric margins, designed for the softmax function, do not naturally extend to sparsity-inducing α-divergences.

To address this, we propose Q-Margin, a formulation that introduces a probabilistic margin by encoding penalties directly into the non-uniform reference measure $\mathbf{q}$. This approach bridges information-theoretic loss design with margin-based learning, providing a principled alternative to geometric logit manipulation that recovers CosFace as the special case $\alpha \to 1$.

Extensive evaluations on large-scale benchmarks (IJB-B/C, VoxCeleb1-H) demonstrate highly competitive performance, particularly at strict low-FAR operating points critical for high-security applications. Furthermore, we showed that the inherent sparsity of Q-Margin can be leveraged to optimize the iterative bisection algorithm, yielding exact and highly efficient training throughput even on datasets with millions of identities. Our analysis also highlighted the interplay between α, scale (s) and margin (m), offering practical guidance for tuning these generalized loss functions. We note that most of our WebFace42M experiments were run a single time due to computational cost, so the reported gains, while consistent across benchmarks and modalities, lack confidence intervals. Additionally, Q-Margin introduces one extra hyperparameter (α) compared to standard margin-penalty softmax losses, and we observed that α and s interact in ways that require joint tuning.

Future work. Natural extensions include dynamically updating the reference measure based on sample difficulty, connecting Q-Margin with adaptive strategies such as AdaFace, and evaluating on additional corpora for fully controlled comparisons.

Ethical considerations. Face and speaker verification rely on biometric data that is personal, hard to revoke once compromised, and prone to unequal error rates across demographic groups, and such models can be repurposed for surveillance without consent. Q-Margin's low-FAR, high-security focus can sharpen these concerns, so deployments should use consensually collected data, audit demographic fairness at the relevant operating points, and follow applicable privacy regulation.

Acknowledgements

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.
AWS resources were provided by the National Infrastructures for Research and Technology GRNET and funded by the EU Recovery and Resiliency Facility.
The work was supported by Ministry of Education, Youth and Sports of the Czech Republic (MoE) through the OP JAK project "Linguistics, Artificial Intelligence and Language and Speech Technologies: from Research to Applications" (ID:CZ.02.01.01/00/23_020/0008518).

References

1. An, X., Deng, J., Guo, J., Feng, Z., Zhu, X., Yang, J., Liu, T.: Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4032–4041 (2022). https://doi.org/10.1109/CVPR52688.2022.00401
2. An, X., Zhu, X., Gao, Y., Xiao, Y., Zhao, Y., Feng, Z., Wu, L., Qin, B., Zhang, M., Zhang, D., Fu, Y.: Partial fc: Training 10 million identities on a single machine. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 1445–1449 (2021). https://doi.org/10.1109/ICCVW54120.2021.00166
3. Blondel, M., Martins, A., Niculae, V.: Learning classifiers with fenchel-young losses: Generalized entropies, margins, and algorithms. In: Chaudhuri, K., Sugiyama, M. (eds.) Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 89, pp. 606–615. PMLR (16–18 Apr 2019), https://proceedings.mlr.press/v89/blondel19a.html
4. Boutros, F., Damer, N., Kirchbuchner, F., Kuijper, A.: Elasticface: Elastic margin loss for deep face recognition. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 1577–1586 (2021). https://doi.org/10.1109/CVPRW56347.2022.00164
5. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: VGGFace2: A Dataset for Recognising Faces across Pose and Age . In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 67–74. IEEE Computer Society, Los Alamitos, CA, USA (May 2018). https://doi.org/10.1109/FG.2018.00020, https://doi.ieeecomputersociety.org/10.1109/FG.2018.00020
6. Chan, C.H., Kittler, J.: Angular sparsemax for face recognition. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 10473–10479. IEEE (2021)
7. 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)
8. Chung, J.S., Nagrani, A., Zisserman, A.: VoxCeleb2: Deep speaker recognition. In: Proc. Interspeech. pp. 1086–1090 (2018)
9. Dan, J., Liu, Y., Deng, J., Xie, H., Li, S., Sun, B., Luo, S.: Topofr: A closer look at topology alignment on face recognition. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 37213–37240. Curran Associates, Inc. (2024). https://doi.org/10.52202/079017-1174, https://proceedings.neurips.cc/paper_files/paper/2024/file/419b6c974712adb884bfbbeea8e94d1b-Paper-Conference.pdf
10. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4685–4694 (2019). https://doi.org/10.1109/CVPR.2019.00482
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

12. 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)
13. 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: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5901–5910 (2020)
14. Huh, J., Chung, J.S., Nagrani, A., Brown, A., Jung, J.w., Garcia-Romero, D., Zisserman, A.: The VoxCeleb speaker recognition challenge: A retrospective. IEEE/ACM Transactions on Audio, Speech, and Language Processing **32**, 3850–3866 (2024). https://doi.org/10.1109/TASLP.2024.3444456
15. Kim, M., Jain, A.K., Liu, X.: Adaface: Quality adaptive margin for face recognition. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18729–18738 (2022). https://doi.org/10.1109/CVPR52688.2022.01819
16. Kim, M., Liu, F., Jain, A.K., Liu, X.: Cluster and aggregate: Face recognition with large probe set. vol. 35, pp. 36054–36066 (2022)
17. Li, Y., Turner, R.E.: Rényi divergence variational inference. Advances in neural information processing systems **29** (2016)
18. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6738–6746 (2017). https://doi.org/10.1109/CVPR.2017.713
19. Liu, W., Wen, Y., Yu, Z., Yang, M.: Large-margin softmax loss for convolutional neural networks. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48. p. 507–516. ICML'16, JMLR.org (2016)
20. Martins, A., Astudillo, R.: From softmax to sparsemax: A sparse model of attention and multi-label classification. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1614–1623. PMLR, New York, New York, USA (20–22 Jun 2016), https://proceedings.mlr.press/v48/martins16.html
21. 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: 2018 international conference on biometrics (ICB). pp. 158–165. IEEE (2018)
22. Meng, Q., Zhao, S., Huang, Z., Zhou, F.: Magface: A universal representation for face recognition and quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14225–14234 (June 2021)
23. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: the first manually collected, in-the-wild age database. In: proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 51–59 (2017)
24. Novello, N., Tonello, A.M.: f-divergence based classification: beyond the use of cross-entropy. In: Proceedings of the 41st International Conference on Machine Learning. ICML'24, JMLR.org (2024)
25. Oinar, C., M. Le, B., Woo, S.S.: Kappaface: Adaptive additive angular margin loss for deep face recognition. IEEE Access **11**, 137138–137150 (2023). https://doi.org/10.1109/ACCESS.2023.3338648

26. Peters, B., Niculae, V., Martins, A.F.T.: Sparse sequence-to-sequence models. In: Korhonen, A., Traum, D., Màrquez, L. (eds.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 1504–1519. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1146, https://aclanthology.org/P19-1146/
27. Rényi, A.: On measures of entropy and information. In: Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics. vol. 4, pp. 547–562. University of California Press (1961)
28. Roulet, V., Liu, T., Vieillard, N., Sander, M.E., Blondel, M.: Loss functions and operators generated by f-divergences. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=V1YfPJDliw
29. Saadabadi, M.S.E., Malakshan, S.R., Zafari, A., Mostofa, M., Nasrabadi, N.M.: A quality aware sample-to-sample comparison for face recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6129–6138 (2023)
30. Sengupta, S., Chen, J.C., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: 2016 IEEE winter conference on applications of computer vision (WACV). pp. 1–9. IEEE (2016)
31. Thienpondt, J., Desplanques, B., Demuynck, K.: The IdLab VoxSRC-20 submission: Large margin fine-tuning and quality-aware score calibration in DNN based speaker verification. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5814–5818 (2021). https://doi.org/10.1109/ICASSP39728.2021.9414600
32. Wang, C., Jiang, Y., Yang, C., Liu, H., Chen, Y.: Beyond reverse KL: Generalizing direct preference optimization with diverse divergence constraints. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=2cRzmWXK9N
33. 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: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5265–5274 (2018). https://doi.org/10.1109/CVPR.2018.00552
34. Wang, K., Wang, S., Zhang, P., Zhou, Z., Zhu, Z., Wang, X., Peng, X., Sun, B., Li, H., You, Y.: An efficient training approach for very large scale face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4083–4092 (2022)
35. Wang, S., Chen, Z., Han, B., Wang, H., Liang, C., Zhang, B., Xiang, X., Ding, W., Rohdin, J., Silnova, A., et al.: Advancing speaker embedding learning: Wespeaker toolkit for research and production. Speech Communication **162**, 103104 (2024)
36. 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: proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 90–98 (2017)
37. Wuerkaixi, A., Yan, K., Zhang, Y., Duan, Z., Zhang, C.: Dyvise: Dynamic vision-guided speaker embedding for audio-visual speaker diarization. In: 2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP). pp. 1–6. IEEE (2022)
38. Xiang, X., Wang, S., Huang, H., Qian, Y., Yu, K.: Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and

Conference (APSIPA ASC). pp. 1652–1656 (2019). https://doi.org/10.1109/APSIPAASC47483.2019.9023039
39. 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), 5 (2018)
40. 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)
41. Zhou, J., Jia, X., Li, Q., Shen, L., Duan, J.: Uniface: Unified cross-entropy loss for deep face recognition. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20673–20682 (2023). https://doi.org/10.1109/ICCV51070.2023.01895
42. Zhu, Z., Huang, G., Deng, J., Ye, Y., Huang, J., Chen, X., Zhu, J., Yang, T., Du, D., Lu, J., Zhou, J.: Webface260m: A benchmark for million-scale deep face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(2), 2627–2644 (2023). https://doi.org/10.1109/TPAMI.2022.3169734