Hyperspherical Consistency Regularization

Cheng Tan[1,2,*], Zhangyang Gao[1,2,*], Lirong Wu[1,2], Siyuan Li[1,2], Stan Z. Li[1,2]
[1] AI Lab, School of Engineering, Westlake University
[2] Institute of Advanced Technology, Westlake Institute for Advanced Study
*{tancheng,gaozhangyang,wulirong,lisiyuan,stan.zq.li}@westlake.edu.cn

Abstract

Recent advances in contrastive learning have enlightened diverse applications across various semi-supervised fields. Jointly training supervised learning and unsupervised learning with a shared feature encoder becomes a common scheme. Though it benefits from taking advantage of both feature-dependent information from self-supervised learning and label-dependent information from supervised learning, this scheme remains suffering from bias of the classifier. In this work, we systematically explore the relationship between self-supervised learning and supervised learning, and study how self-supervised learning helps robust data-efficient deep learning. We propose hyperspherical consistency regularization (HCR), a simple yet effective plug-and-play method, to regularize the classifier using feature-dependent information and thus avoid bias from labels. Specifically, HCR first project logits from the classifier and feature projections from the projection head on the respective hypersphere, then it enforces data points on hyperspheres to have similar structures by minimizing binary cross entropy of pairwise distances' similarity metrics. Extensive experiments on semi-supervised and weakly-supervised learning demonstrate the effectiveness of our method, by showing superior performance with HCR.

1. Introduction

The last decade has witnessed revolutionary advances in deep learning across various computer vision fields such as image classification [26,29,38,72], object detection [48,64–66], and semantic segmentation [25,47,67] in the presence of large-scale labeled datasets. However, massive collection and accurate annotation of datasets are time-consuming and expensive. In many practical situations, only small-scale high-quality labeled datasets are available. For this reason, semi-supervised learning (SSL) that learning from few labeled data and a large number of unlabeled data has received broad attention [4, 5, 42, 62, 63, 69, 71, 73, 84, 85].

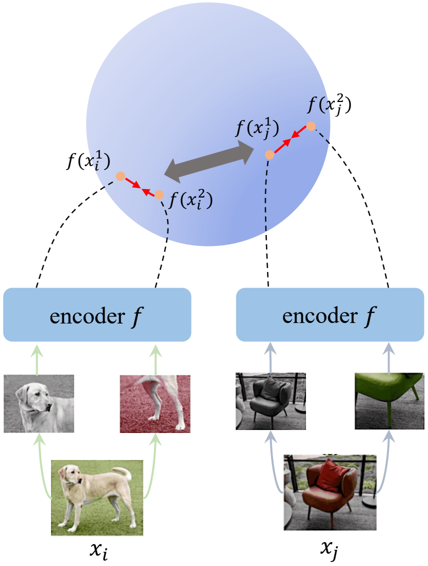


Figure 1. Illustration of contrastive learning on the hypersphere. Red arrows denote positive pairs tend to attract each other, and the gray arrow denotes negative pairs tend to repel each other.

With the development of contrastive learning [7, 9–11, 20, 24, 27, 28, 45, 74, 80, 90, 94], recent SSL algorithms [23, 34, 43, 70, 79, 81, 92] tend to extend self-supervised learning into supervised learning by adding a branch network as a projection head that jointly learns from feature-dependent and label-dependent information. Though the feature encoder is supposed to learn better by making agreements from different views on latent spaces, the classifier which determines the ultimate predictions still suffers from the bias of semi-supervision or weak-supervision. Typically, [33, 93] found that data imbalance is not the key issue in learning high-quality representations from long-tail data, while simply adjusting the classifier with balanced sampling can effectively alleviate the imblanced bias. This phenomenon suggests that decent representation may help but not be enough for robust learning, while regularizing the classifier is necessary to improve learning performance.

[*] Equal contribution

A vast number of current empirical contrastive learning methods [9–11,20,24,28,45] project feature embeddings on a hypersphere through ℓ_2 normalization while maximizing distances between negative pairs and minimizing distances between positive pairs, as shown in Figure 1. Restricting the output space to a unit hypersphere can improve training stability in machine learning where dot products are ubiquitous [77, 80, 86]. Besides, well-clustered features on the hypersphere are linearly separable from the rest of the feature space. The above desirable traits are considered to be useful while regularizing the classifier.

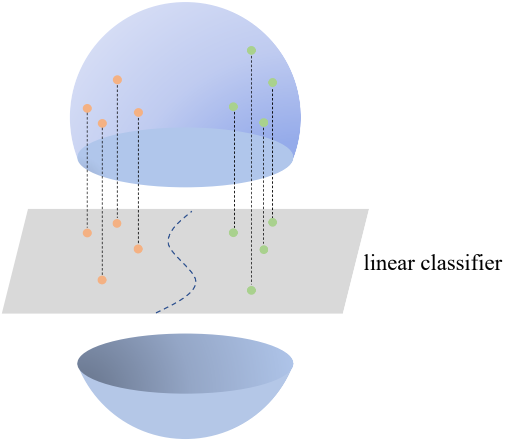


Figure 2. Linear classifier learns to separate the hypersphere through the hyperplane.

In this work, we analyze the relationship between the projection head and the classifier, and propose hyperspherical consistency regularization (HCR) to constrain the latent hyperspherical space. As shown in Figure 2, a decent classifier is able to find an optimal hyperplane in a hypersphere manifold, and data points on the classifier's hyperplane can be reprojected on a hypersphere. HCR assumes data points on the projection head's hypersphere and the classifier's hypersphere have similar geometric structures, and preserves such structures by making distributions of pairwise distances consistent. Experiments on semi-supervised learning and weakly-supervised learning indicate HCR can considerably improve the generalization ability.

2. Related works

2.1. Contrastive learning

Self-supervised learning designs pretext tasks [19, 58, 61, 91] to produce supervision signals derived from the data itself, while contrastive learning is its subset that aims to group similar samples closer and diverse samples away from each other [7,9–11,20,24,27,28,45,74,80,83,90,94]. Inspired by the important technique ℓ_2 normalization on metric learning [52, 68, 77], [83] takes the class-wise supervision to the extreme of instance-wise supervision and tries to maximally scatter the features of samples over the unit hypersphere. Most of the subsequent works [7, 9–11, 20, 24, 27, 28, 45] on contrastive learning employ ℓ_2 normalization as a standard setting, while [80] highlights ℓ_2 normalization helps contrastive learning optimize uniformity of the induced distribution of the features on the hypersphere together with the alignment of features from positive pairs. Representation learning benefits from the desirable traits of placing features on the unit hypersphere that improving training stability and separable ability.

Extending contrastive learning to semi-supervised learning or weakly-supervised learning is straightforward. SupCon [34] proposes class-wise contrastive loss under fully-supervised setting and inspires researchers to focus on the power of contrastive learning in supervised scenarios. Self-Tuning [81] explores group contrastive learning and tackles confirmation bias and model shift issues in an efficient one-stage framework towards data-efficient transfer learning and semi-supervised learning. CoMatch [43] unifies contrastive learning, consistency regularization, entropy minimization and graph-based SSL to mitigate confirmation bias in pseudo-label-based semi-supervised learning. PSC [79] proposes a hybrid network that jointly performs both self-supervised learning and prototypical supervised contrastive learning in a cumulative learning manner. BalFeat [32] combines strengths of supervised methods and contrastive methods to learn representations that are both discriminative and balanced. MoPro [44] simultaneously optimizes classical supervised loss and prototypical contrastive loss using momentum prototypes and tries to achieve robust weakly-supervised learning. Co-learning [71] rethinks that the co-training-based noisy label learning methods provide limited information gain since the differences between two networks of the same architecture mainly come from random initialization. Thus, this method explores intrinsic similarity and structural similarity to combat noisy labels.

These methods [12, 32, 43, 71, 79, 81] have similar architectures that concurrently optimize typical contrastive learning and supervised learning with certain techniques. We can see this manner from a different perspective: both label-dependent supervised learning and feature-dependent contrastive learning are pretext tasks that aim to learn proper representations. Previous works [2, 17] combine different pretext tasks to boost self-supervised learning performance and find there exist relationships between pretext tasks in which regularization is in need. Thus, our method regularizes the implicit connections between supervised learning and self-supervised learning in hypersphere space as shown in Figure 3. HCR builds a bridge between classical supervised learning and pretext tasks in self-supervised learning, and the regularization is plug-and-play to apply in these joint-learning methods.

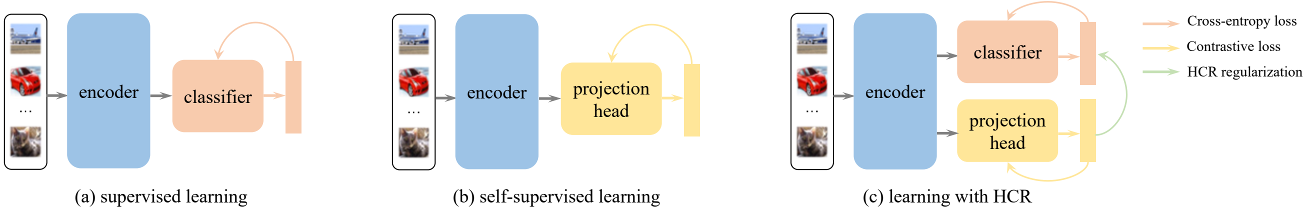


Figure 3. Conceptual illustration of different learning paradigms. We suppose supervised learning and self-supervised learning learn proper representations through different pretext tasks relying on label-dependent and feature-dependent information respectively. HCR takes supervised learning as the primary task and forces self-supervised learning to assist it from another perspective.

2.2. Learning on the hypersphere

There are quite a number of methods that learn representations on the hypersphere [8,15,16,30,46,49–51,53,54,60,78,87] and show that the key semantics in neural networks is angular information instead of magnitude. MHE [49] draws inspiration from the Thomson problem to regularize networks with a minimum hyperspherical energy objective for improving the generalization ability of networks. CoMHE [46] shows that naively minimizing hyperspherical energy suffers from difficulties due to highly nonlinear and non-optimization, and proposes projecting neurons to suitable subspaces where hyperspherical energy can get minimized efficiently. Moreover, Johnson-Lindenstrauss lemma [14] establishes a guarantee for CoMHE's projections. SphereGAN [60] remaps Euclidean feature spaces into the hypersphere by geometric transformation and calculates geometric moments for minimizing the multiple Wasserstein distances of probability measures on the hypersphere. Our work reprojects the Euclidean feature space of the classifier into a hypersphere and explores its connection with the projection head's hypersphere.

3. Methods

3.1. Preliminaries

HCR focuses on regularizing the manner that is jointly training supervised learning and self-supervised learning and tries to find their relationships. Suppose $\mathcal{X} \subset \mathbb{R}^n$ is the n-dimensional Euclidean image space, and $\mathcal{Y} = \{0,1\}^c$ is the ground-truth label space with c classes in an one-hot manner if the label exists. The usual framework consists of a classifier $g : \mathbb{R}^{D_f} \to \mathbb{R}^{D_g}$ and a projection head $h : \mathbb{R}^{D_f} \to \mathbb{S}^{D_h-1}$ with their shared feature encoder $f : x \in \mathcal{X} \to \mathbb{R}^{D_f}$, where D_f, D_g, D_h denote the dimension of output Euclidean spaces from f, g, h respectively. HCR imposes constraints in hyperspherical spaces so that the classifier $g : \mathbb{R}^{D_f} \to \mathbb{S}^{D_g-1}$ outputs a (D_g-1)-dimensional hypersphere $\mathbb{S}^{D_g-1}$ by mapping the original outputs to ℓ_2 normalized feature vectors of dimension D_g.

3.2. Hyperspherical consistency regularization

As the classifier $g(\cdot)$ and the projection head $h(\cdot)$ perform different tasks according to the same features from the feature encoder $f(\cdot)$, HCR assumes there exists a distance-preserving mapping $\mathcal{F} : \mathbb{R}^{D_h} \to \mathbb{R}^{D_g}$ and its inverse mapping $\mathcal{F}^{-1} : \mathbb{R}^{D_g} \to \mathbb{R}^{D_h}$ that establish the connections of points on the hyperspheres. We argue that the relationship of the hyperspheres from different tasks can be characterized by the geometric property. Here, we consider the pairwise distance as the key geometry property, and force points on the classifier's hypersphere to have a similar structure as the projection head's, as shown in Figure 4.

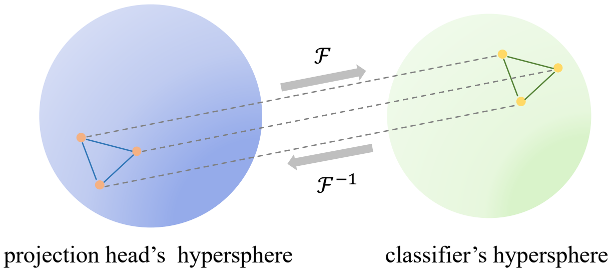


Figure 4. Preserve the geometry structure of data points lies on hyperspheres by their pairwise distances.

We first define the pairwise distances on the respective hyperspheres as:

$$\begin{aligned} d_g(x_i, x_j) &= \|g \circ f(x_i) - g \circ f(x_j)\|, \\ d_h(x_i, x_j) &= \|h \circ f(x_i) - h \circ f(x_j)\|, \end{aligned} \tag{1}$$

where $\|\cdot\|$ denotes Euclidean distance. While $x_i, x_j \in \mathcal{X}$ and $i \neq j$, we use d_g and d_h to represent the sets of pairwise distances on $\mathcal{X}$ for notation simplicity. To measure the pairwise distances on hyperspheres with different dimensions, we define similarity metrics $p(d_g)$ and $q(d_h)$ that are considered to be normal distributions multiplied by constant terms (see Sec 4.1):

$$
\begin{aligned}
p(d_g) &= C_g \frac{1}{\sigma_g\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(d_g-\mu_g)^2}{\sigma_g^2}\right], \\
q(d_h) &= C_h \frac{1}{\sigma_h\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(d_h-\mu_h)^2}{\sigma_h^2}\right],
\end{aligned}
\tag{2}
$$

where C_g, C_h are constants that forces the similarity metric to be in $[0,1]$. $\sigma_g, \mu_g, \sigma_h, \mu_h$ can be chosen according to the situations. For the convenience of optimization, we empirically assumes $p(d_g), q(d_h) \sim N(0, \frac{1}{2})$ in all experiments except especially mentioned.

As contrastive learning tries to push away samples from different classes and pull together samples from the same classes, the respective similarity metrics are supposed to be approaching either zero or one. Thus, we define the objective of HCR that minimizing the binary cross entropy (BCE) between $p(d_g)$ and $q(d_h)$:

$$
\begin{aligned}
&\mathrm{HCR}(p(d_g), q(d_h)) = \mathrm{BCE}(p(d_g) \mid\mid q(d_h)) \\
&= -p(d_g)\log q(d_h) - (1-p(d_g))\log(1-q(d_h)),
\end{aligned}
\tag{3}
$$

Through minimizing Equation 3, the mutual information $I(g \circ f(x), h \circ f(x))$ between logits $g \circ f(x)$ and feature projections $h \circ f(x)$ is implicitly maximized (see Sec 4.2).

3.3. HCR as a regularization for learning

Now that we have introduced the formulation of HCR, here we propose HCR as a regularization for semi-supervised or weakly-supervised learning. While HCR imposes the consistency between the classifier and the projection head on the hyperspherical latent spaces, it is suitable for the jointly learning manner that performs supervised learning and self-supervised learning simultaneously. In such a setting, the entire objective function can be represented as:

$$
\begin{aligned}
\mathcal{L} = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \mathcal{L}_s(x,y) + \sum_{x\in\mathcal{X}} \mathcal{L}_u(x) \\
+ \mathrm{HCR}(p(d_g), q(d_h)),
\end{aligned}
\tag{4}
$$

where $\mathcal{L}_s$ denotes the supervised loss for the labeled data, and $\mathcal{L}_u$ denotes the contrastive loss (i.e., the commonly-used InfoNCE [59]) for the unlabeled data. HCR regularizes supervised learning and explores its connections with contrastive learning so that both $\mathcal{L}_s$ and $\mathcal{L}_u$ are needed.

3.4. Relations to supervised contrastive learning

Similar to SupCon [34], HCR leverages contrastive learning to benefit vanilla supervised learning. While SupCon explicitly pushes apart clusters of samples from different classes and pulls together clusters of samples belonging to the same class in a self-supervised contrastive manner, HCR implicitly makes agreements on latent hyperspherical spaces between contrastive learning and supervised learning. Moreover, SupCon directly imposes contrastive loss to standard cross entropy, which requires a full exploration of label information so that it is limited in supervised learning fashion. HCR forces supervised learning to imitate contrastive learning in latent spaces without extra label information so that it can conveniently fit into semi-supervised learning and weakly-supervised learning frameworks.

4. Theoretical Insights

This section is inspired by rigorous theoretical results from [1, 3, 6, 13, 14, 18, 21, 31, 35, 39, 55, 75] and provides theoretical and intuitive perspectives about HCR.

4.1. Distribution of distance on a hypersphere

Theorem 1. (Asymptotic form of the distribution of Euclidean distance in a hypersphere for large dimensions). *Given the D-dimensional hypersphere $\mathbb{S}^D(a)$ with radius a, x_i and $x_j (\forall i \neq j)$ are any two points chosen at random in $\mathbb{S}^n(a)$ whose Euclidean distance is denoted by $r (0 \leq r \leq 2a)$. Then, the asymptotic distribution of r is $N(\sqrt{2}a, \frac{a^2}{2D})$ as $D \to \infty$.*

Theorem 1, which has been heavily studied in [1,21,55], tells us the distribution of Euclidean distance in a hypersphere obeys normal distribution as the dimension of the hypersphere becomes large. Though HCR only considers the case that points on the hypersphere surface and ignores the points inside, it still agrees with this theorem. HCR models the pairwise distance distribution as normal distribution and tries to utilize pairwise distances as a key property for preserving the geometric structure of the projection head's outputs. Thus, HCR builds a bridge between feature-dependent information and label-dependent information.

4.2. Connections between hyperspheres

Theorem 2. (Johnson-Lindenstrauss lemma). *Let $\epsilon \in (0,1)$. Let $N, D_g \in \mathbb{N}$ such that $D_g \leq C\epsilon^{-2}\log N$, for a large enough absolute constant C. Let $H \subseteq \mathbb{R}^{D_h}$ be a set of N points. There exists a linear mapping $\mathcal{F} : \mathbb{R}^{D_h} \to \mathbb{R}^{D_g}$, such that for all $h_i, h_j \in H$:*

$$
(1-\epsilon)||h_i - h_j||^2 \leq ||\mathcal{F}\circ h_i - \mathcal{F}\circ h_j||^2 \leq (1+\epsilon)||h_i - h_j||^2. \tag{5}
$$

The famous Johnson-Lindenstrauss lemma has found numerous applications that includes searching for graph embedding, manifold learning and dimension reduction. Here, this lemma guarantees the projection of two points from high-dimensional space to low-dimensional space preserves their Euclidean distance with high probability. Though the dimensions of feature projections $h \circ f(x)$ and logits $g \circ f(x)$ are usually not the same, HCR preserves their relative distances under this theorem.

Theorem 3. (Mutual information's invariance property to reparametrization of the marginal variables). If $H' = \mathcal{F}(H)$ and $G' = \mathcal{T}(G)$ are homeomorphisms (i.e. smooth uniquely invertible maps), then the mutual information $I(H, G) = I(H', G')$.

This theorem [13,35,75] reveals the possible connections beween $g \circ f(x)$ and $h \circ f(x)$. We discuss that Equation 3 is preserving pairwise distances between the projection head's and the classifier's hyperspherical spaces. For those limited data points, we consider the mapping $\mathcal{F}$ from the projection head h to the classifier g is approaching bijective through preserving pairwise distances. Since when $\mathcal{F}$ is invertible, the mutual information is:

$$\begin{aligned} I(h \circ f(x), g \circ f(x)) &= I(h \circ f(x), \mathcal{F} \circ (h \circ f(x))) \\ &= I(h \circ f(x), h \circ f(x)) \end{aligned} \tag{6}$$

so that it is maximized. Thus, HCR preserves the distributions of pairwise distances that implicitly maximizes the mutual information $I(h \circ f(x), g \circ f(x))$.

5. Experiments

To validate the effectiveness of our proposed HCR, we conduct experiments on various tasks, such as semi-supervised learning, fine-grained classification, and noisy label learning, among which the latter two tasks belong to weakly-supervised learning.

5.1. Baselines

We take the recent typical works Self-Tuning [81] and Co-learning [71] as our baselines on account of their jointly learning manner, that is, both of them build the network architecture using a shared feature encoder with a classifier and a projector head and train two heads simultaneously though in different ways.

Self-Tuning unifies the exploration of labeled data and unlabeled data and the transfer of a pretrained model in a pseudo group contrast (PGC) mechanism. The vanilla contrastive learning maximizes the similarity between query q with its corresponding positive key k_0 (a different view of the same data sample):

$$\mathcal{L}_{\text{CL}} = -\log \frac{\exp(q \cdot k_0/\tau)}{\exp(q \cdot k_0/\tau) + \sum_{d=1}^{D} \exp(q \cdot k_d/\tau)}, \tag{7}$$

where τ is a hyperparameter for temperature scaling. Self-Tuning modifies the contrastive mechanism by introducing a group of positive keys from samples with the same pseudo label to contrast with other samples as follows:

$$\mathcal{L}_{\text{PGC}} = -\frac{1}{D+1} \sum_{d=0}^{D} \log \frac{\exp(q \cdot k_d^{\hat{y}}/\tau)}{\exp(q \cdot k_0^{\hat{y}/\tau}) + \text{Neg}}, \tag{8}$$

where $\text{Neg} = \sum_{c=1}^{\{1,2,\ldots,C\}\setminus\hat{y}} \sum_{j=1}^{D} \exp(q \cdot k_j^c/\tau)$, D is the number of classes, $\hat{y}$ denotes the pseudo label. Equation 7 and Equation 8 are corresponding to the unlabeled loss $\mathcal{L}_u$ in Equation 4.

Self-Tuning has bridged supervised learning and self-supervised learning by guiding the contrastive mechanism of the projection head through pseudo labels produced by the classifier. However, this scheme can only help the feature encoder obtain decent representations while the bias of the classifier remains unavoidable. Thus, we expect the projection head to benefit the classifier as well through HCR as shown in Figure 5.

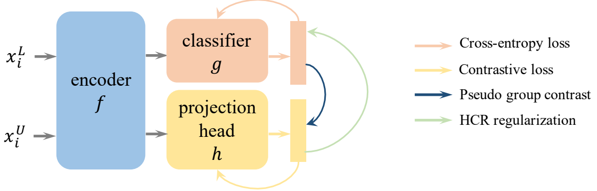
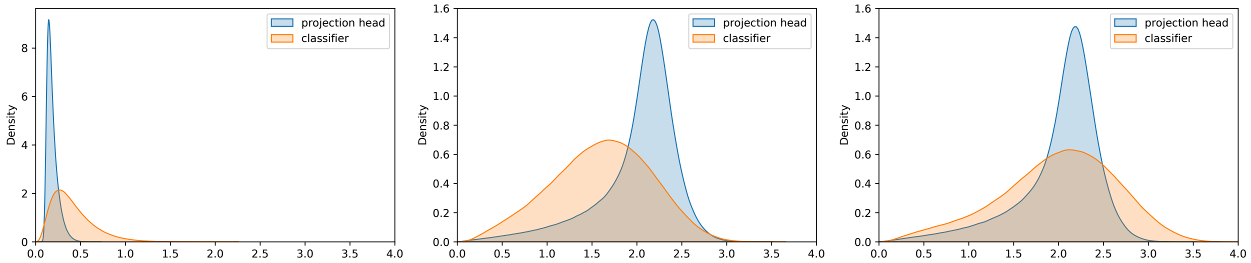


Figure 5. The illustration of Self-Tuning with HCR. x_i^L denotes data samples from the labeled dataset, and x_i^U denotes data samples from the unlableed dataset. PGC uses pseudo label information from the classifier to guide the projection head, while HCR reversely uses projections to correct the classifier.

Co-learning is a recent work that challenges the co-training scheme in noisy label learning. It provides perspectives from both supervised learning and self-supervised learning through the above mentioned jointly learning manner. As the classifier is extremely unreliable in noisy learning settings, Co-learning also proposes a structural similarity that imposes structure-preserving constraints similar to HCR. Co-learning directly assumes the pairwise distance follows a normal distribution and minimizes the Kullback–Leibler (KL) divergence of the distributions between the projection head and the classifier.

However, Co-learning cannot theoretically guarantees that the structure-preserving constraints are well performed. As shown in Figure 6, we train Co-learning and HCR on the CIFAR-10 dataset with 80% symmetric noise for a hundred epochs, and present the hyperspherical distance distributions. The initial hyperspherical pairwise distances of the projection head are roughly in $[0, 0.5]$. After training, they become large, which suggests the conclusion of [80] is correct, i.e., contrastive learning is trying to make features uniformly distributed on the hypersphere rather than concentrating in a local area. Moreover, the distance distribution learned by HCR preserves the structure better than Co-learning's because the classifier's distance distribution is balanced referring to the projection head's.



(a) The initial distance distribution.

(b) The distance distributions learned by Co-learning.

(c) The distance distributions learned by HCR.

Figure 6. The hyperspherical distance distributions of the CIFAR-10 dataset with 80% symmetric noise.

5.2. Semi-supervised learning

Following the same experimental setting as [81], we compare Self-Tuning with HCR against three classical semi-supervised learning methods: Pseudo-Labeling [41], Π-model [40], and Mean Teacher [73], as well as three recent methods UDA [85], FixMatch [69], SimCLRv2 [10], and Self-Tuning itself [81]. Experiments are conducted on three mainstream visual datasets: Stanford Cars [36], FGVC Aircraft [56], CUB-200-2011 [76], and CIFAR-100 [37]. Stanford Cars contains 16185 images of 196 classes of cars, and the pixel resolution is 360×240. FGVC Aircraft consists of 10000 images of 100 different aircraft model variants. The image resolution is about 1-2M pixels, but its width and height are not fixed. CUB-200-2011 is a dataset with totally 6033 images of 200 bird species, and each image has about less than 250 thousand pixels. CIFAR-100 is a classical visual dataset with 100 classes and 600 images per class, and the image resolution is 32×32.

For a fair comparison, all these methods implement a ResNet-50 model and initialize it from ImageNet-pretrained weights. Besides, we remove the last layer of the pretrained model and add the projection head h and classifier g with randomly initialized weights. The default temperature τ is 0.07, and the learning rate is 0.001. The optimizer follows the original Self-Tuning, which is SGD with a momentum of 0.9. Experiments are repeated three times with different random seeds, and we report the average test accuracy of three trials for each experiment. When we reproduce Self-Tuning, we unexpectedly find the results are better than its paper reported, thus we honestly *report the reproduced results rather than using results from its paper*.

As reported in Table 1, HCR significantly improves the performance of Self-Tuning by an average of 2.30% in different label proportions. Moreover, HCR obtains averagely 3.77% improvements under the condition of 15% label proportion, which indicates the effectiveness of HCR with extremely few labels.

Table 1. Classification accuracy (%) ↑ of semi-supervised learning methods on Stanford Cars dataset (ResNet-50 pretrained).

Method		Label Proportion	
	15%	30%	50%
Pseudo-Labeling	40.93±0.23	67.02±0.19	78.71±0.30
Π-model	45.19±0.21	57.29±0.26	64.18±0.29
Mean Teacher	54.28±0.14	66.02±0.21	74.24±0.23
UDA	39.90±0.43	64.16±0.40	71.86±0.56
FixMatch	49.86±0.27	77.54±0.29	84.78±0.33
SimCLRv2	45.74±0.16	61.70±0.18	77.49±0.24
Self-Tuning	74.99±0.11	85.87±0.04	89.83±0.01
Self-Tuning+HCR	**78.76**±0.08	**87.70**±0.07	**91.14**±0.06

Table 2 shows results on FGVC Aircraft dataset. As we can see, the observations are consistently the same as those for Stanford Cars dataset, which is, HCR is still able to obtain large gains (averagely 3.03%) even Self-Tuning has achieved a very high accuracy. Also, the lower the label proportion, the larger the improvements brought by HCR.

Table 2. Classification accuracy (%) ↑ of semi-supervised learning methods on FGVC Aircraft dataset (ResNet-50 pretrained).

Method		Label Proportion	
	15%	30%	50%
Pseudo-Labeling	46.83±0.30	62.77±0.31	73.21±0.39
Π-model	37.72±0.25	58.49±0.26	65.63±0.36
Mean Teacher	51.59±0.23	71.62±0.29	80.31±0.32
UDA	43.96±0.45	64.17±0.49	67.42±0.53
FixMatch	55.53±0.26	71.35±0.35	78.34±0.43
SimCLRv2	40.78±0.21	59.03±0.29	68.54±0.30
Self-Tuning	66.68±0.17	79.94±0.09	84.35±0.08
Self-Tuning+HCR	**70.54**±0.02	**82.64**±0.04	**86.89**±0.15

We report results on CUB-200-2011 dataset in Table 3. It can be seen that applying HCR yields better performance than the original Self-Tuning.

Note that the improvement on CUB-200-2011 dataset is slightly less than the former two datasets. The reason is that the average number of samples per class of CUB-200-2011 is much less than Stanford Cars and FGVC Aircraft

Table 3. Classification accuracy (%) ↑ of semi-supervised learning methods on CUB-200-2011 dataset (ResNet-50 pretrained).

Method	Label Proportion 15%	 30%	 50%
Pseudo-Labeling	45.33±0.23	56.20±0.29	64.07±0.32
Π-model	45.20±0.25	58.49±0.26	65.63±0.36
Mean Teacher	53.26±0.19	66.66±0.20	74.37±0.30
UDA	46.90±0.31	61.16±0.35	71.86±0.43
FixMatch	44.06±0.23	63.54±0.18	75.96±0.29
SimCLRv2	45.74±0.15	62.70±0.24	71.01±0.34
Self-Tuning	64.79±0.06	74.31±0.07	78.45±0.31
Self-Tuning+HCR	**66.42**±0.24	**75.06**±0.13	**79.48**±0.16

datasets. It's difficult for HCR to capture the structure of data while those data are sparsely distributed on the hypersphere. We analyze the relationship between average samples per class and the improvements as shown in Figure 7.

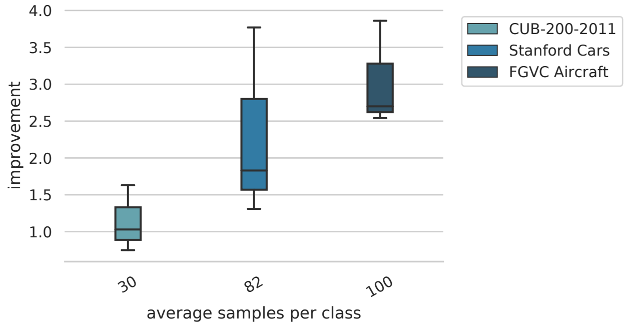


Figure 7. The improvements brought by HCR are proportional to the average number of samples per class.

Except for the visual datasets evaluated in Self-Tuning, we perform experiments on the standard semi-supervised learning benchmark CIFAR-100 dataset. The results are shown in Table 4 and Table 5. Self-Tuning implements EfficientNet-B2 model [72] as the pretrained weights of WRN-28-8 [89] are not available. While FixMatch obtains a higher error rate with EfficientNet-B2 than WRN-28-8, Self-Tuning outperforms those methods on WRN-28-8. HCR here further increases its leading.

Besides, under extremely few labels condition, HCR strongly outperforms other methods through promoting Self-Tuning by 4.46% which is definitely a large margin. We believe our proposed HCR can play a significant role in bridging feature-dependent and label-dependent information, especially in the case of few labels being available.

Table 4. Error rates (%) ↓ of semi-supervised learning methods on CIFAR-100 dataset with 2500 labes, and 10000 labels.

Method	Network	2.5K	10k
Pseudo-Labeling	WRN-28-8 #Para: 11.76M	57.38	36.21
Π-Model		57.25	37.88
Mean Teacher		53.91	35.83
MixMatch		39.94	28.31
UDA		33.13	24.50
ReMixMatch		27.43	23.03
FixMatch		28.64	23.18
FixMatch	EfficientNet-B2 #Para: 9.43M	29.99	21.69
Fine-Tuning		31.69	21.74
Co-Tuning		30.94	22.22
Self-Tuning		24.16	17.57
Self-Tuning+HCR		**23.93**	**16.24**

Table 5. Error rates (%) ↓ of semi-supervised learning methods on CIFAR-100 dataset with only 400 labels (EfficientNet-B2 pretrained). CT: Co-Tuning, PL: Pseduo Labeling, MT: Mean-Teacher, FM: FixMatch.

Fine-Tuning	L^2-SP	DELTA	BSS
60.79	59.21	58.23	58.49
Co-Tuning	Pseudo Labeling	Π-model	Mean Teacher
57.58	59.21	60.50	60.68
FixMatch	UDA	SimCLRv2	CT+PL
57.87	58.32	59.45	56.21
CT+MT	CT+FM	Self-Tuning	Self-Tuning+HCR
56.78	57.94	47.17	**42.71**

5.3. Fine-grained classification

We conduct experiments on fine-grained classification using fully-labeled Stanford Cars, FGVC Aircraft, and CUB-200-2011 datasets. The results in Table 6 show that HCR performs consistently better than the baseline. FGVC Aircraft still gets the most improvements, benefited from its large average number of samples per class as we have mentioned above. We believe HCR can not only help semi-supervised conditions but also difficult supervised learning.

Table 6. Classification accuracy (%) ↑ of transfer learning methods on fine-grained datasets.

Method	Stanford Cars	Aircraft	CUB200
Fine-Tuning	87.20±0.19	81.13±0.21	78.01±0.16
L^2-SP	86.58±0.26	80.98±0.29	78.44±0.17
DELTA	86.32±0.20	80.44±0.20	78.63±0.18
BSS	87.63±0.27	81.48±0.18	78.85±0.31
Co-Tuning	89.53±0.09	83.87±0.09	81.24±0.14
Self-Tuning	92.33±0.10	88.96±0.21	81.60±0.11
Self-Tuning+HCR	**93.03**±0.06	**90.41**±0.03	**82.63**±0.19

5.4. Noisy label learning

We follows the same experimental settings as [71] to compare Co-learning with HCR against other co-training-based noisy label learning methods: Decoupling [57], Co-teaching [22], Co-teaching+ [88], JoCoR [82], and Co-learning itself [71]. We conduct experiments on CIFAR-100 dataset with three different types of noise, i.e., *symmetric*, *asymmetric*, and *instance-dependent*. The details of these noise types are in Appendix. Among these noisy types, we recognize instance-dependent (or feature-dependent) noise as a more realistic setting because human annotations are prone to different levels of errors for tasks with varying difficulty levels. Following [71], we report the average test accuracy over the last 10 epochs of five trials for each experiment. The base model is ResNet-18.

Tabel 7 shows the results on symmetric noise. As it is the simplest synthetic noise type, we perform experiments on high noise ratios as 50% and 80%. While Co-learning has already shown amazing results on high noise ratios, HCR further improves Co-learning by averagely 3.92% under different noise ratios.

Table 7. Average test accuracy (%) on CIFAR-100 with symmetric noise over the last 10 epochs.

Method	sym-20%	sym-50%	sym-80%
Standrad CE	57.79±0.44	33.75±0.46	8.64±0.22
Decoupling	56.18±0.32	31.58±0.54	7.71±0.23
Co-teaching	64.28±0.32	32.62±0.51	6.65±0.71
Co-teaching+	55.40±0.71	26.49±0.45	8.57±1.55
JoCoR	62.29±0.71	30.19±0.60	6.84±0.92
Co-learning	66.58±0.15	55.54±0.43	35.45±0.79
Co-learning+HCR	**70.27**±0.32	**59.93**± 0.25	**39.14**±0.47

We report the results on asymmetric noise in Table 8. HCR significantly improves Co-learning by 3.74% on average. Moreover, HCR presents more stable results than Co-learning as the standard deviations are minor.

Table 8. Average test accuracy (%) on CIFAR-100 with asymmetric noise over the last 10 epochs.

Method	asym-20%	asym-30%	asym-40%
Standrad CE	59.36±0.36	51.06±0.44	42.49±0.23
Decoupling	57.97±0.24	49.86±0.54	41.51±0.67
Co-teaching	59.76±0.53	49.53±0.79	40.62±0.79
Co-teaching+	56.11±0.60	47.12±0.73	38.98±0.54
JoCoR	58.58±0.51	49.04±0.91	39.72±0.76
Co-learning	65.26±0.76	56.97±1.22	47.62±0.79
Co-learning+HCR	**68.85**±0.22	**61.94**±0.17	**50.29**±0.69

When it comes to instance-dependent noise, HCR assists Co-learning in obtaining 1.52% gains averagely as shown in Table 9. As Co-learning has utilized feature-dependent information, HCR's improvements on instance-dependent noise are slightly less than the former two noise types.

We also conduct experiments on CIFAR10 with extremely high noise and show the results in Figure 8. While Co-learning suffers from overfitting on noisy labels, HCR consistently works well and performs great robustness.

Table 9. Average test accuracy (%) on CIFAR-100 with instance-dependent noise over the last 10 epochs.

Method	ins-20%	ins-30%	ins-40%
Standrad CE	55.45±0.54	48.77±0.47	41.30±0.27
Decoupling	52.20±0.48	45.32±0.83	36.33±0.47
Co-teaching	55.16±0.61	45.24±0.37	34.64±1.00
Co-teaching+	50.37±0.85	40.73±0.58	32.15±0.80
JoCoR	54.21±0.34	45.03±0.52	34.08±1.05
Co-learning	69.42±0.42	65.45±0.86	60.40±1.37
Co-learning+HCR	**70.03**±0.31	**66.89**±0.41	**62.91**±0.84

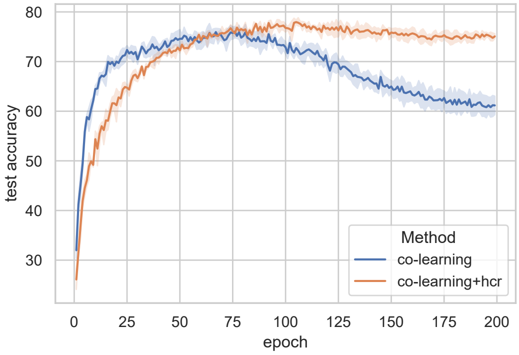


Figure 8. Results on CIFAR-10 with 80% symmetric noise.

6. Conclusion

In this paper, we borrow several theoretical insights from geometry and propose a novel consistency regularization method for semi-supervised and weak-supervised learning, called hyperspherical consistency regularization (HCR), to encourage the pairwise distance distribution of the classifier to be similar to the distribution of the projection head in the latent space. HCR can be conveniently implemented to those jointly learning methods as a plug-in regularization, or be applied to a vanilla supervised learning network with only an additional projection head. Through extensive experiments on semi-supervised learning, fine-grained classification and noisy label learning, HCR shows consistent improvements on these tasks. In general, HCR cast a novel view on leveraging self-supervised learning to assist data-efficient and robust deep learning by introducing hyperspherical consistency.

Acknowledgemnet This work is supported in part by the Science and Technology Innovation 2030 - Major Project (No. 2021ZD0150100) and National Natural Science Foundation of China (No. U21A20427).

References

[1] Vangalur S Alagar. The distribution of the distance between random points. *Journal of Applied Probability*, 13(3):558–566, 1976. 4

[2] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in Neural Information Processing Systems*, 32:15535–15545, 2019. 2

[3] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008. 4

[4] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *International Conference on Learning Representations*, 2019. 1

[5] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[6] Emmanuel J Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, 52(12):5406–5425, 2006. 4

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 1, 2

[8] Beidi Chen, Weiyang Liu, Zhiding Yu, Jan Kautz, Anshumali Shrivastava, Animesh Garg, and Animashree Anandkumar. Angular visual hardness. In *International Conference on Machine Learning*, pages 1637–1648. PMLR, 2020. 3

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 1, 2

[10] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in Neural Information Processing Systems*, 33:22243–22255, 2020. 1, 2, 6

[11] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 2

[12] Hao Cheng, Zhaowei Zhu, Xing Sun, and Yang Liu. Demystifying how self-supervised features improve training from noisy labels. *arXiv preprint arXiv:2110.09022*, 2021. 2

[13] Thomas M Cover and Joy A Thomas. Elements of information theory second edition solutions to problems. *Internet Access*, pages 19–20, 2006. 4, 5

[14] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003. 3, 4

[15] Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak. Hyperspherical variational autoencoders. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pages 856–865. Association For Uncertainty in Artificial Intelligence (AUAI), 2018. 3

[16] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 3

[17] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2051–2060, 2017. 2

[18] Robert J Durrant and Ata Kabán. Random projections for machine learning and data mining: Theory and applications. In *ECML PKDD*, 2012. 4

[19] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. 2

[20] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. 1, 2

[21] John M Hammersley. The distribution of distance in a hypersphere. *The Annals of Mathematical Statistics*, pages 447–452, 1950. 4

[22] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, Montreal, 2018. Curran Associates, Inc. 8

[23] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems*, 33:5679–5690, 2020. 1

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1, 2

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2961–2969, 2017. 1

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1

[27] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020. 1, 2

[28] Qianjiang Hu, Xiao Wang, Wei Hu, and Guo-Jun Qi. Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1074–1083, 2021. 1, 2

[29] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. 1

[30] Mengmeng Jing, Jingjing Li, Lei Zhu, Zhengming Ding, Ke Lu, and Yang Yang. Balanced open set domain adaptation via centroid alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8013–8020, 2021. 3

[31] William B Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space 26. *Contemporary mathematics*, 26, 1984. 4

[32] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2020. 2

[33] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2019. 1

[34] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:1565—-1576, 2020. 1, 2, 4

[35] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004. 4, 5

[36] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 6

[37] A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009. 6

[38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 1

[39] Arno Kuijlaars and E Saff. Asymptotics for minimal discrete energy on the sphere. *Transactions of the American Mathematical Society*, 350(2):523–538, 1998. 4

[40] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017. 6

[41] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 6

[42] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2019. 1

[43] Junnan Li, Caiming Xiong, and Steven Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. *arXiv preprint arXiv:2011.11183*, 2020. 1, 2

[44] Junnan Li, Caiming Xiong, and Steven Hoi. Mopro: Webly supervised learning with momentum prototypes. In *International Conference on Learning Representations*, 2020. 2

[45] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*, 2020. 1, 2

[46] Rongmei Lin, Weiyang Liu, Zhen Liu, Chen Feng, Zhiding Yu, James M Rehg, Li Xiong, and Le Song. Regularizing neural networks via minimizing hyperspherical energy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6917–6927, 2020. 3

[47] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 1

[48] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 1

[49] Weiyang Liu, Rongmei Lin, Zhen Liu, Lixin Liu, Zhiding Yu, Bo Dai, and Le Song. Learning towards minimum hyperspherical energy. *Advances in Neural Information Processing Systems*, 31:6222–6233, 2018. 3

[50] Weiyang Liu, Rongmei Lin, Zhen Liu, James M Rehg, Liam Paull, Li Xiong, Le Song, and Adrian Weller. Orthogonal over-parameterized training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7251–7260, 2021. 3

[51] Weiyang Liu, Rongmei Lin, Zhen Liu, Li Xiong, Bernhard Schölkopf, and Adrian Weller. Learning with hyperspherical uniformity. In *International Conference On Artificial Intelligence and Statistics*, pages 1180–1188. PMLR, 2021. 3

[52] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 212–220, 2017. 2

[53] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *International Conference on Machine Learning*, pages 507–516. PMLR, 2016. 3

[54] Weiyang Liu, Yan-Ming Zhang, Xingguo Li, Zhiding Yu, Bo Dai, Tuo Zhao, and Le Song. Deep hyperspherical learning. *Advances in Neural Information Processing Systems*, 31:3953–3963, 2017. 3

[55] Reginald Douglas Lord. The distribution of distance in a hypersphere. *The Annals of Mathematical Statistics*, 25(4):794–798, 1954. 4

[56] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classi-

fication of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6

[57] Eran Malach and Shai Shalev-Shwartz. Decoupling ”when to update” from ”how to update”. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 960–970, Long Bench, 2017. Curran Associates, Inc. 7

[58] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 2

[59] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[60] Sung Woo Park and Junseok Kwon. Sphere generative adversarial network based on geometric moment matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4292–4301, 2019. 3

[61] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016. 2

[62] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11557–11568, 2021. 1

[63] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. *Advances in Neural Information Processing Systems*, 28:3546–3554, 2015. 1

[64] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016. 1

[65] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7263–7271, 2017. 1

[66] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 1

[67] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[68] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 2

[69] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33, 2020. 1, 6

[70] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in Neural Information Processing Systems*, 33:11839–11852, 2020. 1

[71] Cheng Tan, Jun Xia, Lirong Wu, and Stan Z. Li. Co-learning: Learning from noisy labels with self-supervision. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM ’21, page 1405–1413, New York, NY, USA, 2021. Association for Computing Machinery. 1, 2, 5, 7, 8

[72] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 1, 7

[73] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30, 2017. 1, 6

[74] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pages 10268–10278. PMLR, 2021. 1, 2

[75] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2019. 4, 5

[76] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 6

[77] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017. 2

[78] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 3

[79] Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. Contrastive learning based hybrid networks for long-tailed image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 943–952, 2021. 1, 2

[80] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020. 1, 2, 5

[81] Ximei Wang, Jinghan Gao, Mingsheng Long, and Jianmin Wang. Self-tuning for data-efficient deep learning. In *International Conference on Machine Learning*, pages 10738–10748. PMLR, 2021. 1, 2, 5, 6

[82] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with

co-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13726–13735, Washington, 2020. IEEE. 8

[83] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 2

[84] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. *Advances in Neural Information Processing Systems*, 33, 2020. 1

[85] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33, 2020. 1, 6

[86] Jiacheng Xu and Greg Durrett. Spherical latent spaces for stable variational autoencoders. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4503–4513, 2018. 2

[87] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1426–1435, 2019. 3

[88] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7164–7173, Long Bench, 09–15 Jun 2019. PMLR. 8

[89] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 7

[90] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021. 1, 2

[91] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 2

[92] Jincheng Zhong, Ximei Wang, Zhi Kou, Jianmin Wang, and Mingsheng Long. Bi-tuning of pre-trained representations. *arXiv preprint arXiv:2011.06182*, 2020. 1

[93] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9719–9728, 2020. 1

[94] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Advances in Neural Information Processing Systems*, 33:1–13, 2020. 1, 2