HyperSpaceX: Radial and Angular Exploration of HyperSpherical Dimensions

Chiranjeev Chiranjeev, Muskan Dosi, Kartik Thakral Mayank Vatsa, and Richa Singh

Indian Institute of Technology Jodhpur, India
{chiranjeev.1,dosi.1,thakral.1,mvatsa,richa}@iitj.ac.in
https://github.com/IAB-IITJ/HyperSpaceX

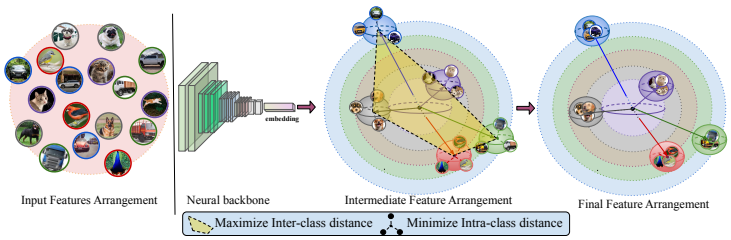

Fig. 1: Visual concept of the proposed HyperSpaceX framework: It utilizes a novel radial-angular latent space on hyperspherical manifolds to differentiate features effectively. Initially, class features are indistinguishable due to overlap. The proposed DistArc loss exhibits two feature arrangement learning: high inter-class variation employing multi-radial angular arrangement and minimal intra-class distance, leading to highly separable-discriminable class features.

Abstract. Traditional deep learning models rely on methods such as softmax cross-entropy and ArcFace loss for tasks like classification and face recognition. These methods mainly explore angular features in a hyperspherical space, often resulting in entangled inter-class features due to dense angular data across many classes. In this paper, a new field of feature exploration is proposed known as *HyperSpaceX* which enhances class discrimination by exploring both angular and radial dimensions in multi-hyperspherical spaces, facilitated by a novel *DistArc* loss. The proposed DistArc loss encompasses three feature arrangement components: two angular and one radial, enforcing intra-class binding and inter-class separation in multi-radial arrangement, improving feature discriminability. Evaluation of *HyperSpaceX* framework for the novel representation utilizes a proposed predictive measure that accounts for both angular and radial elements, providing a more comprehensive assessment of model accuracy beyond standard metrics. Experiments across seven object classification and six face recognition datasets demonstrate state-of-the-art *(SoTA)* results obtained from *HyperSpaceX*, achieving up to a 20% performance improvement on large-scale object datasets in lower dimensions and up to 6% gain in higher dimensions.

Keywords: Representation Learning · Image Classification · Face Recognition

1 Introduction

The advancement in image classification and face recognition has greatly benefited from the introduction of innovative loss functions, designed to better differentiate between class features. These functions are crafted to learn unique representations for each class by employing proxies or weight vectors to increase the distinction between classes. Traditional image classification techniques [14], [12], [24], [34] primarily rely on cross-entropy loss [1], [35] to create intricate class boundaries and improve model generalization. However, a common limitation of these approaches is their tendency to neglect the reduction of within-class feature distances while failing to adequately separate features between different classes. This results in the blending of feature points from dissimilar classes, thus leading to reduced accuracies in object classification and face verification tasks.

To address this gap, several modifications to softmax-based loss functions have been introduced, utilizing proxy-based methods to cluster similar class features while concurrently expanding the separation between dissimilar ones within the latent space. For face recognition, existing approaches [3], [8], [16], [26], [28] leverage proxy-based objectives to achieve better feature separability on a hypersphere through angular dimensions. Methods such as SphereFace [16], [29] and ArcFace [3] introduce angular margins to promote large-margin discriminative feature learning, while CosFace [28] incorporates marginal cosine functions to modify softmax loss by equalizing radial variations, introducing a cosine margin for angular decision boundaries. Furthermore, face recognition has benefited from proxy-free training strategies that focus on deep metric learning techniques [19], [22], [25], [31]. These strategies generate embeddings that bring similar faces closer together while keeping dissimilar ones apart, adapting effectively to various conditions without the need for explicit class labels. However, methods like triplet learning [22] and contrastive loss [31] depend heavily on the availability of extensive image pairs or triplets for effective training and often face difficulties with hard-pair mining. Other discriminative loss functions like Center loss [30] and Git loss [7] encounter difficulties in updating class center parameters with a large identity count, adding computational cost. Orthogonal Projection Loss (OPL) [21] and Learnable Subspace Orthogonal Projection [15] employ orthogonal projection constraints to improve separability between classes, but this approach restricts the number of classes to 2^d in d-dimensional space.

The limitations of traditional loss functions, including those modified for angular space, have led to challenges in effectively differentiating between features. The issue occurs when features of closely related classes overlap, causing uncertainties in the correct identification or classification of subjects[1]. This overlap of class features makes it difficult to discern between unique identities or classes, thereby impacting the overall performance. To overcome these limitations, we propose the *HyperSpaceX* framework, a novel approach that extends the exploration of feature space beyond the angular to include radial dimensions in multi-hyperspherical latent spaces. It introduces a loss function that emphasizes

[1] It is further analyzed in section 3

the arrangement of features to increase inter-class separability and intra-class compactness. Through the effective combination of angular and radial organization of features, the framework aims to minimize the overlap among inter-class data points, presenting a novel discriminative feature representation approach for image classification and face recognition. The key highlights are:

1. Introducing *HyperSpaceX*, a novel discriminative feature representation and arrangement learning that explores both radial and angular dimensions within multi-hyperspherical spaces, enabling effective feature learning.
2. Developing the *DistArc* loss function, aimed at enhancing the discriminative power of deep learning models. It improves feature representation by promoting better separation between different classes and tighter clustering of features within the same class in the latent space.
3. Presenting a predictive measure to evaluate the model's performance, incorporating both radial and angular dimensions in multi-hyperspherical settings. This measure provides a more comprehensive understanding of the model's capabilities in handling complex feature distributions.
4. Evaluating the proposed approach using seven object datasets (MNIST, FashionMNIST, CIFAR-10, CIFAR-100, CUB-200, TinyImageNet and ImageNet1K) and six face datasets (LFW, CFP-FP, AgeDB-30, CA-LFW, CP-LFW and D-LORD), showing its effectiveness in improving feature distinction and model accuracy across various data types and complexities, often achieving state-of-the-art results.

2 The HyperSpaceX Framework

In traditional hyperspherical angular latent spaces, identities or classes are delineated by distinct angular directions. However, with the increase in the number of classes and the volume of data for each class, overlaps and intersections in these class representations increases, leading to significant classification challenges, as they blur the distinction between different classes. To address the issue of feature overlap and the necessity for more defined class distinctions, we propose the HyperSpaceX framework. This framework utilizes a multi-hyperspherical space, drawing on both radial hyperspheres and angular dimensions to promote a discriminative distribution of class features. This learnable arrangement distribution is facilitated by the novel DistArc loss that significantly enhances class differentiation and separation along with the grouping of same-category features. DistArc strategically arranges features by considering both their angle and distance from the center, across various spherical layers. This method utilizes the diversity of spatial dimensions to establish clear and distinct areas within the spherical model, making it easier to define precise boundaries between classes. Such a multi-dimensional approach generates a more elaborate hyperspherical subspace, facilitating the creation of more efficient decision boundaries.

2.1 Preliminaries

Deep learning research has explored a variety of loss functions, among which Cross-entropy loss is most commonly used for classification tasks. Let x_i and ω be the feature embedding and weight or proxy matrix, respectively. Cross-entropy loss is defined as $L_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\omega_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{K} e^{\omega_j^T x_j + b_j}}$. Here, ω_{y_i} signifies the proxy vector of the y_i-th class. The term b refers to the bias, N represents the number of samples distributed over K number of classes. This function calculates the expected loss between the predicted and actual class distributions, focusing on optimizing class separations without necessarily bringing similar class features closer or significantly distancing different classes. Such limitations can adversely affect image classification and face recognition performance, particularly under high intra-class variations. There have been attempts to enhance inter-class distinctions through loss functions like Orthogonal Projection Loss (OPL) [21], which introduces perpendicular margins between class features. However, these efforts are often overshadowed by the dominance of cross-entropy in optimization processes. Softmax-based losses, including those used in deep face recognition like SphereFace [16] [29], CosFace [28] and ArcFace [3], primarily rely on cross-entropy based formulation. Among these is ArcFace loss, one of the most popular angular-loss functions, which distinguishes itself by allocating features around class proxies within a hyperspherical space and applying an angular margin m to enhance the cohesion within classes and the distinction between different classes. The mathematical formulation of ArcFace loss is represented as $L_{\text{ArcFace}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{cos(\theta_{y_i}+m)}}{e^{cos(\theta_{y_i}+m)}+\sum_{j=1,j\neq y_i}^{K} e^{cos\theta_j}}$.
In this, $cos\theta$ is calculated using cosine similarity between feature embedding x and proxy ω, that is defined by computing a dot product between unit vectors of x and ω as represented through $cos(\theta_{y_i}) = \hat{x}_i \cdot \hat{\omega}_{y_i} \quad cos(\theta_j) = \hat{x}_i \cdot \hat{\omega}_j$. However, focusing solely on angular dimensions can lead to ambiguity among classes in densely populated identity spaces, highlighting a need for broader exploration beyond angular metrics.

2.2 Proposed DistArc Loss

The proposed *DistArc* loss navigates through radial and angular dimensions across multiple hyperspheres centered at the origin to improve class discrimination. The *DistArc* loss is formulated as,

$$L_{\text{DistArc}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\cos(\theta_{y_i}+m)} + \cos(\phi_{y_i}) - \lambda\delta_{y_i}}{e^{\cos(\theta_{y_i}+m)} + \sum_{j=1,j\neq y_i}^{K} e^{\cos(\theta_j)} - \lambda\delta_j} \tag{1}$$

where λ represents the weighing factor and $\cos\theta$ defines the cosine of angle between embedding x and a proxy vector ω. An angular-margin penalty of m is introduced in $\cos\theta$ to enhance the feature discriminability. Fig. 2a represents the geometric illustration of $\cos(\theta_{y_i} + m)$. The DistArc loss minimizes the angle

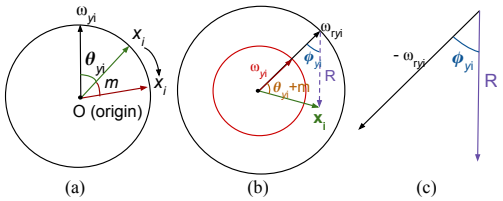

Fig. 2: The geometric interpretation of angular and radial formations in multi-hyperspherical dimensions in the training phase through (a) θ and angular-margin penalty m, and (b) angle ϕ between scaled proxy vector ω_{y_i} and resultant vector R. (c) Shows the vector representation of R and ω_{y_i} in a reverse direction for computing angle ϕ using cosine of an angle ϕ.

θ_{y_i}, thereby increasing the cosine similarity between embeddings x_i and their corresponding proxy's ω_{y_i}. However, to include the radial dimension also we incorporate the $\cos\phi_{y_i}$ which represents the cosine of angle between radii scaled proxy vector $\omega_{r_{y_i}}$ and a resultant vector R_{y_i} of $\omega_{r_{y_i}}$ and x_i. Fig. 2b describes the geometric representation of the resultant vector R_{y_i}. R_{y_i} is modified while learning/training such that it minimizes the angle between R_{y_i} and $\omega_{r_{y_i}}$ which is defined using $\omega_{r_{y_i}} = \hat{\omega}_{y_i} \times r_{y_i}$ and R_{y_i} is defined as $R_{y_i} = -\omega_{r_{y_i}} + x_i$. Here, $R_{y_i} \in \mathbb{R}^d$, $\omega_{r_{y_i}} \in \mathbb{R}^{d\times 1}$, and $x_i \in \mathbb{R}^d$. The resultant vector helps to maintain the embedding magnitude within the hyperspherical radii of magnitude $||\omega_{r_{y_i}}||_2$ and further optimizes the embeddings x_i*'s* to cluster in the angular direction by minimizing the angle (ϕ_{y_i}) between R_{y_i} and $\omega_{r_{y_i}}$: $\boldsymbol{\cos(\phi_{y_i}) = \hat{R}_{y_i} \cdot -\hat{\omega}_{r_{y_i}}}$. This equation, while calculating the $\cos\phi_{y_i}$, the origin is taken at $\hat{\omega}_{r_{y_i}}$ which results $\hat{\omega}_{r_{y_i}}$ vector to be in reverse direction leading to $-\hat{\omega}_{r_{y_i}}$. We can visualize its vector representation from Fig. 2c representing the angle computation process between the reversely directed scaled proxy vector $\omega_{r_{y_i}}$ and a resultant vector R_{y_i}. The term $\cos\phi_{y_i}$ within the proposed loss function $L_{DistArc}$ is designed to enhance the intra-class compactness of features with their respective scaled class proxies, $\omega_{r_{y_i}}$ by optimizing both the radial and angular dimensions of different hyperspheres. This term also has a major significance in letting the distribution of each class to stay within particular radii hyperspheres in the feature space.

Through *DistArc* loss, the $\cos\theta_{y_i}$ optimizes to align the embeddings with class proxies ω_{y_i} in the angular space and $\cos\phi_{y_i}$ optimizes to increase the compactness among embeddings and their respective scaled proxies $\omega_{r_{y_i}}$. It also ensures that the length of embeddings does not extend beyond their radial space r_i; further, minimizes the angle between normalized resultant vector ($\hat{R}_{y_i}$) and normalized scaled proxy ($\hat{\omega}_{r_{y_i}}$) to bring closer the embeddings at the point of scaled proxy, $\omega_{r_{y_i}}$ in the angular space. The overall geometric interpretation of *DistArc* loss during the training phase is shown in Fig. 2b.

In the proposed DistArc loss, δ_{y_i} act as a pulling force to attract embeddings x_i towards $\omega_{r_{y_i}}$. This loss component optimizes to shift the distribution of each class towards the scaled proxies $\omega_{r_{y_i}}$ in the radial space, leading to an increase in the embedding's magnitude. Leading to the enhancement of intra-class compact-

ness, δ_j in the denominator of the loss function, helps to distant the x_i's away from other scaled proxies ω_{r_j}'s excluding $\omega_{r_{y_i}}$. The overall combination of δ_{y_i} and δ_j increases the inter-class distance between dissimilar class scaled proxies and embeddings while decreasing the intra-class distance between similar class scaled proxy and embeddings. The δ_{y_i} term is defined using, $\boldsymbol{\delta_{y_i} = ||\omega_{r_{y_i}} - x_i||_2^2}$ and in case when $j \neq y_i$, $\boldsymbol{\delta_j = ||\omega_{r_j} - x_i||_2^2}$. A negative (-) sign is used in the δ component of loss Eq. 1 to maximize the logarithmic distribution overall and subsequently minimize the *DistArc* loss.

2.3 Analytical Ablation of DistArc Loss and Inductive Bias for HyperSpaceX Framework

In the proposed *DistArc* loss, the first component $\cos(\theta_{y_i} + m)$ represents the cosine of the angle θ_{y_i} between ω_{y_i} and x_i, incorporating an additional penalty m for the angular margin applied to angle θ_{y_i}, to strengthen coherence within the class while increasing differentiation between different classes. The second term, $\cos\phi_{y_i}$, calculates the angle between the scaled proxy $\omega_{r_{y_i}}$ and its corresponding resultant vector R_{y_i}. This can be expressed as the dot product between the two vectors: $\omega_{r_{y_i}}$ in reverse direction and R_{y_i}. This aims to minimize the angle ϕ_{y_i}, ensuring that feature points remain close to their scaled proxies and do not extend beyond the boundaries of their respective radial hyperspheres. The last two terms, δ_{y_i} and δ_j, help to shift the features to their corresponding scaled proxy $\omega_{r_{y_i}}$ points over different radii hyperspheres and diverging x_i features from different class proxies ω_{r_j}'s. It ensures inter-class separability in the radial direction across multi-hyperspherical manifolds. The δ_{y_i} term also results in the tighter clustering of class features around their respective $\omega_{r_{y_i}}$'s leading to high class discrimination. The combined importance of the components $\cos\theta_{y_i}$ and $\cos\phi_{y_i}$ manifests itself in the minimization of angles θ and ϕ, using a more concentrated angular distribution of features aligning to the corresponding class proxies. Further, it prevents features from extending beyond their respective class radii. The loss's additional δ component contributes to the compact clustering and length amplification of x_i vectors, shifting them to specific hypersphere radii[2].

Additionally, in the context of *HyperSpaceX* framework, the inductive bias effectively shapes the decision boundaries based on both radial and angular factors. We define the decision boundary as,

$$\begin{aligned}(\cos(\theta_1 + m) + ||x_1||_2) - (\cos(\theta_2) + ||x_2||_2) = 0 \text{ for class 1} \\ (\cos(\theta_1) + ||x_1||_2) - (\cos(\theta_2 + m) + ||x_2||_2) = 0 \text{ for class 2}\end{aligned} \tag{2}$$

This analysis of inductive bias in decision boundaries is built on binary classification and can effectively be extended to multi-class classification tasks. Consequently, the features learned through the radial-angular DistArc loss exhibit more separable decision boundaries and enhanced discriminative capabilities

[2] A detailed visual and theoretical ablation is included in the supplementary material.

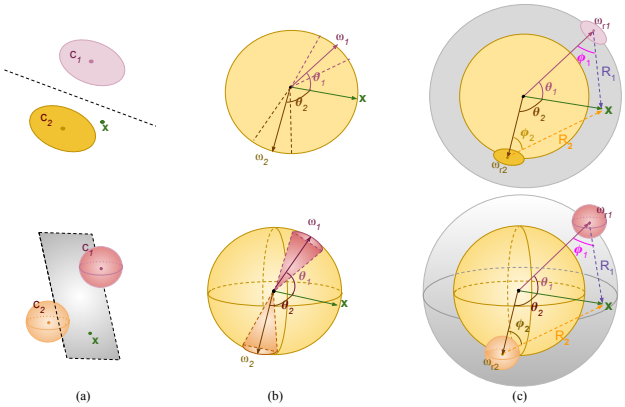

Fig. 3: The 2-D (in first row) and 3-D (in second row) latent space visualization of features learnt through (a) metric-based loss functions, (b) Angular-softmax-based loss functions, and (c) the proposed Radial-Angular DistArc loss. Further showing the decision-making process of assigning a test sample x to the most favourable class distribution represented using either class center c or proxy vector ω.

among class features. A more comprehensive derivation of the decision boundary influenced by inductive bias, elucidating its dependency on angles ϕ, θ, and resultant vector R is present in the supplemental.

3 Predictive Measure and Latent Space Visualization

The most prevailing choice for evaluating and categorizing a sample involves determining the category by calculating the maximum prediction from the model's predicted probability distribution. The proposed framework identifies the most favourable class distribution across multi-hyperspherical manifolds by considering a combination of angular and radial factors. An optimal class prediction is determined through a distinct approach that helps the evaluation metric, taking into account the radial and angular aspects of hyperspheres. The procedure encompasses the computation of resultant vectors between a sample's feature vector and each scaled proxy through mathematical formulation defined as $R_i = x - \omega_{r_i}$ $(\forall i \in 1, 2, 3, ...)$. The resultant vector with the smallest magnitude determines the most favourable class proximity. Cosine law of triangle length computation is employed to determine the length of the resultant vector R based on learned and computed angles θ, ϕ and a magnitude of feature vector x. Eq. 3 defines the formulation for R_i*'s* length computation for every i^{th} category.

$$||R_i||_2 = ||x||_2 \cos\phi_i \; + \; ||\omega_{r_i}||_2 \cos(\pi - (\theta_i + \phi_i)) \tag{3}$$

We introduce a favourable class determining predictive measure for the *HyperSpaceX* framework with its radial-angular based formulation, *i.e.*,

$$\hat{y} = \underset{R_m}{\arg\min}\{R_m \in \mathbb{R}^K : R_m\} \tag{4}$$

where $R_m \in \mathbb{R}^K$ refers to a vector comprising magnitude of resultant vectors between a sample's feature vector $x \in \mathbb{R}^d$ and scaled proxy matrix $\omega_r \in \mathbb{R}^{d \times K}$ representing the radii scaled K-class proxies, and $\hat{y}$ defines a predicted class distribution. The predicted class distribution for a specific sample is effectively represented by the class having the smallest resultant vector magnitude in relation to that sample. The derivation of a predictive measure formulation is provided in the supplementary material.

During testing, the class of a test sample is determined by its feature representation predicted by the model, influenced by the loss function during training. Fig. 3 displays visual representations of 2D and 3D latent space features, elucidating the decision-making process for classifying a test sample. This is achieved by employing a nearest distance class measure, calculated as the Euclidean distance, as shown in Fig. 3a. In Fig. 3b, the class prediction process is shown, based on angular distance within a unit hypersphere. Lastly, Fig. 3c describes the probable class selection process by finding the vector with the least magnitude across multi-hyperspherical manifolds. As shown in Fig. 4a.i, the cross-entropy loss facilitates class feature separation, leaving intermixed features that affect prediction confidence due to lack of discriminability. Conversely, angular-softmax losses, introducing angular-margin penalties to enhance cosine similarity among embeddings, aiming separation and discriminability still causes inter-mixed feature representations (depicted via Fig. 4a.ii). *HyperSpaceX* framework with *DistArc* loss, presents a solution by distributing features across various angles on multiple hyperspheres, achieving better separation and discriminability by integrating both angular and radial dimensions. This approach results in a more distinct feature arrangement between classes, as demonstrated in Fig. 4a.iii), indicating the effectiveness of *DistArc* loss in creating a more discriminable and separable feature space.

4 Experimental Setup

The experimental approach is designed to analyze the distribution of features in the latent space across various image classification and face recognition tasks. We have trained and tested the proposed *HyperSpaceX* framework on well-known benchmark datasets: (a) seven image classification datasets including MNIST [4], FashionMNIST [32], CIFAR-10 [11], CIFAR-100 [11], CUB-200 [27], TinyImageNet [13] and ImageNet-1K [2], and (b) six face recognition datasets including LFW [10], CFP-FP [23], AgeDB-30 [18], CA-LFW [37], CP-LFW [36] datasets (CASIA-WebFace [33] is used for initial model training). Model is also trained on 1500 subjects of D-LORD [17] dataset and tested on 600 different test subjects.
Implementation Details: The HyperSpaceX framework undergoes training using various deep neural network backbones, including the iResNet50 architecture [6], [9], the ViT and RN101 backbones [5] from the CLIP foundation model [20]. The models are trained using DistArc loss and several variants of softmax-based loss functions, including Cross-entropy, CosFace and ArcFace loss, and also with Orthogonal projection loss. For **image classification**, the model

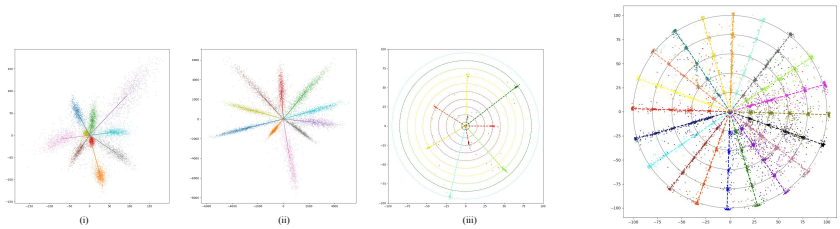

Fig. 4: (a) Illustrating comparative visual analysis of the organization of MNIST class feature distribution in the latent space. The feature representations are learned using (i) Cross-entropy loss, (ii) ArcFace loss, and (iii) the proposed *DistArc* loss on the MNIST database, where each color represents a unique class. While (b) depicts the subclass organization of the CIFAR-100 dataset on 2D multi-hyperspherical manifolds. The color of each line denotes a distinct superclass, facilitating angular separability. Subclasses within each superclass are distinguished radially, with each subclass represented as blobs extending radially from the superclass center.

undergoes both training and testing in a classification setting. We train models for small-class simple datasets like MNIST and FashionMNIST with an SGD optimizer at a learning rate of $1e-2$ and a weight decay of $5e-4$. We further set the value of $m = 0.4$ and $\lambda = 0.003$ in the DistArc loss components. For complex datasets like CIFAR-10, CIFAR-100, CUB-200 and TinyImageNet, the value of λ is set to 0.005. Other hyperparameters remain the same. For **face recognition**, the training is conducted by employing a classification setup, while during testing, face verification is performed using image pairs with class-set disjoint from training. In the face recognition task, the training is performed on the CASIA-WebFace dataset in the classification setting and testing is performed in the verification setup on LFW, CFP-FP, AgeDB-30, CA-LFW and CP-LFW datasets. The SGD optimizer with a learning rate of $1e-3$ and weight decay of $5e-4$ is utilized for both the classifier and backbone network. However, in this task λ is varied based on the number of epochs. We initialize by setting λ to 0.001 and increasing it by an addition of 0.001 after every tenth epoch and stop updating after reaching the value of 0.005.

5 Results and Analysis

Image Classification Results: We first present the results pertaining to how class features are organized across a range of embedding sizes, from low (2D) to high (512D) and very high (2048D) dimensions, distributed across radial-hyperspherical manifolds. Table 1 showcases the performance on multi-class datasets like CIFAR-100 [11] and TinyImageNet [13], achieved by employing different backbone architectures trained with a variety of loss functions. This highlights the effectiveness of our HyperSpaceX framework, which utilizes DistArc loss for training. This method enhances class distinctiveness, typically outper-

Table 1: Performance comparison (accuracy in %) on image classification task between the proposed DistArc and other loss functions utilized for training models with multiple backbones. The best and second best performances are **bolded** and underlined.

		CIFAR-100					Tiny-Imagenet				
Embedding Size	**Backbone**	**Cross Entropy**	**ArcFace**	**CosFace**	**OPL**	**DistArc**	**Cross Entropy**	**ArcFace**	**CosFace**	**OPL**	**DistArc**
2	**iResNet50** [6]	47.12	51.48	46.92	45.20	**62.06**	14.03	16.03	13.78	17.73	**20.89**
	RN101 [20]	37.84	36.02	35.12	38.85	**53.95**	17.96	16.22	18.10	20.55	**32.97**
	ViT-B [20]	58.36	56.27	53.96	57.04	**77.93**	37.58	42.94	36.25	36.49	**54.11**
	ViT-L [20]	67.89	66.44	63.81	66.27	**79.56**	45.93	44.06	49.34	46.20	**68.21**
512	**iResNet50** [6]	**83.05**	80.22	80.01	81.28	81.62	61.83	60.02	58.39	63.06	**63.27**
	RN101 [20]	69.39	70.84	68.37	69.11	**75.58**	53.04	53.01	51.77	52.85	**54.40**
	ViT-B [20]	81.99	81.58	79.46	80.92	**82.62**	**78.38**	74.22	72.61	74.42	76.03
	ViT-L [20]	88.75	87.27	85.40	84.83	**89.48**	86.49	85.55	81.84	85.94	**86.79**
2048	**iResNet50** [6]	85.22	84.29	86.82	86.14	**89.10**	61.39	63.92	62.40	62.71	**64.20**
	RN101 [20]	79.40	78.15	77.06	78.22	**80.37**	54.13	53.28	53.87	54.10	**59.06**
	ViT-B [20]	81.04	79.33	79.29	80.72	**83.00**	79.54	78.37	79.26	77.00	**80.37**
	ViT-L [20]	87.59	85.93	84.74	90.20	**90.20**	85.03	81.39	82.48	86.04	**86.94**

forming models trained with conventional loss functions such as Cross-entropy, ArcFace, CosFace, and Orthogonal Projection Loss (OPL). These results also showcase that the DistArc loss leads to features with enhanced separability and discrimination due to its effective exploration of both radial and angular dimensions within each hyperspherical latent space. This approach results in a significant performance increase: in TinyImageNet classification tasks, we observe an increase of over 22% in 2D, around 1.44% in 512D, and 4.93% in 2048D spaces, when benchmarked against cross-entropy loss and various softmax-angular margin-based loss functions. A similar trend of substantial improvement is observed for the CIFAR-100 dataset, showing improvements of up to 19.57% in 2D, 6.19% in 512D, and 2.61% in 2048D spaces.

Fig. 4b illustrates the subclass and class-based feature distribution for CIFAR-100 categories, arranging each subclass angularly within its superclass group across different radial spaces. This setup enhances radial separability for subclasses and angular separability for superclasses, leading to clear distinctions between them. Furthermore, subclasses are compactly positioned according to their radial and angular coordinates, showcasing precise class feature discrimination. This organized feature distribution demonstrates the capability for a more defined feature arrangement, optimizing representation in lower-dimensional spaces and simplifying model complexity. Consequently, Fig. 4b uses color differentiation to distinguish between unrelated classes, while same-colored points on different radii indicate subclasses belonging to the same superclass, thereby facilitating a methodical learning of class arrangements in multi-radial spaces for refined differentiation between classes and their subclasses.

The results for the CUB-200 dataset [27], a large multi-class image classification challenge focusing on bird categorization, are shown in Table 2. Classifying images in the CUB-200 dataset is particularly demanding due to the subtle, unique features each bird species has, despite the commonality of features like beaks and wings. This demands an emphasis on fine-grained, distinct characteris-

Table 2: Comparative performance (accuracy in %) analysis on CUB-200 dataset between proposed DistArc and other loss functions used for training models with multiple backbones. The best and second best performances are **bolded** and underlined.

Embedding Size	Backbone	Cross Entropy	ArcFace	CosFace	OPL	DistArc
2	**iResNet50** [6]	7.50	6.19	7.07	11.59	**18.09**
	RN101 [20]	11.44	10.47	11.42	12.84	**31.50**
	ViT-B [20]	18.70	20.63	19.04	17.06	**25.93**
	ViT-L [20]	25.48	22.85	24.19	27.77	**48.11**
512	**iResNet50** [6]	61.29	63.47	60.29	62.83	**70.48**
	RN101 [20]	51.40	50.85	50.11	54.26	**58.44**
	ViT-B [20]	67.93	68.70	68.18	70.36	**76.06**
	ViT-L [20]	80.79	78.27	79.66	80.65	**83.28**

Table 3: Evaluating model performance (accuracy in%) using conventional and proposed predictive measures in the HyperSpaceX framework when trained using DistArc loss on backbones with 512-D embedding size. The best performances are **bolded**.

Backbone	CIFAR-100		CUB-200		TinyImageNet	
	Classification layer	Radial-Angular (proposed)	Classification layer	Radial-Angular (proposed)	Classification layer	Radial-Angular (proposed)
iResNet50 [6]	80.29	**81.62**	69.56	**70.48**	61.93	**63.27**
RN101 [20]	71.05	**75.58**	58.19	**58.44**	54.33	**54.43**
ViT-B [20]	78.99	**82.62**	74.91	**76.06**	**76.10**	76.03
ViT-L [20]	88.20	**89.48**	82.85	**83.28**	82.51	**86.79**

tics of different bird body parts when developing bird image classification models. In such challenging scenario, as shown in Table 2, the *HyperSpaceX* framework outperforms existing methods. Models trained with DistArc loss achieved a significant performance boost of approximately 20.34% using a ViT-L backbone with a 2D last embedding layer. Additionally, DistArc loss trained models excelled in higher-dimensional latent spaces. This success is attributed to our approach's adeptness in discriminately mapping various bird class features across distinctively separated radii and angles in a multi-hyperspherical space, while simultaneously compacting features within the same bird class.

Reduced Latent Space Representation: Table 1 and Table 2 represent the performance on several backbones having the last embedding layer of size 2. It signifies that on complex datasets such as CIFAR-100, CUB-200 and TinyImageNet, the proposed radial-angular feature learning approach in a 2-D latent space achieves a performance boost of 16% to 20% over traditional cross-entropy loss methods even with large number of classes. Experimental results on the CUB-200 dataset also depict the high separability among various low-variance classes. Quantitatively, we can deduce that while dimensionality reduction for large complex datasets from a higher dimension of 512 and 2048 to the lower dimension of 2 embedding space, cross-entropy loss observes a $40-42\%$ performance gap while DistArc loss successfully reduced this disparity to $20-24\%$.

Predictive Measures Analysis: In the Image classification task, the class predictions are generally made utilizing the model's classification layer, adopted for

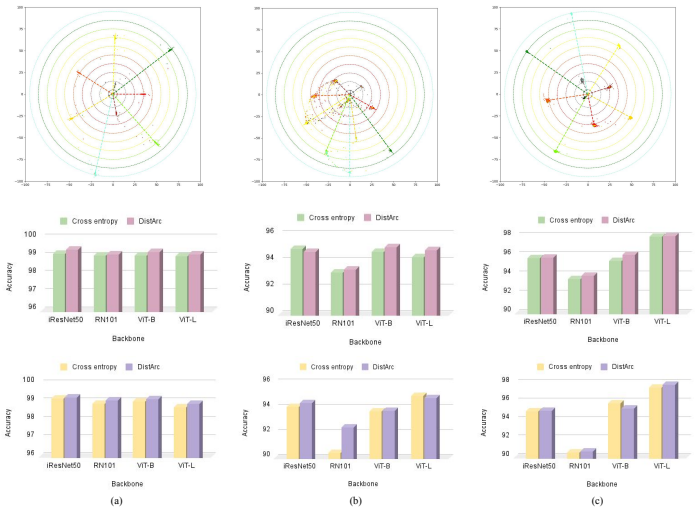

Fig. 5: Analysis of loss functions using small-class simple datasets (a) MNIST and (b) FashionMNIST, and complex dataset (c) CIFAR-10. The first row visualizes the features, showcasing the outcomes of learning with *DistArc* loss over 2D multi-spherical manifolds. The last two rows illustrate classification performance of different backbones with a 2-D and 512-D embedding sizes, trained using Cross-entropy and DistArc loss.

models trained with various existing loss functions except for DistArc loss, for results presented in Table 1 and Table 2.The proposed DistArc loss, the predictions are computed using the newly introduced radial-angular based predictive measure defined in section 3. The comparison of utilizing traditional and proposed predictive measures in the HyperSpaceX framework is shown via Table 3. From the outperforming results shown, it can be analyzed that the proposed radial-angular-based predictive measure is effective over the conventional approach of predictions from the final classification layer for models trained using DistArc loss due to different radii of classes.

Performance of HyperSpaceX on a Large-Scale Image Dataset: Table 4 provides the details the performance of the proposed framework after fine-tuning on the ImageNet-1K dataset [2]. We observe that the proposed DistArc loss outperforms existing methods for embedding sizes of 32 and 128. It also achieves the second-highest performance at 512 dimensions. As the number of classes grows, ArcFace finds it difficult to manage many feature points on a unit hypersphere. This demonstrates the benefit of strategically arranging features in a multi-radial hyperspherical space using DistArc loss during model training.

HyperSpaceX Performance on Datasets with Fewer Classes: Fig. 5 showcases the performance and feature visualization for MNIST, FashionMNIST, and CIFAR-10 datasets. The top row of Fig. 5 illustrates the 2D-latent space feature organization across multi-radial hyperspherical manifolds, highlighting the

Table 4: Classification Accuracy (%) on ImageNet-1K dataset using iResNet50 [6]. The best and second best performances are **bolded** and underlined.

Embedding Size	Cross Entropy	ArcFace	DistArc
32	48.25	43.05	**55.91**
128	65.18	58.62	**65.30**
512	**69.27**	59.67	67.90

Table 5: Ablation study of *DistArc* loss, reporting classification performance (in %) for various image classification datasets.

Loss Components	MNIST	Fashion-MNIST	CIFAR-10
$\cos(\theta)$	98.97	93.69	94.81
$\cos(\theta)$ & $\cos(\phi)$	98.93	93.72	95.16
$\cos(\theta)$ & δ	99.01	94.83	95.31
$\cos(\theta)$ & $\cos(\phi)$ & δ	99.19	95.04	96.03

strategic angular positioning. This arrangement emphasizes the significant distance between features of unrelated classes and the closeness of features within the same class, facilitating accurate class predictions. The subsequent rows display a comparison of classification accuracy across different models trained using cross-entropy and DistArc loss functions in 2D and 512D latent spaces. These findings illustrate the superior performance of the *HyperSpaceX* framework across both simple (MNIST, FashionMNIST) and more complex (CIFAR-10) datasets with a smaller number of classes.

Training Convergence Analysis: The training convergence rates of the DistArc loss were benchmarked against existing loss functions. Training with softmax cross-entropy loss shows only a slight decrease in loss values. In contrast, DistArc loss significantly restructures the feature space layout by dispersing identities across various radial hyperspheres. This approach leads to higher initial loss figures that efficiently reduce to a minimum of 0.22, surpassing the reduction achieved with softmax cross-entropy (0.34), CosFace (2.31), ArcFace (1.51), and OPL (0.68) loss functions. A detailed comparison illustrating the training convergence rates of loss functions is available in the supplementary file.

Ablatic Analysis of *DistArc*: Table 5 shows the ablation analysis on three image classification datasets, MNIST, FashionMNIST, and CIFAR-10, while utilizing ViT-B architecture (embedding size 512). The first and second row of Table 5 shows the efficacy of performance improvement over angular separation through θ and ϕ, and the third row shows the improvement with radial and angular factors due to δ and θ terms. While, the last row shows the highest performance achieved by including all θ, ϕ and δ components into the loss formulation, highlighting the importance of each component.

Face Recognition Results: Table 6 showcases detailed comparative results of the *DistArc* loss against other loss functions in face recognition tasks. Utilizing the iResNet50 architecture and training on the CASIA-WebFace dataset, we have evaluated verification performance with a 512-dimensional embedding. The *DistArc* loss demonstrates state-of-the-art (SoTA) results on LFW and CP-LFW datasets and near state-of-the-art results on other datasets. Additionally, experiments with the iResNet50 backbone on the MS1Mv2 dataset—a derivative of the now-withdrawn MS-Celeb dataset—were conducted to ensure a fair comparison with SoTA methods are included in supplementary materials. The *DistArc* loss trained model achieved following results: 99.82% on LFW and 98.21% on AgeDB-

Table 6: Quantitative results (in %) of face recognition models trained using different loss functions. The best and second best performances are **bolded** and underlined.

Loss	LFW	CFP-FP	AgeDB-30	CA-LFW	CP-LFW
Triplet	98.98	91.90	89.98	-	-
Center loss [30]	99.28	-	-	85.48	77.48
Softmax	99.08	94.39	92.33	88.17	84.85
Norm-Softmax	98.56	89.79	88.72	-	-
SphereFace [16]	99.11	94.38	91.70	92.55	90.90
CosFace (m=0.35) [28]	99.10	95.44	92.98	92.83	91.03
ArcFace (m=0.45) [3]	99.46	**95.47**	**94.93**	92.47	90.85
SphereFace2 [29]	99.50	-	93.68	**93.47**	91.07
DistArc (m=0.4)	**99.54**	95.41	94.03	93.32	**91.10**

30, surpassing results from SphereFace, CosFace, ArcFace, and SphereFace2[3]. These findings highlight the *HyperSpaceX* framework's proficiency in crafting highly discriminative and distinct features within the latent space.

The HyperSpaceX framework is also evaluated on the D-LORD [17] dataset, a large-scale open-set surveillance dataset with 600 test subjects, using a deep metric learning approach [19]. With ArcFace, a Rank-1 identification accuracy of 61.84% was achieved, while DistArc achieved 62.71%. For Rank-5 accuracy, the results were 65.89% and 66.03%, respectively, indicating that DistArc is more effective at learning enhanced discriminative features than ArcFace.

6 Conclusion

In this research, we introduce the *HyperSpaceX* framework, which navigates both angular and radial dimensions, facilitating a uniquely distinguishable and discernible arrangement of class features within multi-hyperspherical manifold space. This is achieved using the novel *DistArc* loss, which leverages the radial and angular components of the feature space. This loss function significantly enhances the cohesion among similar class features while concurrently maximizing the distances between different classes, resulting in more precise and discriminative feature representation. To evaluate the performance of this radial-angular framework, we introduce a predictive measure that utilizes the shortest resultant vector between the embeddings of test samples and proxy vectors, offering a deeper insight into the model's performance. The efficacy of *HyperSpaceX* is demonstrated through comprehensive experiments on seven image classification datasets and six face recognition datasets. The proposed methodology shows improvements in handling a diverse array of image and face datasets, ranging from simpler to more complex and large-scale collections encompassing several classes and samples.

[3] SphereFace (LFW: 99.42%), CosFace (LFW: 99.73%), ArcFace (LFW: 99.82%, AgeDB30: 98.15%) and SphereFace2 (LFW: 99.50%, AgeDB30: 93.68%).

Acknowledgements

Chiranjeev and Thakral received partial support from the PMRF Fellowship, and Vatsa is partially supported by the Swarnajayanti Fellowship.

References

1. Conniffe, D.: Expected maximum log likelihood estimation. Journal of the Royal Statistical Society: Series D (The Statistician) **36**(4), 317–329 (1987)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 248–255. IEEE (2009)
3. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4690–4699 (2019)
4. Deng, L.: The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE signal processing magazine **29**(6), 141–142 (2012)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Duta, I.C., Liu, L., Zhu, F., Shao, L.: Improved residual networks for image and video recognition. In: Int. Conf. Pattern Recog. pp. 9415–9422. IEEE (2021)
7. Gallo, I., Nawaz, S., Calefati, A., Janjua, M.K.: Git loss for deep face recognition. In: Brit. Mach. Vis. Conf. p. 313. BMVA (2018)
8. Guo, G., Zhang, N.: A survey on deep learning based face recognition. Comput. Vis. Image Underst. **189**, 102805 (2019)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 770–778 (2016)
10. 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)
11. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv. Neural Inform. Process. Syst. **25** (2012)
13. Le, Y., Yang, X.: Tiny imagenet visual recognition challenge. CS 231N **7**(7), 3 (2015)
14. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4), 541–551 (1989)
15. Li, L., Zhang, Y., Huang, A.: Learnable subspace orthogonal projection for semi-supervised image classification. In: ACCV. pp. 477–490. Springer (2022)
16. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 212–220 (2017)
17. Manchanda, S., Bhagwatkar, K., Balutia, K., Agarwal, S., Chaudhary, J., Dosi, M., Chiranjeev, C., Vatsa, M., Singh, R.: D-lord: Dysl-ai database for low-resolution disguised face recognition. IEEE Trans. on Biom., Behav., and Ident. Sci. (2023)

18. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: the first manually collected, in-the-wild age database. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. pp. 51–59 (2017)
19. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4004–4012 (2016)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. pp. 8748–8763. PMLR (2021)
21. Ranasinghe, K., Naseer, M., Hayat, M., Khan, S., Khan, F.S.: Orthogonal projection loss. In: Int. Conf. Comput. Vis. pp. 12333–12343 (2021)
22. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 815–823 (2015)
23. Sengupta, S., Chen, J.C., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: Winter Conf. on App. of Comput. Vis. pp. 1–9. IEEE (2016)
24. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229 (2013)
25. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. Adv. Neural Inform. Process. Syst. **29** (2016)
26. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. Adv. Neural Inform. Process. Syst. **27** (2014)
27. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: Caltech-ucsd birds-200-2011 (cub-200-2011). Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
28. 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: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5265–5274 (2018)
29. Wen, Y., Liu, W., Weller, A., Raj, B., Singh, R.: Sphereface2: Binary classification is all you need for deep face recognition. In: Int. Conf. Learn. Represent. (2022)
30. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Eur. Conf. Comput. Vis. Springer (2016)
31. Wu, C.Y., Manmatha, R., Smola, A.J., Krahenbuhl, P.: Sampling matters in deep embedding learning. In: Int. Conf. Comput. Vis. pp. 2840–2848 (2017)
32. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017)
33. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
34. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Eur. Conf. Comput. Vis. pp. 818–833. Springer (2014)
35. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. Adv. Neural Inform. Process. Syst. **31** (2018)
36. 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)
37. Zheng, T., Deng, W., Hu, J.: Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. CoRR **abs/1708.08197** (2017)