

Angular Gap: Reducing the Uncertainty of Image Difficulty through Model Calibration

Bohua Peng
Imperial College London
London, UK
bohua.peng19@imperial.ac.uk

Mobarakol Islam✉
Imperial College London
London, UK
m.islam20@imperial.ac.uk

Mei Tu
Samsung Research
Beijing, China
tumei@outlook.com

ABSTRACT

Curriculum learning needs example difficulty to proceed from easy to hard. However, the credibility of image difficulty is rarely investigated, which can seriously affect the effectiveness of curricula. In this work, we propose Angular Gap, a measure of difficulty based on the difference in angular distance between feature embeddings and class-weight embeddings built by hyperspherical learning. To ascertain difficulty estimation, we introduce class-wise model calibration, as a post-training technique, to the learnt hyperbolic space. This bridges the gap between probabilistic model calibration and angular distance estimation of hyperspherical learning. We show the superiority of our calibrated Angular Gap over recent difficulty metrics on CIFAR10-H and ImageNetV2. We further propose Angular Gap based curriculum learning for unsupervised domain adaptation that can translate from learning easy samples to mining hard samples. We combine this curriculum with a state-of-the-art self-training method, Cycle Self Training (CST). The proposed Curricular CST learns robust representations and outperforms recent baselines on Office31 and VisDA 2017.

CCS CONCEPTS

• **Computing methodologies** → **Curriculum learning.**

KEYWORDS

Example difficulty, hyperspherical learning, model calibration.

ACM Reference Format:

Bohua Peng, Mobarakol Islam✉, and Mei Tu. 2022. Angular Gap: Reducing the Uncertainty of Image Difficulty through Model Calibration. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*, October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3503161.3548289>

1 INTRODUCTION

Ascertaining example difficulty is a critical problem to curriculum learning and self-paced learning, in that curricula rank training samples by difficulty and proceed from easy to hard. In the context of image classification, a natural idea is to quantify such difficulty with human selection frequency[34], i.e., the fraction of annotators

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '22, October 10–14, 2022, Lisboa, Portugal

© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10...\$15.00
<https://doi.org/10.1145/3503161.3548289>

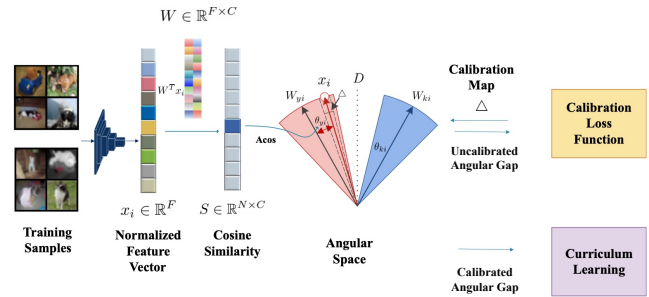


Figure 1: An overview of our Angular Gap image difficulty quantification framework. In the training stage, deep neural networks learn image vectors and class vectors in an angular space with label information, and output raw difficulty scores based on angles. Difficulty scores are then calibrated on a hold-out validation set. In the test stage, the proposed framework can output example difficulty for downstream tasks.

selecting a sample for its target class. However, human labelling effort is not scalable to get fine-grained image difficulty. To measure the difficulty of 10,000 CIFAR10 images, CIFAR10-H[2] recruits 2570 annotators to perform 511,400 trials, with an average of 51 human decisions per image, not including a considerable amount of practice and attention checks. Hence, automating difficulty estimation is crucial to applying curriculum learning to large scale datasets. Probabilistic models are particularly compelling for this automatic estimation demand because of their consistency towards noisy image contents and uncertain labels regularly presented in large scale datasets. Early works have characterized image difficulty with maximum confidence or classification margin, the difference between the predicting probability of the correct class and the largest among others. However, difficulty measurers based on modern neural networks have a reputation of being poorly calibrated. Extensive research have shown that the negative log-likelihood can easily overfit training samples, pushing average predicting probability away from accuracy[13, 21]. This suggests considerable uncertainty of softmax probabilities, and imprecise difficulty measurement undermines the performance of curriculum learning. While probability estimation deteriorates, final classification results actually improve[13]. Very recently, deep ensemble methods [1, 18] measure example difficulty with agreement either from last layers' predictions or from intermediate layers' predictions. Reducing estimation uncertainty with ensembling requires selected treatments and controls such as architectures, number of submodels, and number of

data splits. In this work, we show faithful image difficulty can be efficiently estimated by deep metric learning.

Hyperspherical learning[26], a weakly supervised learning framework, groups instances of the same concept together and pushes instances of different concepts apart by enforcing angular discrimination during training. The framework allows for more robust similarity estimation and has improved representation learning in both computer vision[7] and natural language understanding[8]. Specifically, samples and classes are projected as vectors with constant norms in a hyperbolic space. The normalization operation creates "radial" feature distributions, and the corresponding cosine similarity has been proved to be robust for many downstream tasks[6]. The benevolent properties of hyperspherical similarity estimation give us motivation for difficulty estimation. Angular visual hardness (AVH)[5] initially defines difficulty as the angular distance to its label class normalized by the sum of angular distances to all classes. However, a limitation of this difficulty is that significant angular information can be flushed away by the accumulation of imprecise angular distances. For instance, if an image shows a tabby cat, the distance to its class vector is washed out by distances to unrelated classes, e.g., goldfish or sailboat, resulting in example difficulty with high variance. Based on the assumption that more probable predictions are better calibrated[21], we propose a new difficulty defined as the difference between angular distances of the label class and the smallest of other classes as illustrated in Figure 1. Additionally, we introduce a multi-level calibration method to reduce estimation uncertainty through post-training calibration. In summary, our contributions and findings are summarized below:

- 1.) We propose Angular Gap to measure example difficulty for designing a curriculum learning scheme.
- 2.) We develop multilevel calibration techniques with global and class-wise calibration to produce accurate uncertainty for Angular Gap.
- 3.) We propose a smooth transfer learning curriculum and integrate CST with calibrated Angular Gap for the unsupervised domain adaptation task.
- 4.) We extensively validate calibrated Angular Gap on several SOTA methods and datasets of unsupervised domain adaptation and the results suggest the superior performance.

2 RELATED WORKS

Image difficulty. A wide range of researchs show images possess different amounts of difficulty. It takes tremendous efforts to quantify human perceptual image difficulty. Recently, a line of works model difficulty with the "learning dynamics" of labelling functions. Forgetting events[39] relate example difficulty to catastrophic forgetting [9] by measuring the occurrence of a sample being forgotten during training. The measurement is generalized from discrete domains to continuous domain by averaging the results of ensembles. C-score [18] designs a Monte Carlo method to estimate difficulty w.r.t the probability of correct generalization. Prediction depth [1] employs an ensemble of k-NN classifiers to output intermediate predictions, and defines difficulty as the earliest layer where subsequent intermediate predictions converge. However, difficulty measured by deep ensembles rely on selected treatments and controls such as architectures, data splits and ensembling strategies. Recently, Angular Visual Hardness [5] initially tries to model image difficulty

with angular distance predicted by a single neural network. In this work, we reinforce this idea with hyperspherical learning[26] that emphasizes angular discrimination and ascertain difficulty with model calibration.

Uncertainty estimation. The shared goal of uncertainty estimation and model calibration is to provide trustworthy model confidence for decision making. Expected calibration error (ECE) and reliability diagrams are standard metrics to measure model calibration[33]. Recently, deep ensembles[18, 23] become popular methods for visual uncertainty estimation due to less correlation between individual models. The major drawback is their heavy computational overheads. To ascertain the model confidence of a single model, another line of research focuses on post-training calibration. Platt scaling[33] is a test-by-time parametric approach that rescales output logits with an extra linear layer trained on a hold-out validation set. Temperature Scaling (TS)[13] simplifies this approach with a single learnable parameter. Most recently, variants of Platt scaling[21] and class-wise TS [16] present better calibration performance over vanilla TS. However, Dirichlet Calibration[21] claims that Temperature Scaling mainly focuses on the maximum probability instead of predictions of all classes, which aligns with [30]. In this work, we opt to revisit model calibration and predict plausible similarity in a hyperbolic space.

Curriculum learning. Curriculum learning[3] is a paradigm that favors learning along a curriculum of examples from easy to hard. Starting from this general idea, self-paced learning [22] implements an automatic curriculum that considers examples with small loss as representative examples. With recurrent neural network, MentorNet[17] combines the best of curriculum learning and self-paced learning with a teacher-student architecture that supervises the training of base networks by learning a data-driven curriculum. In the context of supervised learning, deep neural networks learn transferrable features from representative examples before overfitting specific features[18]. We extend the above ideas to hyperspherical learning and propose curricula that prioritize large Angular Gap.

Unsupervised domain adaptation. Unsupervised domain adaptation (UDA) presents a challenging transfer learning problem where data from the source domain are labeled while data from the target domain are unlabeled. A shared assumption between feature alignment methods[11, 45] and self-training algorithms[29, 46] is that shared knowledge exists between domains, which allows for the same labelling function. On the one hand, shared knowledge exists as similar features between domains in the feature adaptation literature[12, 27]. On the other hand, shared knowledge is modelled by the parameters of feature extractors in the self-training algorithms[25, 46, 47]. Using source models to label target data, CBST[46] initially performs pseudo-label selection with class-wise confidence thresholds, which is then improved by confidence regularization as CRST[47]. To handle large domain discrepancy, FixMatch[38] applies a pair of weak and strong data augmentations to target image and enforce consistency regularization when the weakly-augmented image prediction is confident. FixBi[29] uses a fixed mixup ratio to train twin feature extractors as "bridges" between domains. As an alternative, curriculum learning has been applied to domain adaptation from task-level[44] or instance-level[36] using feature adaptation. We borrow these ideas

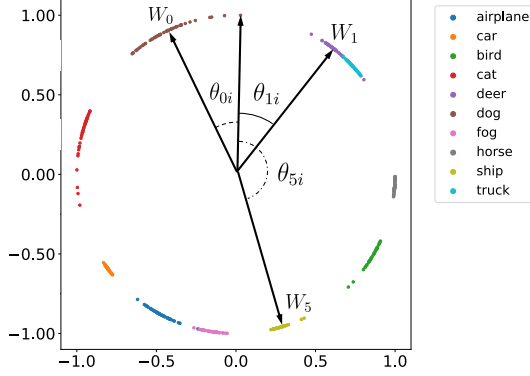


Figure 2: An example of Angular Visual Hardness (AVH). Three out of ten angles are shown for visual clarification.

and propose a transfer learning curriculum to create infinite bridges that gradually decrease discrepancy between different domains. Different from existing works, our transfer learning curriculum does not purely focus on easy samples, but choose the optimal from searching.

3 PRELIMINARY

Angular Visual Hardness. Figure 2 shows an example of image difficulty measured by Angular Visual Hardness (AVH). This metric is automatically estimated by a neural network formed by a feature extractor and a linear classifier. For an individual image, AVH is defined as the angular distance between its feature vector and label weight vector divided by the sum of angular distance between its feature vector and all class weights.

$$AVH(x_i) = \frac{\mathcal{A}(x_i, w_{y_i})}{\sum_{k=1}^C \mathcal{A}(x_i, w_k)} \quad (1)$$

$$\mathcal{A}(x_i, w_k) = \arccos\left(\frac{x_i^T w_k}{\|x_i\| \|w_k\|}\right) \quad (2)$$

where $x_i \in \mathbb{R}^d$ denotes the d dimensional image feature extracted by a backbone. The image is categorized into one of C classes and labeled as y . $w_k \in \mathbb{R}^d$ is the k -th column of the linear classifier's weight W .

Subdomain feature alignment. Deep Subdomain Adaptation Network[45] performs fine-grained feature alignment by dynamically weighing up samples from less representative classes. In their method, local maximum mean discrepancy (MMD) can be measured as follows

$$d_{\mathcal{H}}(P, Q) \triangleq \sum_{k=1}^C u_k \left\| \frac{1}{|X^S|} \sum_{x_s \in S} \phi(x_s) - \frac{1}{|X^T|} \sum_{x_t \in T} \phi(x_t) \right\|_{\mathcal{H}} \quad (3)$$

where u_k is the class ratio to characterize subdomains defined in [45]. x^s and y^s are features and labels from the source domain S , and x^t are unlabeled features from the target domain T . ϕ are kernel functions that measure the distance between source feature x_s and target feature x_t on a Hilbert space. Deep features are optimized with the classification loss and this transfer loss.

Cycle Self-Training. As a state-of-the-art single perspective UDA

method, Cycle Self-training (CST)[25] contains an inner loop and an outer loop. Both loops share the same feature extractor but have their own classifier. With input augmentations and consistency regularization, the inner loop focuses on correctly predicting target samples. The outer loop updates feature representations to reduce the difference between source classifier and target classifier.

4 CALIBRATED ANGULAR GAP

In the context of image classification, we propose Calibrated Angular Gap to estimate example difficulty for curriculum learning. The learnt difficulty metric is based on the angular distance between the feature vectors and the class-weight vectors predicted by hyperspherical learning. In the standard curriculum learning[3], we train the model with easier samples determined by Angular Gap, and then gradually feed harder samples. For domain adaptation, we propose a novel curriculum to work with Angular Gap, which provides a smooth transition between adapting easy samples and hard sample mining. We combine this method with cycle self-training (CST).

4.1 Angular Gap

Example difficulty can be considered as modelling "similarity" between examples and abstract concepts. The abstract concepts can be class labels, prototypes, or even text descriptions. For simplicity, we define a new difficulty, Angular Gap, measured as the difference between the similarity to its label class and the largest similarity of all classes. This definition is based on the assumption that larger cosine similarities are more precisely estimated than smaller ones. For example, when searching with an image of a tabby cat, one can probably get many of its kind and some tiger cats because of their common visual properties.

Definition1 Formally, we represent Angular Gap as

$$\mathcal{D}(x, y) = \text{sim}(x, w_y) - \arg \max_{k \neq y} \text{sim}(x, w_k) \quad (4)$$

$$\text{sim}(x, w_k) = \cos \theta_k = \frac{x_i^T w_k}{\|x\| \|w_k\|}, \quad (5)$$

where θ_k is the angle between x and w_k .

Following common practice of image recognition[7, 41], we emphasize angular discrimination on the hyperbolic space with normalized softmax loss (NSL) and feature norm rescaling represented as

$$L_{NSL} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(s \cdot \cos \theta_{y_i})}{\sum_{k=1}^C \exp(s \cdot \cos \theta_k)} \quad (6)$$

s denotes the scaling factor that rescales feature norms to a constant. Unlike [7] that inserts a geodesic margin between the sample and its class center, here we remove the margin to achieve better generalisation. Note that feature normalization has projected features to a hypersphere with a radius of s .

4.2 Multilevel model calibration

In our empirical study, feature norms increase continuously yet slowly when training with NSL, indicating negative log-likelihood

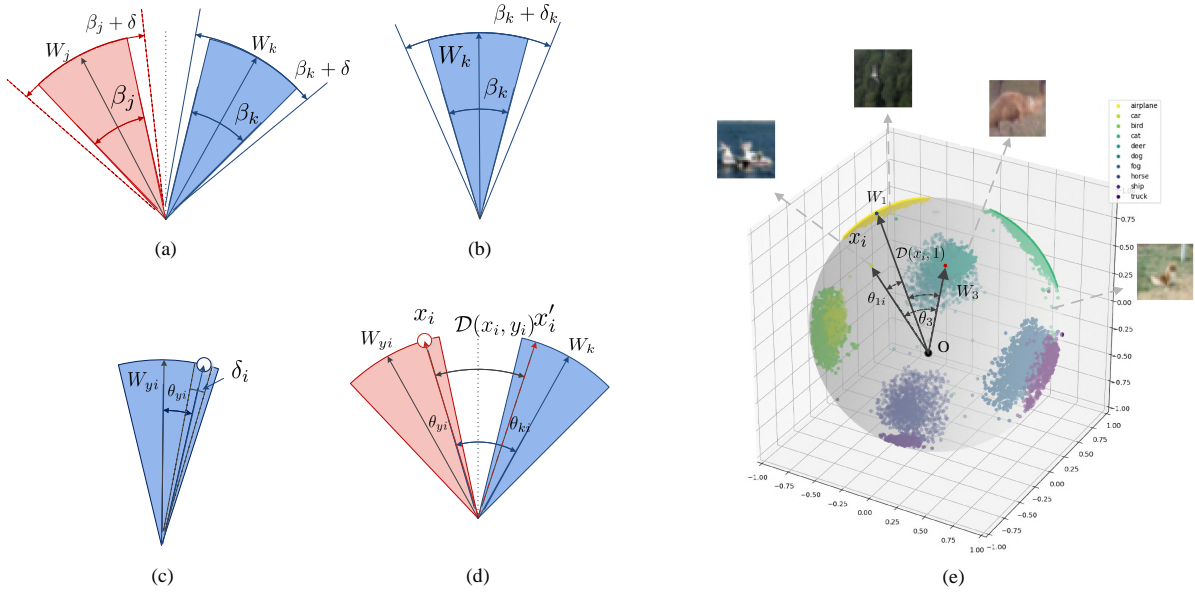


Figure 3: Geometric interpretation of model calibrations on a hyperbolic space. A neural network learns class weights W , features x and angles β and θ during training. These angles are rectified by δ with post-training calibration. (a) Global calibration adds a small angle δ to all samples simultaneously, hence increasing the angular range of class j from β_j to $\beta_j + \delta$. (b) Classwise calibration learns a vector s_d that adds δ_k to the angles of class k . (c) Model calibration rectifies an individual angle θ_{yi} by δ_i . (d) An angular gap is the difference between angle θ_{yi} of the label class and the smallest angle θ_k among other classes. (e) We visualize the image difficulty of CIFAR10 measured by Angular Gap $\mathcal{D}(x_i, y_i)$ on a 3D globe.

overfitting training data. This is partially because the cosine similarity get minimized when feature norms increase. Although overfitting may increase test accuracy, uncertain similarities harm difficulty estimation. As shown in Figure 3, we handle this problem with model calibration from a global level, a class-wise level and an instance level. In general, our idea is to learn multiplicative calibration functions that refine the angles on a hyperbolic space.

$$\cos(\theta + \Delta) = \cos \Delta \cos \theta - \sin \Delta \sin \theta \quad (7)$$

$$\approx \cos \Delta \cos \theta - \Delta \sin \theta \quad (8)$$

$$= \varphi(x, \theta) \cos \theta \quad (9)$$

With small-angle approximation, the nonlinear calibration function $\varphi(x, \theta)$ adds or remove a small angle δ from the original prediction. **Global calibration** Global calibration expands or shrinks all angles on the hyperbolic space simultaneously with a single learnable parameter s_t . This requires a validation dataset with samples $x_j \in \chi$ and labels $y_j \in \mathcal{Y} = \{, \dots, C\}$. The loss function for this calibration method is

$$\min_{s_t} L_{global} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(s \cdot (s_t \cdot \cos \theta_{yi}))}{\sum_{k=1}^C \exp(s \cdot (s_t \cdot \cos \theta_k))}, \quad (10)$$

$$\cos \xi_k = s_t \cdot \cos \theta_k, \quad (11)$$

s_t is an additional parameter learnt during post-training calibration. $\cos \xi_k$ is the refined angular distance at global level.

Class-wise calibration A single parameter is not enough to give us precise example difficulty calibration. To capture class-level difficulty exhibited in behavioral datasets, we let the neural network

learn a vector $s \in \mathbb{R}^C$ that equally rescales angles with another calibration loss function defined as

$$L_{class} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(s \cdot (s_{yi} \cdot \cos \theta_{yi}))}{\sum_{k=1}^C \exp(s \cdot (s_k \cdot \cos \theta_k))}, \quad (12)$$

$$\cos \xi_k = s_k \cdot \cos \theta_k, \quad (13)$$

where s_k is the k th entry of the vector s corresponds to class k . s_{yi} rescales the angular distance to the label class. $\cos \xi_k$ is the calibrated angular distance at class level.

$$\mathcal{D}^*(x, y) = \cos \xi_y - \arg \max_{k \neq y} \cos \xi_k \quad (14)$$

Augmented image difficulty \mathcal{D}^* can be computed by replacing the directly measured similarities $\cos \theta_k$ with refined similarities $\cos \xi_k$, as done with global calibration and class-wise calibration. The proposed calibration methods are natural extensions of Temperature Scaling and Vector Scaling[13] in the hyperspherical setting.

4.3 Curricular Cycle Self-Training

With Angular Gap, we facilitate domain adaption by defining a novel curriculum that prioritizes alignment on easy samples that contain general necessary knowledge, and gradually focus on hard samples that contain specific knowledge. Because alignment can be measured by fixed kernel functions or neural classifiers, we combine Angular Gap with DSAN[45] and CST[25], and propose Curricular DSAN and Curricular CST respectively.

In Figure 4, we design a novel curriculum that control example

Table 1: Accuracy (%) of standard curriculum learning guided by example difficulties on CIFAR10-H. Our methods outperform AVH by 1% and are on par with C-score. Note that C-score was built with a deep ensemble with selected data splits, while we train a single AngularGap model from scratch. Correlations with human selection frequency are measured with Spearman’s rank and Kendall’s Tau, with $p < 0.001$ for all experiments. Calibration is empirically measured with ECE(%). Global calibration, Class-wise calibration and Temperature Scaling calibration is represented by Global, Class-wise and TS respectively.

Methods	Spearman’s rank	Kendall’s Tau	ECE	Top-5 acc.	Top-1 acc.
Maximum Confidence	0.266±0.006	0.148±0.004	11.3±0.2	94.5±0.3	74.8±2.1
Maximum Confidence _{TS}	0.273±0.004	0.145±0.004	9.1±0.2	94.5±0.3	75.3±2.1
Classification Margin[39]	0.279±0.006	0.142±0.004	11.3±0.2	94.6±0.3	75.3±2.0
Classification Margin _{TS}	0.283±0.006	0.242±0.004	9.0±0.2	94.9±0.3	75.7±1.3
MC-Dropout[10]	0.256±0.007	0.176±0.006	9.4±0.4	95.7±0.3	77.0±0.5
AVH[5]	0.368±0.006	0.258±0.004	8.2±0.2	98.2±0.3	81.2±1.0
AVH _{Global}	0.376±0.003	0.263±0.003	<u>7.5±0.2</u>	98.6±0.2	81.3±0.7
AVH _{Class-wise}	0.377±0.003	0.265±0.002	7.4±0.2	98.6±0.2	81.4±0.6
Forgetting Events[39]	0.260±0.003	0.187±0.002	11.5±0.5	98.0±0.4	78.9±1.2
C-score[18]	0.316±0.001	0.243±0.001	9.8±0.3	99.0±0.1	82.4±0.4
Prediction Depth[1]	0.290±0.001	0.183±0.001	9.8±0.3	98.5±0.2	81.2±0.4
Angular Gap (Ours)	0.378±0.003	0.265±0.003	8.2±0.2	98.6±0.2	82.0±0.6
Angular Gap _{Global}	<u>0.382±0.003</u>	<u>0.268±0.002</u>	<u>7.5±0.2</u>	98.8±0.2	82.3±0.4
Angular Gap _{Class-wise}	0.384±0.002	0.269±0.002	7.4±0.2	<u>98.9±0.2</u>	82.4±0.4

weights according to pacing functions and example difficulty with sigmoid functions. We choose sigmoid functions to work with Angular Gap $\mathcal{A}(\mathbf{W}, x_s, y_s)$ because this combination allows for efficiently searching pacing functions λ in a symmetric space. Moreover, this curriculum can smoothly transfer between easy-to-hard and hard sample mining.

$$d_s = \sigma(\lambda \cdot \mathcal{D}(x_s, y_s)), \quad (15)$$

$$\lambda_{(a,b)}(t) = N \frac{1-b}{aT} t + Nb. \quad (16)$$

where d_s is example weight ranged between 0 and 1. a and b denote the parameters of pacing functions λ . t is the current time step and T is the number of total iterations.

Curricular DSAN. We want to learn generalizable features by prioritizing alignment of easy samples with the same class. We augment the feature alignment process with a dynamic discrepancy scoring function defined as

$$d_{\mathcal{H}}(P, Q) \triangleq \sum_{k=1}^C u_k \left\| \frac{d_s}{|X^S|} \sum_{x_s \in S} \phi(x_s) - \frac{1}{|X^T|} \sum_{x_t \in T} \phi(x_t) \right\|_{\mathcal{H}}. \quad (17)$$

Curricular CST. We apply Angular Gap to CST to improve the optimization of the outer loop. Our curriculum prioritizes model updates for transferring and scoring the features of easy samples, yielding more robust representations for pseudo-labels generation of the inner loop. To this end, we add example weights to the reverse step as follows,

$$L_{rev} \triangleq \sum_{s \in S} \sum_{t \in T} d_s \| \text{sim}(x_s, x_t) \hat{y}_t - y_s \|^2. \quad (18)$$

This term reduces cross-domain discrepancy by explicitly transforming pseudo labels \hat{y} to the source domain, and updates deep representations learnt by the mutual feature extractor.

5 EXPERIMENTS AND RESULTS

We have designed the experiments to evaluate our proposed methods by evaluating the following hypotheses:

- More credible example difficulty: We evaluate the credibility of learnt Angular Gap by analyzing their correlations with human selection frequency on CIFAR10-H and ImageNetV2.
- Better curriculum learning: We evaluate the performance of networks when guided with Angular Gap under the standard curriculum learning framework [3].
- More robust representations: We evaluate how the network is able to learn better generalizable representations with the proposed curriculum on the task of unsupervised domain adaptation.

5.1 Datasets

CIFAR10[19] and ISLRCV 2012 (ImageNet) [20] are standard benchmarks for image classification. For human evaluation, CIFAR10-H[2] and ImageNetV2[34] are two recently popular behavioral datasets that report human selection frequency. Human selection frequency models instance-level difficulty with the fraction of people that correctly classify an image. CIFAR10-H is composed of 10,000 images from 10 classes with 511,400 decisions given by 2570 annotators. ImageNetV2 is composed of 10,000 large-scale images from 1000 classes, where each image is labelled by more than 10 annotators.

For domain adaptation, we consider Office-31[35] and VisDA 2017[32] as standard benchmarks. Office-31 consists of images of 31 classes from three domains - Amazon (A), DSLR (D) and Webcam (W). Each domain has 2, 817, 498 and 795 images respectively. We compare our Curricular DSAN and Curricular CST with recent baselines across all six transfer learning tasks. VisDA 2017 considers 152, 409

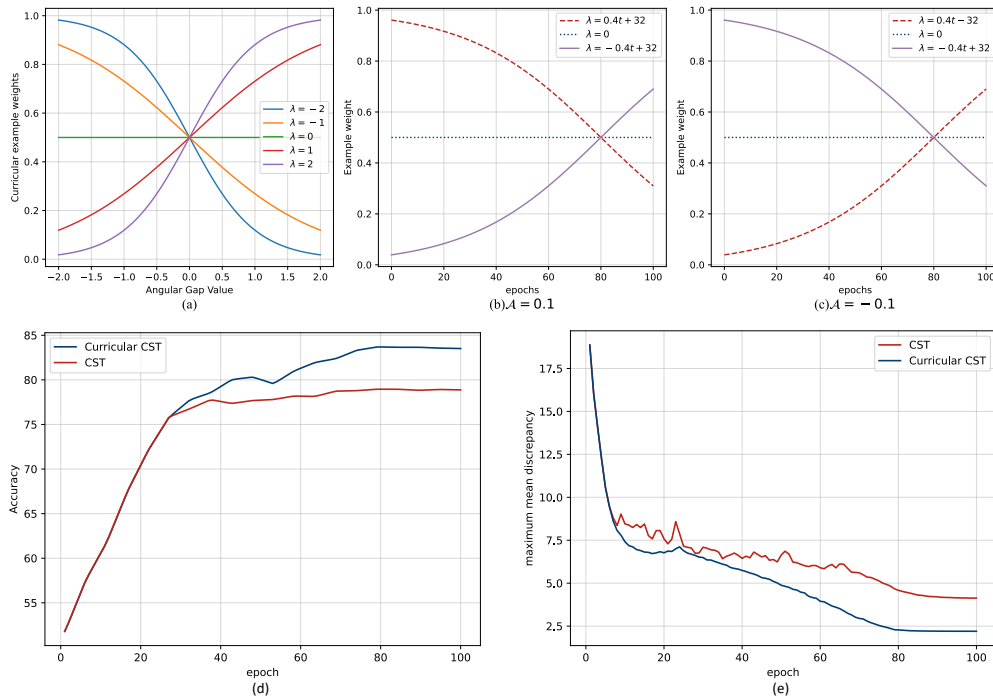


Figure 4: Curricular cycle self-training. The pacing function λ schedules example weights as training proceeds. The optimal transfer learning curriculum, shown as the purple line, decreases the importance of an easy sample (b) and increases the importance of a hard sample (c) as training proceeds. Note that the purple curriculum enforces easy-to-hard at the beginning and hard sample mining at the end. (d) and (e) compare model accuracy and the maximum mean discrepancy between CST (red) and Curricular CST (blue) for all target data points.

Table 2: Correlations between difficulty and HSF on ImageNetV2.

	Spearman's rank		Kendall's Tau	
	ρ	p-value	τ	p-value
Maximum Confidence	0.273	<0.001	0.201	<0.001
Classification margin	0.275	<0.001	0.204	<0.001
Forgetting Events[39]	0.260	0.048	0.187	0.054
Prediction Depth[1]	0.308	<0.001	0.192	<0.001
Classification margin (TS)	0.293	<0.001	0.242	<0.001
AVH (Global)[5]	0.377	<0.001	0.257	<0.001
Angular Gap (Global)	<u>0.379</u>	<0.001	<u>0.269</u>	<0.001
Angular Gap (Class)	0.382	<0.001	0.271	<0.001

labeled synthetic images as the source domain and 55,400 unlabeled real-world images as the target domain.

5.2 Implementation details

Image difficulty estimation. To measure image difficulty, we employ popular convolutional neural networks. We emphasize Angular discrimination with normalized softmax loss and a rescaling factor s of 30. For data preprocessing, we follow PyTorch examples[31] to generate random image crops. We train the models from scratch on CIFAR datasets for 100 epochs with a batch size of 128. We set

the initial learning rate as 0.1 and a cosine learning rate annealing strategy. On ImageNet, we finetune pretrained ResNet50 with SGD optimization and set hyperparameters as stated in PyTorch examples[31].

Multilevel calibration. For calibration, we follow recent papers [13][30] and set the initial weights of vector scaling as an identity matrix. We randomly select 10% of training samples of CIFAR10 for post-training calibration. For ImageNet, we use the validation datasets provided. We optimize the loss functions with LBFGS optimization for 10 epochs with the learning rate set as 0.01. For all experiments, except ImageNet, we report the mean correlation coefficient over 5 different seeds. For ImageNet, we report mean experimental results over 3 different seeds.

Curriculum learning evaluation. Following the fixed easy-to-hard data order as stated in [3], we use the paced learning (PL) setup to fairly compare Angular Gap with other example difficulty measurements. During data loading, we add a fraction of harder examples at the tail of our training data sequence after the current data loader is consumed. Following [42], we employ ResNet18, linear pacing functions and curriculum learning search grids. Note that, for CIFAR10-H, we use the standard image augmentation and weight decay regularization as stated in [31] instead of AugMix[15] to ensure consistency. To amplify the effects of image difficulty, we apply cosine learning rate annealing to SGD optimization. As shown in Figure A.7, we have evaluated image difficulty metrics

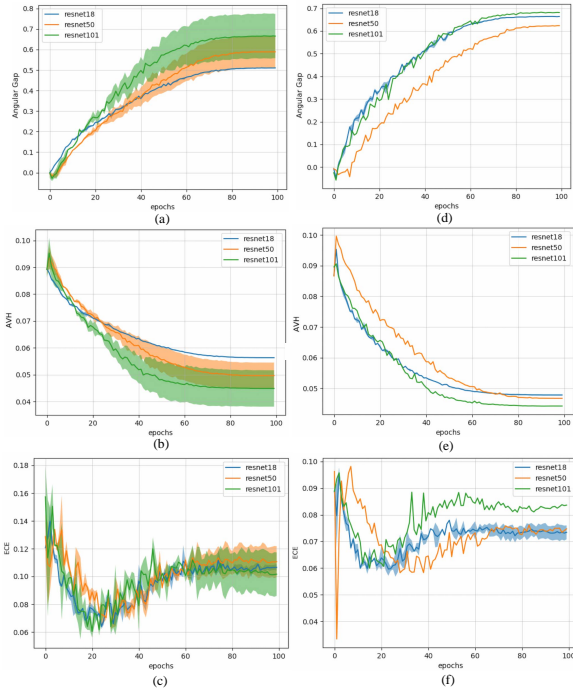


Figure 5: Comparison between Angular Gap and AVH before and after class-wise calibration on CIFAR10-H. Shadows in the plots denote the corresponding standard deviations. The first column shows the training dynamics of uncalibrated Angular Gap (a) and AVH (b) and calibration (c) of the hyperspherical model. The second column shows class-wise calibrated Angular Gap (d) and AVH (e) and calibration (f). Note that the magnitude of the average Angular Gap is more observable than the average AVH across different models. Class-wise model calibration reduces uncertainty in image difficulty estimation.

on standard curriculum learning with Weights and Biases[4] and Google Cloud Platform. The results are summarized in Table 1.

5.3 Results

The results for ImageNetV2 are listed in Table 2 and the results for CIFAR10-H are tabulated as the first columns in Table 1. Figure 5 shows that model calibration has reduced the uncertainty of image difficulty. Our experiments show that simply scaling up models has caused overfitting, and the corresponding image difficulty becomes more uncertain. In Figure 3, we have projected the latent features to a unit globe. Easy samples are closer to their class centers and have larger Angular Gap, while difficult samples with negative Angular Gap can be misclassified due to low-resolved ambiguous content. The 3d visualization reveals more complex angular information, i.e., ambiguous samples related to multiple classes.

We analyze difficulty uncertainty and model capacity using ResNet18, ResNet50 and ResNet101[14]. In general, Figure 5 shows that larger CNNs are more poorly calibrated and have shown more uncertainty

in image difficulty estimation. Scaling up models will cause overfitting, and the corresponding image difficulty will be less plausible. We also analyze the effects of feature normalization by comparing the training dynamics of AlexNet [20], VGG16 [37] and ResNet50 in the supplementary material as shown Figure A.1 and A.2. The improvements on difficulty estimation may probably come from lower feature norms. By emphasizing angular discrimination, feature normalization and rescaling improve model calibration as shown by the reliability diagrams in Figure A.5. Class-wise calibration further improves the class-wise similarity estimation which is the model confidence in the hyperspherical learning setting.

Regarding correlation with human predictions, measured by Spearman’s rank and Kendall’s Tau, calibrated Angular Gap significantly outperforms other baselines except C-score which is an intensively computational ensembling method. We conclude there are two main reasons. Firstly, the Angular Gap regularizes its difficulty estimation with hyperspherical learning. Secondly, the Angular Gap avoids the hazard of uncertain angular distance by using the largest similarity among other classes, which is powerful when the data point is near the boundary of two classes. Interestingly, Figure A.3 shows that class-wise calibration is similar to human judgements to the classes of CIFAR10-H. Another noteworthy observation is that other difficulty measures show improvement after calibration, suggesting that the standard curriculum learning benefits from better example difficulty estimation. Our results also align with experiments of [21] that show multiclass probabilities can be improved by fine-grained calibration.

5.4 Domain Adaptation

We investigate Angular Gap on the domain adaptation tasks for further insights. Following standard protocols of UDA, we use ResNet50 as the backbone for image classification tasks on Office31 and ResNet101 for image classification tasks on VisDA2017. For all methods mentioned above, we project latent features extracted by the backbones to 256d embeddings for discrepancy estimation. For Office31, we use mini-batch stochastic gradient descent (SGD) with an initial learning rate of 0.001, a momentum of 0.9, a batch size of 64, and a weight decay of $5e-4$. The pacing function λ linearly decreases from 4 to -2 for 100 epochs. For VisDA2017, we set the initial learning rate as 0.0001, a momentum of 0.9, a batchsize of 64, and a weight decay of $5e-4$. The pacing function λ linearly decreases from 32 to -8 for 100 epochs. For difficulty estimation, we use Adam to finetune the Angular Gap with 80 percent of source data and perform vector scaling calibration with LBFGS optimization on the rest of source data. The search space is symmetric with a and b chosen from $\{-32, -16, \dots, -2, -1, 2, \dots, 16, 32\}$. We speed up searching with random search and Hyperband[24] algorithms.

Curricular cycle self-trainig. Table 3 shows the classification accuracy of our curricular UDA methods on Office31. Table 4 shows results on *VisDA-2017*. Curricular CST significantly outperforms state-of-the-art baselines with observable margins, indicating the benefits of the proposed curriculum on domain adaptation tasks. Curricular DSAN has also surpassed several feature alignment methods presented in the left part of the Table 4. The optimal curriculum reported by grid search suggests aligning easy samples in the first stage and then focusing on hard samples in the second stage. Note

Table 3: Accuracy (%) on Office31. Our Curricular CST method outperforms baselines on six domain adaptation tasks. $A \rightarrow W$ denotes the transformation task from the Amazon domain to Webcam domain, and $D \rightarrow W$ denotes the transformation from the Webcam domain to DSLR domain. The best accuracy is indicated in bold, and the second best is underlined.

Method	$A \rightarrow W$	$D \rightarrow W$	$W \rightarrow D$	$A \rightarrow D$	$D \rightarrow A$	$W \rightarrow A$	Avg
ResNet	68.44±0.4	96.7±0.1	99.3±0.1	68.9±0.2	62.5±0.3	60.7±0.3	76.1
DANN[11]	84.5±0.4	96.35±0.2	99.5±0.1	80.93±0.5	69.03±0.4	68.98±0.5	83.2
CBST[46]	87.8±0.8	98.5±0.1	100.0±0.0	86.5±1.0	71.2±0.4	70.9±0.7	85.8
MSTN[43]	91.3±0.2	98.9±0.1	100.0±0.0	90.4±0.3	72.7±0.3	65.6±0.5	86.5
CRST[47]	89.4±0.7	98.9±0.4	100.0±0.0	88.7±0.8	72.6±0.7	70.9±0.5	86.8
DSAN[45]	93.0±0.4	97.8±0.2	100.0±0.0	89.3±0.7	73.5±0.5	74.3±0.4	88.0
CST[25]	95.6±0.3	98.4±0.2	100.0±0.0	95.1±0.3	77.8±0.7	78.9±0.2	91.0
FixBi[29]	96.1±0.2	<u>99.3±0.2</u>	100.0±0.0	95.0±0.4	<u>78.7±0.5</u>	<u>79.4±0.3</u>	<u>91.4</u>
Curricular DSAN	93.8±0.2	98.3±0.1	100.0±0.0	90.3±0.5	74.0±0.3	75.2±0.4	88.6
Curricular CST	<u>96.0±0.1</u>	99.5±0.2	100.0±0.0	<u>94.9±0.2</u>	78.9±0.5	80.4±0.1	91.6

Table 4: Accuracy (%) for sythetic-to-real on VisDA2017

Method	Acc.	Method	Acc.
DANN[11]	55.3	CBST[46]	76.4
DAN[27]	61.1	CRST[47]	78.1
MSTN[43]	65.0	FixMatch[38]	76.7
JAN[28]	65.7	CST[25]	79.9
DSAN[45]	74.8	FixBi[29]	<u>87.2</u>
Curricular DSAN	75.4	Curricular CST	88.1

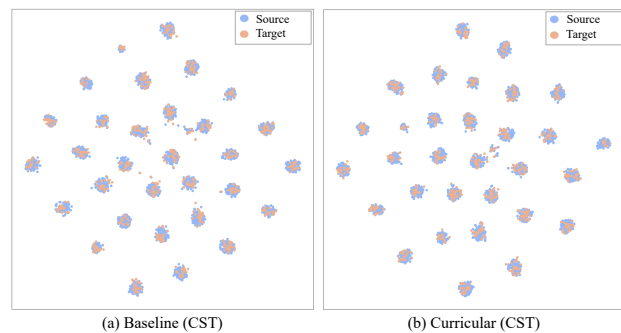


Figure 6: Visualization of features on $A(\text{source}) \rightarrow W(\text{target})$.

that Curricular CST does not need manually tuning confidence threshold for pseudo-label generation as done with CBST, CRST and FixBi.

Curricular domain discrepancy. Figure 4 shows the dynamics of accuracy (%) and discrepancy measured by MMD. Curricular CST is able to achieve better training dynamics and final accuracy. There is a noticeable fluctuation between 20 to 40 epoch, although sigmoid-shape curriculum provides "smooth" transitions. This indicates the model is able to transit from aligning easy samples to hard sample mining. This aligns with the finding that mining more "informative" samples close to the boundary contributes to better classification results. CST methods have larger MMD than DSAN, but better final performance. We claim that this occurs because neural classifiers can measure discrepancy as logits which is an nonlinear measurement.

Feature visualization. On the task $A \rightarrow W$, we visualize the image embeddings with t-SNE[40] in Figure 6. For CST, both the source and target domain features have formed clusters, but many target samples fall out from their class clusters as outliers. For Curricular CST, although the alignment between source cluster centers and target cluster centers is weaker than CST, target embeddings have successfully formed more compact clusters. As a result, there are less outliers than the baseline.

Limitations. The proposed methods have potential limitations.

Measuring difficulty with Angular Gap inevitably generates computational overheads before curriculum learning. Using hyperspherical learning, Angular Gap based curriculum learning improves model generalization with additional complexity according to [7].

6 CONCLUSIONS

In this paper, we propose Angular Gap to address the uncertain image difficulty estimation on a hyperbolic space. We further propose multilevel calibration methods to improve the credibility of the learnt angular metric and look at the calibration problem from hyperspherical learning with geometric interpretations. A curricular cycle self-training method is boosted by the Angular Gap and a curriculum that provides smooth transitions between aligning easy sample and hard sample mining. In our experiments, we show that calibrated Angular Gap is highly correlated with human judgments. On the standard curriculum learning task, the results of Angular Gap are comparable with deep ensembles. We observe the uncertainty of difficulty estimation reduces after calibration. The proposed curriculum results in better optimization of feature discrepancy and significantly improves baselines on the domain

adaptation task. Future work can generalize this framework for standard classification tasks and delve into validating model robustness on synthetic data shift with corruption and perturbation.

REFERENCES

- [1] Robert Baldock, Hartmut Maennel, and Behnam Neyshabur. 2021. Deep learning through the lens of example difficulty. *Advances in Neural Information Processing Systems* 34 (2021).
- [2] R. Battleday, Joshua C. Peterson, and T. Griffiths. 2020. Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature Communications* 11 (2020).
- [3] Yoshua Bengio, J. Louradour, Ronan Collobert, and J. Weston. 2009. Curriculum learning. In *ICML '09*.
- [4] Lukas Biewald. 2020. Experiment Tracking with Weights and Biases. <https://www.wandb.com/> Software available from wandb.com.
- [5] Beidi Chen, Weiyang Liu, Zhiding Yu, Jan Kautz, Anshumali Shrivastava, Animesh Garg, and Animashree Anandkumar. 2020. Angular visual hardness. In *International Conference on Machine Learning*. PMLR, 1637–1648.
- [6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.
- [7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4690–4699.
- [8] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852* (2020).
- [9] Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences* 3, 4 (1999), 128–135.
- [10] Yarín Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*. PMLR, 1050–1059.
- [11] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*. PMLR, 1180–1189.
- [12] Muhammad Ghifary, W Bastiaan Kleijn, and Mengjie Zhang. 2014. Domain adaptive neural networks for object recognition. In *Pacific Rim international conference on artificial intelligence*. Springer, 898–904.
- [13] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*. PMLR, 1321–1330.
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [15] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. 2019. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781* (2019).
- [16] Mobarakol Islam, Lalithkumar Seenivasan, Hongliang Ren, and Ben Glocker. 2021. Class-distribution-aware calibration for long-tailed visual recognition. *arXiv preprint arXiv:2109.05263* (2021).
- [17] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. Mentor-net: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*. PMLR, 2304–2313.
- [18] Ziheng Jiang, Chiyuan Zhang, Kunal Talwar, and Michael C Mozer. 2020. Characterizing structural regularities of labeled data in overparameterized models. *arXiv preprint arXiv:2002.03206* (2020).
- [19] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).
- [20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012).
- [21] Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems* 32 (2019).
- [22] M Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-paced learning for latent variable models. *Advances in neural information processing systems* 23 (2010).
- [23] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems* 30 (2017).
- [24] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. 2017. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research* 18, 1 (2017), 6765–6816.
- [25] Hong Liu, Jianmin Wang, and Mingsheng Long. 2021. Cycle self-training for domain adaptation. *Advances in Neural Information Processing Systems* 34 (2021).
- [26] Weiyang Liu, Yan-Ming Zhang, Xingguo Li, Zhiding Yu, Bo Dai, Tuo Zhao, and Le Song. 2017. Deep hyperspherical learning. *Advances in neural information processing systems* 30 (2017).
- [27] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*. PMLR, 97–105.
- [28] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. 2017. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*. PMLR, 2208–2217.
- [29] Jaemin Na, Heechul Jung, Hyung Jin Chang, and Wonjun Hwang. 2021. Fixbi: Bridging domain spaces for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1094–1103.
- [30] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring Calibration in Deep Learning. In *CVPR Workshops*, Vol. 2.
- [31] Adam Paszke and Gross. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems* 32. Curran Associates, Inc., 8024–8035. <https://github.com/pytorch/examples/blob/master/imagenet/main.py>
- [32] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. 2017. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924* (2017).
- [33] John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10, 3 (1999), 61–74.
- [34] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet?. In *International Conference on Machine Learning*. PMLR, 5389–5400.
- [35] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. 2010. Adapting visual category models to new domains. In *European conference on computer vision*. Springer, 213–226.
- [36] Yang Shu, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. 2019. Transferable curriculum for weakly-supervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 4951–4958.
- [37] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [38] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems* 33 (2020), 596–608.
- [39] Mariya Toneva, Alessandro Sordani, Rémi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. 2019. An Empirical Study of Example Forgetting during Deep Neural Network Learning. *ArXiv abs/1812.05159* (2019).
- [40] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
- [41] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5265–5274.
- [42] Xiaoxia Wu, Ethan Dyer, and Behnam Neyshabur. 2020. When Do Curricula Work? *arXiv preprint arXiv:2012.03107* (2020).
- [43] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. 2018. Learning semantic representations for unsupervised domain adaptation. In *International conference on machine learning*. PMLR, 5423–5432.
- [44] Yang Zhang, Philip David, and Boqing Gong. 2017. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE international conference on computer vision*. 2020–2030.
- [45] Yongchun Zhu, Fuzhen Zhuang, Jindong Wang, Guolin Ke, Jingwu Chen, Jiang Bian, Hui Xiong, and Qing He. 2020. Deep subdomain adaptation network for image classification. *IEEE transactions on neural networks and learning systems* 32, 4 (2020), 1713–1722.
- [46] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*. 289–305.
- [47] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. 2019. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5982–5991.

Angular Gap: reducing the uncertainty of image difficulty through model calibration

Supplementary material

A CALIBRATED ANGULAR GAP

In this section, we give additional information for Angular Gap and multilevel calibration. First, we clarify the difference between confidence and Angular Gap. Then we show some examples of feature norm plots, class-wise calibration maps and reliability diagrams that can help to understand calibrated Angular Gap.

A.1 Model confidence and Angular Gap

In the supervised multi-class classification with neural networks, model confidence refers to the outputs from the softmax layer. Given a class prediction \hat{y}_i of softmax probabilities $y_i \in \mathbb{R}^C$, the confidence can often be computed as follows

$$\mathbb{P}(\hat{y}_i | \mathbf{x}_i, \mathbf{W}, \mathbf{b}) = \frac{\exp(\mathbf{w}_{\hat{y}_i}^T \mathbf{x}_i + b_{\hat{y}_i})}{\sum_{k=1}^C \exp(\mathbf{w}_k^T \mathbf{x}_i + b_k)}. \quad (\text{A.1})$$

By contrast, Angular Gap, defined in Equation 4, uses the cosine similarities between feature vectors and class vectors before the softmax operation. Although the similarity can be considered as a nonlinear version of confidence, we focus on ascertaining the values of these similarities for individual samples during hyperspherical learning. These generalized similarities are enforced by feature normalization and removing bias from the classification layer. To connect Angular Gap with confidence, we write the normalized softmax loss as

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{x}_i} L_{NSL} &= -\frac{1}{N} \sum_{i=1}^N \log \mathbb{P}(y_i | \mathbf{x}_i, \mathbf{W}, s), \\ &= -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(s \cdot (\mathbf{w}_{y_i}^T \mathbf{x}_i / \|\mathbf{w}_{y_i}\| \|\mathbf{x}_i\|))}{\sum_{k=1}^C \exp(s \cdot (\mathbf{w}_k^T \mathbf{x}_i / \|\mathbf{w}_k\| \|\mathbf{x}_i\|))}. \end{aligned} \quad (\text{A.2})$$

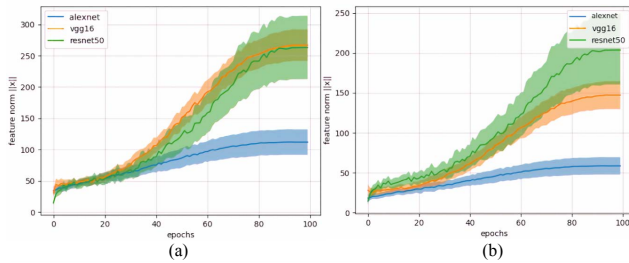


Figure A.1: Feature norm $\|\mathbf{x}\|$ on CIFAR10-H with shadows represent the standard deviation. (a) Training with cross-entropy loss. (b) Training with normalized softmax loss (NSL). Uncertainty increases as model capacity increases.

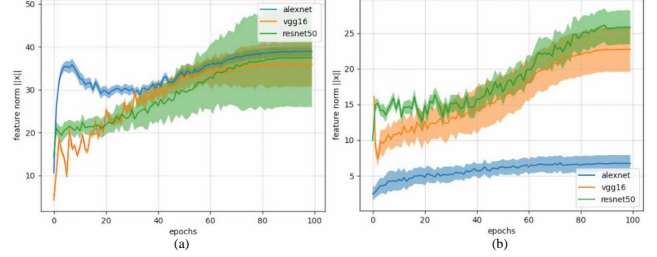


Figure A.2: Feature norm $\|\mathbf{x}\|$ on ILVRC 2012. (a) Training with cross-entropy loss. (b) Training with NSL.

A.2 Feature norm

Figure A.1 and Figure A.2 show the dynamics of feature norm trained on CIFAR10 with cross-entropy loss and normalized softmax loss (NSL) respectively. For CIFAR10-H, the mean and standard deviation are reported over five random seeds, while for ImageNetV2 the statistics are computed over three random seeds. The feature norms slowly diverge as the negative log likelihood minimizes. This indicates the necessity of model calibration when using the probabilities output by a single neural classifier. Although training with NSL makes feature norm smaller, the magnitude of uncertainty cannot be ignored. This can add a double-edged effect on the learnt representations. On the one hand, representations may become more robust because small perturbations to \mathbf{x} cannot easily affect classification results. On the other hand, Angular Gap is more likely to fall into local optimum because it takes more effort to update cosine similarities when feature norms are large. Therefore, in this work we propose global calibration and class-wise calibration to directly refine cosine similarities in the post-training.

A.3 Calibration map

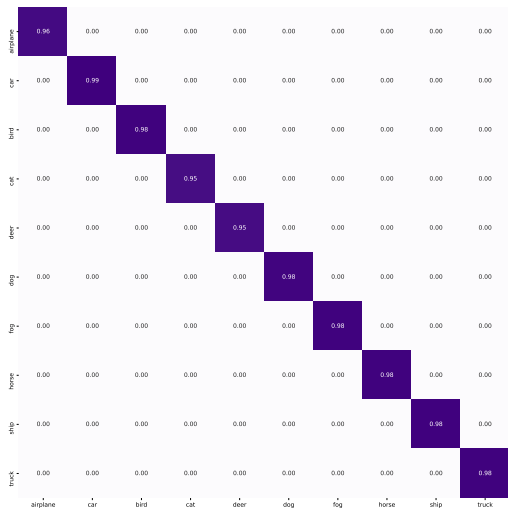
Figure A.3 (a) shows the confusion matrix of human classifiers reported by CIFAR10-H. The class level image difficulty can be represented by the precision of each class. Figure A.3 (b) shows that class-wise calibration is able to capture class level difficulty in which the diagonal entries are learnt during the post-training and others are forced to be zeros.

A.4 Reliability diagrams

As visual representations of model calibration, reliability diagrams plot sample accuracy as a function of confidence. Figure A.5 shows reliability diagrams on the original difficulty estimators (a), and results of applying temperature scaling (b), feature normalization (c), global calibration (d) and class-wise calibration (e). These diagrams also explain the uncertainty of Angular Gap in the sense that Angular Gap can be nonlinearly mapped as confidence in hyperspherical learning. The first column shows the calibration results of the most probable softmax predictions, while the remaining columns show



(a)



(b)

Figure A.3: Comparison of human predicted confusion matrix (a) and a class-wise calibration map (b).

the calibration results of class-wise softmax predictions. The red bars visualize the gaps between expected accuracy with sample accuracy less than expected, the original model shows overconfidence to its predictions. After temperature scaling, the max output probabilities are calibrated but the gaps of the class-wise reliability diagrams are still obvious, which indicates considerable uncertainty exists in the sample confidence of less possible classes. Compared with the previous models trained with cross-entropy shown, models (c), (d) and (e) show less over-confidence, indicating feature normalization is able to reduce the uncertainty of example difficulty of CIFAR10-H samples. With a single learnable parameter, global calibration (d) is not enough to give well-calibrated confidence. Figure A.5 (e) shows

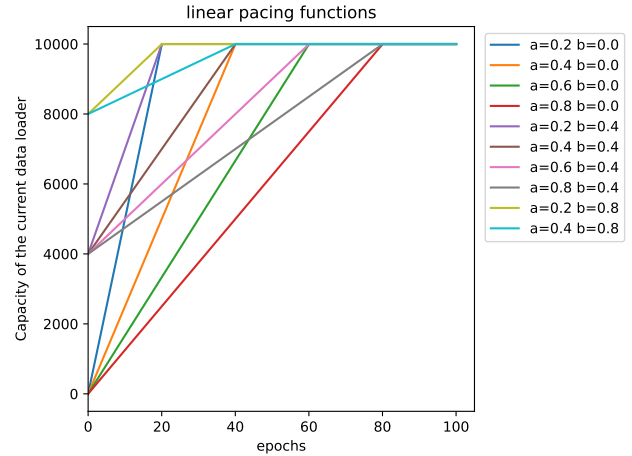


Figure A.4: Linear pacing functions for curriculum learning

that class-wise calibration is able to better calibrate sample confidence on all classes. Figure A.6 (a) shows the calibration results of early stopping applied when the negative log likelihood stagnates. The model has become underconfident. Figure A.6 (b) shows the results of label smoothing (0.1), and class-wise reliability diagrams reveal its instability.

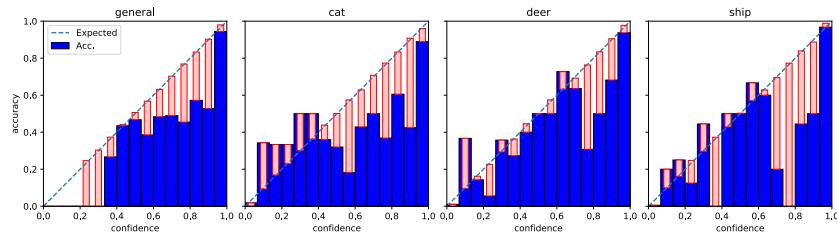
B STANDARD CURRICULUM LEARNING

In this section, we give more details about the standard curriculum learning evaluation that compares image difficulty metrics head-to-head mentioned in Section 5.2. Standard curriculum learning, or Paced learning, designs a simple yet flexible curriculum with precomputed difficulty scores. This curriculum learning scheme contains two main steps. In the first step, data samples are sorted with precomputed difficulty scores in a fixed order. In the second step, a pacing function loads in a scheduled proportion of hard samples every epoch. We opt to use fixed training orders and linear pacing functions to simplify the evaluation. The linear pacing functions can be formally written as,

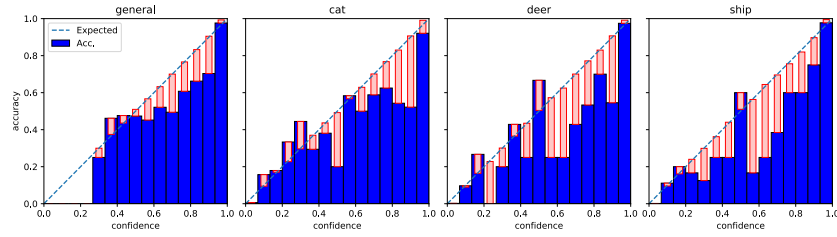
$$g_{(a,b)}(t) = N \frac{1-b}{aT} t + Nb, \quad (A.4)$$

where the pacing functions start with b percentage of training data and gradually add in samples until a percentage of total iterations when the entire training set is fed, before the training continuing to the end. Figure A.4 shows some examples of linear pacing functions for CIFAR10-H. For example, when a is 0.2 and b is 0.8, the linear pacing function corresponds to the olive line on the upper left.

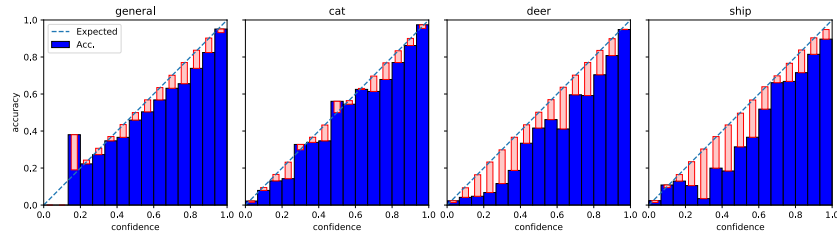
The results for standard curriculum learning evaluation are shown in Figure A.7. Angular Gap outperforms other image difficulty baselines of single-perspective methods, and is on par with the best performing ensemble method, C-score. The error rates of the upper left part of the heat-maps are generally lower than others. This shows that, in order to perform well, the model needs complex enough training materials during the early stage of training.



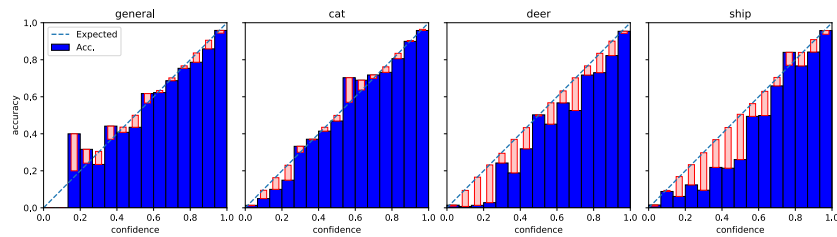
(a) Uncalibrated w/o FN



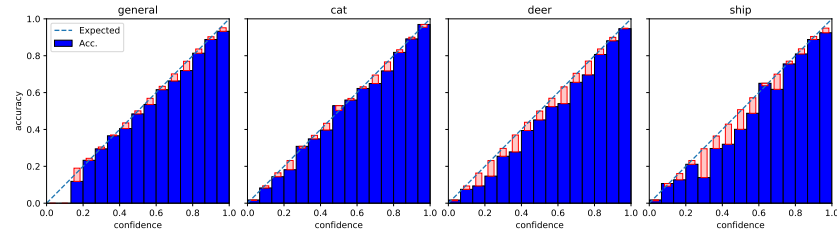
(b) Temperature scaling w/o FN



(c) Uncalibrated with FN



(d) Global calibration with FN



(e) Class-wise calibration with FN

Figure A.5: General reliability diagrams in the first column and class-wise reliability diagrams in the remaining columns. The diagrams visualize the calibration of ResNet18 models by comparing predictive confidence with observed accuracy of CIFAR10-H samples. Red bars indicate the gaps between expected accuracy (dash lines) and observed accuracy (blue bars) of the current confidence bin. (a) and (b) are pre-trained with cross-entropy loss and without feature normalization(FN). (b) uses temperature scaling to calibrate confidence. (c), (d) and (e) are trained with NSL. (d) and (e) applies the proposed global and class-wise calibration during the post-training respectively. Three out of ten classes are shown for visual clarification.

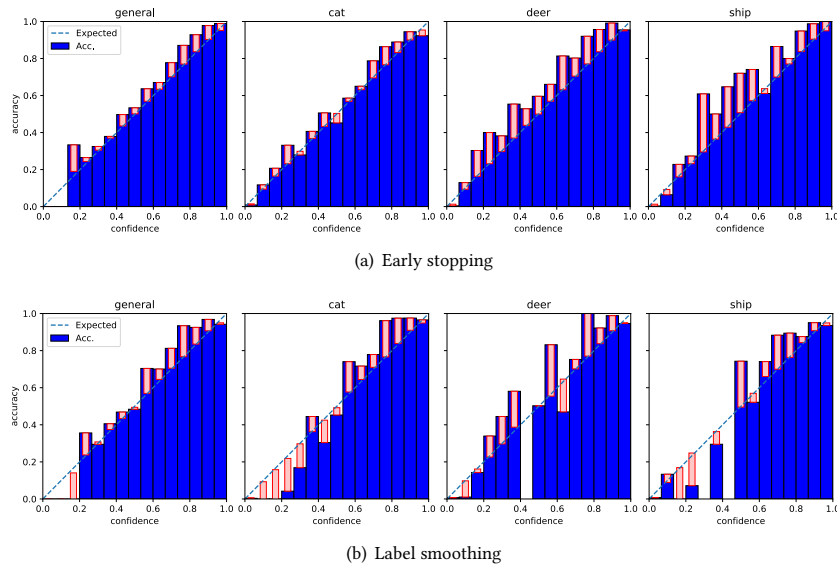


Figure A.6: Additional reliability diagrams of ResNet18 models trained with early stopping (a) and label smoothing (b). The first column shows the calibration results of the most probable softmax predictions, while the remaining columns show the calibration results of class-wise softmax predictions.

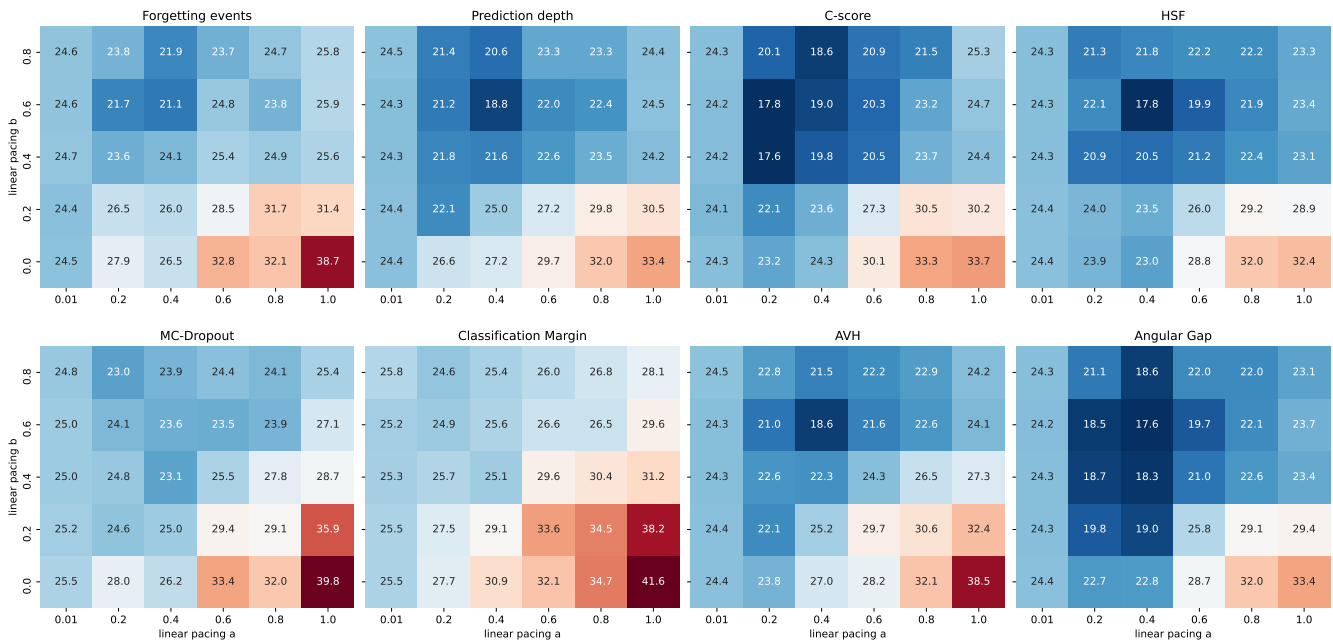


Figure A.7: Test error rate (%) of different difficulty metrics applied to standard curriculum learning on CIFAR10-H. Each difficulty metric corresponds to a heat-map for the parameter $a \in \{0.01, 0.2, 0.4, 0.6, 0.8, 1.0\}$ and the parameter $b \in \{0.0, 0.2, 0.4, 0.6, 0.8\}$. In each heat-map, a cell represents the median accuracy over five runs. Results in the first row show accuracy of image difficulty predicted by ensemble methods: (a) Forgetting events, (b) Prediction depth, (c) C-score, (d) HSF. The second row reports accuracy of image difficulty predicted by single-perspective methods: (e) MC-dropout, (f) Classification margin, (g) AVH, (h) Angular Gap.