




# EntAugment: Entropy-Driven Adaptive Data Augmentation Framework for Image Classification

Suorong Yang<sup>1,2</sup>, Furao Shen<sup>1,3</sup><sup>†</sup>, and Jian Zhao<sup>4</sup>

<sup>1</sup> State Key Laboratory for Novel Software Technology, Nanjing University, China

<sup>2</sup> Department of Computer Science and Technology, Nanjing University, China

<sup>3</sup> School of Artificial Intelligence, Nanjing University, China

<sup>4</sup> School of Electronic Science and Engineering, Nanjing University, China  
sryang@smail.nju.edu.cn, {frshen, jianzhao}@nju.edu.cn

**Abstract.** Data augmentation (DA) has been widely used to improve the generalization of deep neural networks. While existing DA methods have proven effective, they often rely on augmentation operations with random magnitudes to each sample. However, this approach can inadvertently introduce noise, induce distribution shifts, and increase the risk of overfitting. In this paper, we propose EntAugment, a tuning-free and adaptive DA framework. Unlike previous work, EntAugment dynamically assesses and adjusts the augmentation magnitudes for each sample during training, leveraging insights into both the inherent complexities of training samples and the evolving status of deep models. Specifically, in EntAugment, the magnitudes are determined by the information entropy derived from the probability distribution obtained by applying the softmax function to the model’s output. In addition, to further enhance the efficacy of EntAugment, we introduce a novel entropy regularization term, EntLoss, which complements the EntAugment approach. Theoretical analysis further demonstrates that EntLoss, compared to traditional cross-entropy loss, achieves closer alignment between the model distributions and underlying dataset distributions. Moreover, EntAugment and EntLoss can be utilized separately or jointly. We conduct extensive experiments across multiple image classification tasks and network architectures with thorough comparisons of existing DA methods. Importantly, the proposed methods outperform others without introducing any auxiliary models or noticeable extra computational costs, highlighting both effectiveness and efficiency. Code is available at <https://github.com/Jackbrocp/EntAugment>.

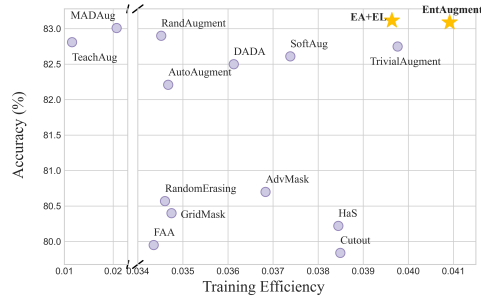
**Keywords:** Data augmentation, image classification, generalization

## 1 Introduction

Data augmentation (DA) has been widely utilized in training deep neural networks to alleviate overfitting and enhance models’ generalization performance [18,

---

<sup>†</sup> Corresponding authors.



**Fig. 1: Training efficiency vs. classification accuracy.** Training efficiency is calculated using the inverse of the average per epoch training time costs, where higher values are preferable. Regarding the efficiency-effectiveness trade-off, our method performs better than prior SOTA DA methods. EA+EL: EntAugment+EntLoss.

22, 29, 33, 44]. However, the prevalent DA frameworks employed for training deep networks commonly adhere to a statically invariant strategy, which is uniformly applied to all training samples throughout the training process [6, 7, 9, 21, 23, 28, 43]. Regardless of specific augmentation operations employed, these methods typically incorporate randomly sampled parameters to adjust the strengths of augmentation. Consequently, the variability in the augmented data is stochastic and non-adaptive to training samples or target model training status. For instance, information deletion-based DA methods [3, 34, 43, 51] adopt deterministic strategies to randomly erase some information in images during training. Moreover, automatic DA approaches [6, 7, 23, 24, 28] search for the augmentation operation space and parameter space prior to the actual training. During on-line training, these methods utilize randomly chosen operations and parameters to generate augmented data, thus non-adaptive to different samples and model training progression. Meanwhile, for different datasets and models, the augmentation parameters have to be manually customized or searched before task model training [6, 7, 9, 43, 50], which hinders practicality.

Despite the efficacy, existing DA methods often neglect the critical aspect of customizing augmentation magnitudes for individual images based on the model training progression, leading to uncontrolled variations in training data. More importantly, the uncontrolled variations in training data can potentially lead to severe drawbacks [42]. Specifically, such variations in the current DA frameworks may inevitably result in excessive or insufficient data manipulation, as evidenced in previous studies [12, 24, 43]. Excessive data manipulation engenders significant variations, introducing noisy samples and distribution shifts. Conversely, insufficient data manipulation may elevate the overfitting risks. Thus, if the augmentation is not aptly modulated, it can deteriorate the overall model performance, which underscores the intrinsic limitations of the existing DA mechanisms.

Given these challenges, it is beneficial to dynamically customize the extent of DA manipulation for each sample. This customization should take into account

both the sample’s inherent difficulty and the models’ current training status (i.e., generalization capacity). For instance, in the early phases of training, when the model’s capacity is limited, and it struggles to classify the majority of samples correctly, augmenting training data with slight magnitudes can expedite performance improvement. As the model progresses and enhances its ability to generalize, a notable portion of samples becomes easier to classify. In this stage, intensifying the augmentation manipulations helps provide augmented data in more diverse scenarios, thereby enabling models to capture more discriminative features. This adaptive DA framework also aligns with the principles of curriculum learning [14, 35, 39, 40] and has demonstrated superior efficacy in the context of model training [3, 16, 43].

In this paper, we propose EntAugment, a novel, tuning-free, and adaptive DA framework. EntAugment incorporates a dynamic adjustment mechanism based on the inherent complexities of individual training samples and the evolving state of models throughout training. However, an accurate and efficient assessment of sample difficulty during online training poses a challenge, primarily due to the limitations in dynamically assessing the model’s training progression based solely on data-derived difficulty measures. To address this challenge, we introduce an entropy-driven data augmentation mechanism that leverages information entropy derived from the probability distribution obtained by applying the softmax function to the model’s output. Although straightforward in concept, the entropy metric derived for each sample dynamically evolves throughout model training. It serves as a robust indicator of the model’s classification confidence for individual samples. More specifically, samples with inherent complexity and those encountered early in training typically exhibit higher entropy values, indicating greater difficulty. As training advances, samples that are initially challenging may exhibit reduced difficulty over time, highlighting dynamic shifts in difficulty measures. Consequently, determining sample difficulty becomes an adaptive and dynamic process shaped by the interplay between the model’s evolving capabilities and the training data’s inherent characteristics. In EntAugment, more challenging samples receive more conservative augmentations to enhance the model’s learning while mitigating the introduction of excessive noise. Conversely, simpler samples benefit from more extensive variations, thereby reducing the risk of overfitting. Meanwhile, it is also noteworthy that EntAugment only incurs minimal computational overhead and obviates the need for any auxiliary models or extra training overhead, highlighting its superior efficiency.

To further enhance the effectiveness of EntAugment, we extend EntAugment by introducing an entropy regularization term, denoted as EntLoss, into the standard cross-entropy (CE) loss. EntLoss is primarily engineered to bolster classification confidence (i.e., minimizing entropy) and enhance the overall performance of EntAugment. Furthermore, EntLoss is underpinned by a rigorous theoretical analysis, wherein we prove that it helps models better fit the underlying dataset distribution (see Proposition 1). Precisely, the dissimilarity induced by the vanilla CE loss serves as an upper bound for the dissimilarity induced by the combined application of CE loss and EntLoss. Therefore, EntLoss brings

additional benefits of enhancing the model’s ability to fit the dataset distribution, thereby improving generalization performance and surpassing the efficacy of existing approaches based on vanilla CE loss. Experiments across a variety of deep models and datasets, e.g., CIFAR-10/100 [19], Tiny-ImageNet [5], ImageNet [20], ImageNet-LT [26], Places-LT [26] demonstrate that our methods bring greater improvement to task models than existing SOTA DA methods in terms of test-set performance while incurring minimal training cost.

Our main contributions can be summarized as follows:

1. We propose EntAugment, a tuning-free and adaptive DA framework, which dynamically adjusts the magnitudes of DA for individual training samples based on sample difficulty and the model training progression.
2. To the best of our knowledge, we are the first to pinpoint that the sample difficulty is not static but varies through model training. This is used to mitigate the side effects caused by the uncontrollable randomness of DA, which is overlooked in prior works.
3. We introduce EntLoss, an entropy regularization loss term for classification model training, accompanied by theoretical foundations that elucidate its role in enhancing model alignment with dataset distribution, thereby contributing to generalization capabilities.
4. EntAugment and EntLoss can be employed both separately and jointly. Extensive experimental results demonstrate the superior effectiveness and efficiency of our methods compared to existing SOTA DA methods.

## 2 Related Work

### 2.1 Data Augmentation

Recently, many data augmentation methods have emerged with the primary objective of enhancing the generalization capabilities of deep neural networks. Image erasing-based data augmentation approaches typically erase some random information within images. For instance, Cutout [9], GridMask [3], HaS [34], and Random Erasing [51] randomly mask out one or more structural regions within images. Since these methods neglect the structural characteristics of training images, AdvMask [43] identifies the critical regions within images offline and selectively drops some structural sub-regions with critical points during online augmentation. However, these methods neglect the training dynamics of deep models. This oversight may inadvertently hamper the optimization potential and efficacy of the learning process, resulting in suboptimal performance.

Image mixing-based data augmentation methods, such as Mixup [47] and CutMix [45], typically mix random information from two or more images to generate the augmented data during training. While effective in generating diverse augmented data, the magnitude of variations introduced by these methods is random, which may inevitably introduce distribution shifts and noises. Meanwhile, the above methods need expert knowledge to design the operations and parameters for specific datasets [21].

Automated data augmentation methods have demonstrated superior performance [6, 21, 24, 28, 36, 48]. These methods typically try to search augmentation policies and parameters automatically based on some metrics before task model training. For instance, AutoAugment [6] employs reinforcement learning to search for the optimal combination of data augmentation policies tailored to individual datasets. Fast-AutoAugment [23], motivated by density matching principles between training and test datasets, introduces an inference-only metric for the evaluation of data augmentation operations. RandAugment [7] leverages grid search to select a combination of augmentation operations for various datasets. AWS [37] designs an augmentation-wise weight-sharing strategy to search augmentation operations. In contrast, TrivialAugment [28] adopts the same augmentation space obtained by these methods but opts for a simpler approach by applying a single augmentation operation to each image during training. SelectAugment [24] utilizes a two-step Markov decision process and hierarchical reinforcement learning to learn the augmentation policy and select samples to be augmented online. However, as SelectAugment employs AutoAugment, Mixup, or CutMix for data augmentation, the strengths of the augmentation applied remain uncontrollable. Moreover, AdDA [49] is proposed for contrastive learning, which learns to adaptively adjust the augmentation compositions and achieves more generalizable representations. SoftAug [25] generates augmentation with invariant transforms to soft augmentations. MADAug [16] jointly trains an augmentation policy network through a bi-level optimization scheme to select augmentation operations for samples. TeachAugment [36] leverages a teacher model to generate the transformed data based on the adversarial strategy. DADA [21] relaxes the discrete DA policy selection to a differentiable optimization problem, facilitating efficient and accurate DA policy learning. Since these methods adopt augmentation operations with random or fixed parameters, the magnitude of variations in the augmented data is difficult to adjust. Our proposed method distinguishes itself by adaptively determining the augmentation magnitudes based on image characteristics and the dynamics of the model training process, thereby enhancing the flexibility and efficacy of data augmentation.

## 2.2 Entropy-Regularized Loss

In the realm of deep learning and machine learning, entropy-regularized loss functions are typically designed to encourage models to produce predictions that are more evenly spread or uncertain [4, 32]. Particularly within the domains of reinforcement learning and decision-making processes, entropy regularization finds widespread application to improve the policy optimization process, e.g., obtaining high-entropy output to encourage exploration [2, 8, 11]. Minimum entropy regularizers have been used in other contexts, such as semi-supervised learning problem [13], unsupervised clustering [38], structured output prediction [1], and unsupervised domain adaption [41], etc. In contrast, our work advances the proposition of integrating entropy-regularized loss into the conventional CE loss to expedite model convergence and mitigate model fitting errors.

### 3 The Proposed Method

In this section, we first propose EntAugment. We then introduce the entropy-regularized loss EntLoss and study how it improves the performance of EntAugment and the overall generalization capabilities of task models.

#### 3.1 EntAugment

EntAugment is motivated by a straightforward intuition that conventional data augmentation, which employs a uniform strategy characterized by random magnitudes, may lead to sub-optimal performance. Regardless of specific DA operations, it is imperative to dynamically adjust the magnitude of these operations in accordance with the progress of the target model training and the inherent difficulty of individual samples. For instance, for samples that are classified by the current target model with high confidence, suggesting relative ease of classification, augmentation should encompass a broad spectrum of scenarios to diversify the training data. Conversely, samples that models struggle to classify with high confidence scores, indicating a significant challenge in learning, necessitate more subtle variations in augmentation. Nevertheless, it is essential to acknowledge that the effective and efficient assessment of sample difficulty, while concurrently considering the model’s training status, poses a formidable challenge. In light of this challenge, we propose EntAugment.

EntAugment leverages the augmentation space that has been utilized in previous works [6, 24, 28]. Let  $\mathcal{D}$  denote the training dataset comprising  $N$  training samples, each of the form  $(\mathbf{x}, y) \in \mathcal{D}$ .  $\mathbf{x}$  represents the original data, and  $y = 1, \dots, k$  is the corresponding label, where  $k$  is the total number of classes. Given a classification model  $f_\theta$  parameterized by  $\theta$  and an input sample  $\mathbf{x} \in \mathbb{R}^n$ ,  $f_\theta(\mathbf{x})$  is the network output. For simplicity, we use  $g(\mathbf{x}) = \text{softmax}(f_\theta(\mathbf{x})) \in \mathbb{R}^k$  to denote the output of the softmax function. Therefore,  $g(\mathbf{x})$  is indeed a probability distribution, i.e.,  $\sum_{i=1}^k g(\mathbf{x})_i = 1$ . The information entropy of the softmax output is defined as:

$$-\sum_{i=1}^k g(\mathbf{x})_i \log g(\mathbf{x})_i, \quad (1)$$

which indicates the confidence level of model  $f_\theta$  in classifying  $\mathbf{x}$ . In cases where samples pose challenges for classification, the entropy associated with the model’s output tends to exhibit relatively higher values, whereas lower values indicate easily classifiable data. One notable advantage of the entropy measure defined in Eq. (1) lies in its **dual functionality** - encapsulating information regarding sample complexity and insights into model evolution during training. Therefore, dynamic adjustment of DA magnitude becomes viable. Specifically, the magnitude of the DA manipulation is determined by:

$$\text{mag}(\mathbf{x}) = 1 + \frac{1}{\log k} \sum_{i=1}^k g(\mathbf{x})_i \log g(\mathbf{x})_i. \quad (2)$$

**Algorithm 1** General EntAugment Procedure

**Require:** a batch of image and class label pairs  $(\mathbf{x}, y)$ , batch size  $bs$ , total number of classes  $k$ , an augmentation space  $\mathcal{A}$

**Ensure:** a batch of augmented image  $\mathbf{x}'$

**for**  $i = 0: bs - 1$  **do**

    Calculate the magnitude of the DA manipulation

    for  $\mathbf{x}_i$  based on Eq. (2), denoted as  $mag_i$

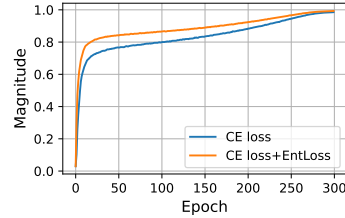
    Randomly sample an augmentation operation

$\mathbf{a}$  from  $\mathcal{A}$

    Augment sample  $\mathbf{x}_i$ :  $\mathbf{x}'_i = \mathbf{a}(\mathbf{x}_i, mag_i)$

**end for**

**return** a batch of augmented data  $\mathbf{x}'$



**Fig. 2:** Values of magnitude defined in Eq. (2) with and without employing EntLoss over the training process.

In this way,  $mag(\mathbf{x})$  scales to  $[0, 1]$  and exhibits an inverse proportional relationship with the entropy measures. In situations where  $mag(\mathbf{x}) \rightarrow 1$ , the augmented samples exhibit a greater degree of variability, while conversely, minor alterations occur as  $mag(\mathbf{x}) \rightarrow 0$ . Additionally, it is noteworthy to highlight that  $mag(\mathbf{x})$  for each sample  $\mathbf{x}$  remains dynamic and evolves continuously throughout the entire training process, **enabling adaptive augmentation**.

Based on Eq. (2), during the initial phase of model training, most samples are difficult to classify, thereby assigned with relatively low magnitudes. Providing relatively simple samples during the prior stage of model training contributes to the enhancement of overall model performance [35, 40]. As models acquire enhanced generalization capabilities, the entropy values in Eq. (1) tend to decrease for more data, indicating higher magnitudes defined by Eq. (2). More augmented data will be provided in more diverse scenarios. Thus, the adaptive data augmentation framework can be ensured. The procedure of EntAugment is outlined in Algorithm 1.

### 3.2 Entropy-Regularized Loss

Although EntAugment performs well theoretically, the model’s confidence scores in data classification may slowly increase and stabilize at a moderate level. As illustrated in Figure 2, the magnitudes defined in Eq. (2) and derived using vanilla CE loss remain consistently low throughout most training epochs. Nonetheless, this could lead to inadequate DA magnitudes and a lack of potential diversity in the augmented data, particularly during later stages of model training.

To address this issue, based on Eq. (1) and Eq. (2), we introduce an entropy regularization loss, which is defined as:

$$\text{EntLoss}(\mathbf{x}) = \frac{1}{\log k} \sum_{i=1}^k g(\mathbf{x})_i \log g(\mathbf{x})_i. \quad (3)$$

When incorporated with vanilla CE loss, EntLoss encourages models to classify samples with higher confidence, i.e., the models trained with EntLoss attain lower entropy values. Let  $\hat{\theta}_{cro}$  denote the empirical risk minimizer employing the conventional CE loss, and  $\hat{\theta}_{ent}$  be the empirical risk minimizer utilizing both the CE loss and EntLoss. Given that the entropy function is denoted as  $\text{Ent}(\cdot)$ , we have

$$\text{Ent}(p_{\hat{\theta}_{cro}}) \leq \text{Ent}(p_{\hat{\theta}_{ent}}), \quad (4)$$

where  $p_{\theta}$  is the probability distribution of the model  $f_{\theta}$ 's output. Furthermore, as shown in Figure 2, the confidence level and magnitudes increase substantially after incorporating EntLoss into the training process. More importantly, EntLoss brings more significant advantages in training classification models.

**Theoretical Analysis** We will theoretically demonstrate that EntLoss can mitigate dissimilarity between the distributions of the model and dataset compared to the conventional CE loss. This facilitates the model to fit the dataset distribution better, thereby enhancing its generalization capability.

Our initial exposition delves into the equivalence between the effect of the vanilla CE loss function and the minimization of the Kullback-Leibler (KL) divergence between the model's and the dataset's distributions. Consider a dataset denoted as  $\mathcal{D} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$ , comprising  $N$  training instances, where each  $\mathbf{z}_i = (\mathbf{x}_i, y_i)$  is independently generated from an unknown potential data distribution, denoted as  $p_{data}(\mathbf{z})$ . Further, let  $p_{model}(\boldsymbol{\theta})$  represent a parametric family of probability distributions over the same space indexed by  $\boldsymbol{\theta}$ . We have the following Lemma.

**Lemma 1.** *From the perspective of maximum likelihood estimation, the empirical risk minimizer can be denoted as  $\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{z} \sim \hat{p}_{data}} \log p_{model}(\mathbf{z}; \boldsymbol{\theta})$ , where  $\hat{p}_{data}$  is the empirical distribution defined by the training set.*

Lemma 1 is formally proven in Appendix B. Based on Lemma 1, the process of model training can be comprehended as minimizing the dissimilarity between the empirical distribution  $\hat{p}_{data}$  and the model distribution  $p_{model}$ . Furthermore, the extent of dissimilarity between these two distributions can be effectively measured by the KL divergence, which is given by:

$$D_{\text{KL}}(\hat{p}_{data} \parallel p_{model}) = \mathbb{E}_{\mathbf{z} \sim \hat{p}_{data}} [\log \hat{p}_{data}(\mathbf{z}) - \log p_{model}(\mathbf{z})]. \quad (5)$$

$\mathbb{E}_{\mathbf{z} \sim \hat{p}_{data}} [\log \hat{p}_{data}(\mathbf{z})]$  is a function only of the training set, not the model. When we train the model to minimize the KL divergence, we need to minimize

$$-\mathbb{E}_{\mathbf{z} \sim \hat{p}_{data}} [\log p_{model}(\mathbf{z})], \quad (6)$$

which can be estimated by  $-\frac{1}{n} \sum_{i=1}^n p_{data}(\mathbf{z}_i) \log p_{model}(\mathbf{z}_i)$ . Therefore, the following Lemma can be derived.

**Lemma 2.** *Minimizing the cross entropy loss is equivalent to minimizing the KL divergence between the model's distribution and the dataset's distribution, as shown in Eq. (5).*

However, the model’s output is indeed a continuous variable, i.e.,  $\theta(\mathbf{z}_i) \in \mathbb{R}^k, i = 1, 2, \dots, n$ . In contrast, the distribution of the dataset is structured as a one-hot variable. For example, if  $\mathbf{x}_i$  belongs to the  $k'$ -th class,  $y_i = [0, \dots, 1, \dots, 0]$ , where  $y_{ij} = 1$  if  $j = k'$  and  $y_{ij} = 0$ , otherwise. This intrinsic inconsistency between the model’s output and the dataset distribution gives rise to a notable theoretical discrepancy. Thus, EntLoss can be used to complement the conventional CE loss function, making the model distribution closer to the data distribution.

**Lemma 3.** *Suppose that  $\hat{\theta}_{cro}$  is the empirical risk minimizer using the vanilla CE loss and  $\hat{\theta}_{ent}$  is the empirical risk minimizer utilizing the CE loss along with the EntLoss, we have*

$$-\mathbb{E}_{\mathbf{z} \sim \hat{p}_{data}} \left[ \log p_{\hat{\theta}_{ent}}(\mathbf{z}) \right] \leq -\mathbb{E}_{\mathbf{z} \sim \hat{p}_{data}} \left[ \log p_{\hat{\theta}_{cro}}(\mathbf{z}) \right] \quad (7)$$

**Proposition 1.** *Based on Eq. (4), (5), and Lemma 3, we have*

$$D_{KL}(\hat{p}_{data} || p_{\hat{\theta}_{cro}}) \geq D_{KL}(\hat{p}_{data} || p_{\hat{\theta}_{ent}}). \quad (8)$$

Proposition 1 is proven in Appendix C. It can be seen that the dissimilarity between the model distribution  $p_{\hat{\theta}_{cro}}$  and the underlying dataset distribution serves as an upper bound for the dissimilarity between the model distribution  $p_{\hat{\theta}_{ent}}$  and the dataset distribution. Proposition 1 reveals the advantageous effect of EntLoss in mitigating the dissimilarity between the model’s distribution and the dataset’s distribution. Thus, EntLoss facilitates models to better fit the inherent distribution of the dataset, thereby achieving enhanced generalization.

In conclusion, EntLoss can be employed to enhance EntAugment or utilized as a standalone optimization method for training classification models.

**Complexity Analysis of EntLoss and EntAugment.** We provide theoretical analysis demonstrating that the utilization of EntLoss and EntAugment does not introduce any notable additional computational cost. Specifically, the computational complexity of CE loss is  $O(k \times N + N)$ , where  $N$  denotes the number of samples and  $k$  is the number of categories. Both EntAugment and EntLoss exhibit a time complexity of  $O(k \times N)$ . Thus, the computational complexity of EntAugment alone is closely equivalent to that of the vanilla CE loss, highlighting its efficiency. Meanwhile, the overall computational complexity is  $O((2k + 1) \times N)$ , which is equivalent to the vanilla CE loss.

## 4 Experiment

**Comparison with state-of-the-arts** We compare our methods with the 11 most representative and commonly used data augmentation methods, including HaS [34], Cutout [9], CutMix [45], GridMask (GM) [3], AdvMask (AM) [43], Random Erasing (RE) [51], AutoAugment (AA) [6], Fast-AutoAugment (FAA) [23], RandAugment (RA) [7], DADA [21], TrivialAugment (TA) [28], TeachAugment (TeachA) [36], MADAug [17], and SoftAug [25].

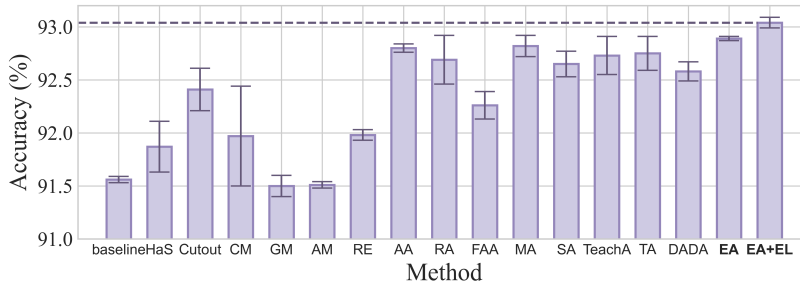
**Table 1:** Image classification accuracy (%) on CIFAR-10/100 (average  $\pm$  std). \* means results reported in the original paper. EA+EL: EntAugment+EntLoss.

Method	R-18 [15]	R-44 [15]	R-50 [15]	WRN [46]	SS-32 [10]	R-18 [15]	R-44 [15]	R-50 [15]	WRN [46]	SS-32 [10]
	CIFAR-10 [19]					CIFAR-100 [19]				
baseline	95.28 $\pm$ .14*	94.10 $\pm$ .40	95.66 $\pm$ .08	95.52 $\pm$ .11	94.90 $\pm$ .07*	77.54 $\pm$ .19*	74.80 $\pm$ .38*	77.41 $\pm$ .27*	78.96 $\pm$ .25*	76.65 $\pm$ .14*
HaS [34]	96.10 $\pm$ .14*	94.97 $\pm$ .27	95.60 $\pm$ .15	96.94 $\pm$ .08	96.89 $\pm$ .10*	78.19 $\pm$ .23	75.82 $\pm$ .32	78.76 $\pm$ .24	80.22 $\pm$ .16	76.89 $\pm$ .33
Cutout [9]	96.01 $\pm$ .18*	94.78 $\pm$ .35	95.81 $\pm$ .17	96.92 $\pm$ .09	96.96 $\pm$ .09*	78.04 $\pm$ .10*	74.84 $\pm$ .56	78.62 $\pm$ .25	79.84 $\pm$ .14	77.37 $\pm$ .28
CutMix [45]	96.64 $\pm$ .22*	95.28 $\pm$ .16	96.81 $\pm$ .10*	96.93 $\pm$ .10*	96.47 $\pm$ .07	79.45 $\pm$ .17	76.09 $\pm$ .15	81.24 $\pm$ .14	82.67 $\pm$ .22	79.57 $\pm$ .10
GridMask [3]	96.38 $\pm$ .17	95.02 $\pm$ .26	96.15 $\pm$ .19	96.92 $\pm$ .09	96.91 $\pm$ .12	75.23 $\pm$ .21	76.07 $\pm$ .18	78.38 $\pm$ .22	80.40 $\pm$ .20	77.28 $\pm$ .38
AdvMask [43]	96.44 $\pm$ .15*	95.49 $\pm$ .17*	96.69 $\pm$ .10*	97.02 $\pm$ .05*	97.03 $\pm$ .12*	78.43 $\pm$ .18*	76.44 $\pm$ .18*	78.99 $\pm$ .31*	80.70 $\pm$ .25*	79.96 $\pm$ .27*
RE [51]	95.69 $\pm$ .10*	94.87 $\pm$ .16*	95.82 $\pm$ .17	96.92 $\pm$ .09	96.46 $\pm$ .13*	75.97 $\pm$ .11*	75.71 $\pm$ .25*	77.79 $\pm$ .32	80.57 $\pm$ .15	77.30 $\pm$ .18
AA [6]	96.51 $\pm$ .10*	95.01 $\pm$ .11	96.59 $\pm$ .04*	96.99 $\pm$ .06	97.30 $\pm$ .11	79.38 $\pm$ .20	76.36 $\pm$ .22	81.34 $\pm$ .29	82.21 $\pm$ .17	82.19 $\pm$ .19
FAA [23]	95.99 $\pm$ .13	93.80 $\pm$ .12	96.69 $\pm$ .16	97.30 $\pm$ .24	96.42 $\pm$ .12	79.11 $\pm$ .09	76.04 $\pm$ .28	79.08 $\pm$ .12	79.95 $\pm$ .12	81.39 $\pm$ .16
RA [7]	96.47 $\pm$ .32	94.38 $\pm$ .22	96.25 $\pm$ .06	96.94 $\pm$ .13*	97.05 $\pm$ .15	78.30 $\pm$ .15	76.30 $\pm$ .16	80.95 $\pm$ .22	82.90 $\pm$ .29*	80.00 $\pm$ .29
DADA [21]	95.58 $\pm$ .06	93.96 $\pm$ .38	95.61 $\pm$ .14	97.30 $\pm$ .13*	97.30 $\pm$ .14*	78.28 $\pm$ .22	74.37 $\pm$ .47	80.25 $\pm$ .28	82.50 $\pm$ .26*	80.98 $\pm$ .15
TA [28]	96.28 $\pm$ .10	95.00 $\pm$ .10	97.13 $\pm$ .08	97.18 $\pm$ .11	97.30 $\pm$ .10	78.67 $\pm$ .19	76.57 $\pm$ .14	81.34 $\pm$ .18	82.75 $\pm$ .26	82.14 $\pm$ .16
TeachA [36]	96.47 $\pm$ .13	95.05 $\pm$ .21	96.40 $\pm$ .14	97.50 $\pm$ .16	97.29 $\pm$ .11	79.27 $\pm$ .24	76.18 $\pm$ .31	80.54 $\pm$ .25	82.81 $\pm$ .26	81.30 $\pm$ .18
MADAug [17]	96.49 $\pm$ .12	95.25 $\pm$ .18	97.12 $\pm$ .17	97.48 $\pm$ .15	97.37 $\pm$ .11	79.39 $\pm$ .18	76.49 $\pm$ .21	81.40 $\pm$ .18	83.01 $\pm$ .23	81.67 $\pm$ .19
SoftAug [25]	96.43 $\pm$ .15	94.51 $\pm$ .20	96.99 $\pm$ .14	97.15 $\pm$ .16	97.22 $\pm$ .19	79.01 $\pm$ .21	76.41 $\pm$ .33	80.94 $\pm$ .33	82.61 $\pm$ .24	80.33 $\pm$ .20
EntAugment	96.71 $\pm$ .05	95.76 $\pm$ .09	97.09 $\pm$ .09	97.47 $\pm$ .10	97.46 $\pm$ .11	79.45 $\pm$ .17	76.40 $\pm$ .18	81.56 $\pm$ .21	83.09 $\pm$ .22	81.60 $\pm$ .13
EA+EL	96.84 $\pm$ .09	95.58 $\pm$ .03	97.15 $\pm$ .09	97.70 $\pm$ .12	97.55 $\pm$ .10	79.82 $\pm$ .12	76.84 $\pm$ .03	82.49 $\pm$ .15	83.16 $\pm$ .23	82.29 $\pm$ .18

**Implementation Details** We closely follow previous works [9, 28, 43] with our setup. Specifically, images are preprocessed by dividing each pixel value by 255 and normalized by the dataset statistics. We train 1800 epochs with cosine learning rate decay for Shake-Shake [10] using SGD with Nesterov Momentum and a learning rate of 0.1, a batch size of 256,  $1e^{-3}$  weight decay and cosine learning rate decay. We train all other models for 300 epochs with a batch size of 256 and a 0.1 learning rate with cosine annealing learning rate decay strategy, SGD optimizer with the momentum of 0.9, and weight decay of  $5e^{-4}$ . We follow the common practice in the field of DA method studies [9, 43] to build the baseline model, i.e., data augmentation of random crop and random horizontal flip is utilized as the baseline. For fairness, all methods are implemented with the same configurations. The augmentation space utilized follows previous work [6, 24, 28], while the magnitude of each strategy is dynamically determined. The experiments are repeated across three independent runs.

#### 4.1 Results on CIFAR-10 and CIFAR-100

To evaluate the effectiveness of EntAugment and EntLoss, in Table 1, we conduct experiments on CIFAR-10/100 utilizing various deep networks, including ResNet18/44/50 (R-18/44/50) [15], Wide-ResNet-28-10 (WRN) [46], and Shake-Shake-26-32 (SS-32) [10]. It can be observed that when employed independently, EntAugment consistently surpasses prior SOTA methods in most cases. Moreover, the integration of EntLoss into the EntAugment framework consistently demonstrates notable performance improvements across deep models on both CIFAR-10 and CIFAR-100 datasets. For instance, on CIFAR-10, despite the high test accuracy already achieved by DA methods, both EntAugment and EntAugment+EntLoss further enhance the model performance. In particular, EntAugment+EntLoss typically yields accuracy improvements of approximately 0.5% on CIFAR-10. On CIFAR-100, the improvements are even more substantial, with the proposed method surpassing previous SOTA methods by nearly



**Fig. 3: Transferability analysis.** Transferred test accuracy (%) on CIFAR-10.

**Table 2: Top-1 accuracy (%) on ImageNet dataset (average  $\pm$  std).**

HaS	GM	Cutout	CutMix	Mixup	AA	FAA	RA	MA	SA	DADA	TA	TeachA	EA	EA+EL
77.2 $\pm$ 0.2	77.9 $\pm$ 0.2	77.1 $\pm$ 0.3	77.2 $\pm$ 0.2	77.0 $\pm$ 0.2	77.6 $\pm$ 0.2	77.6 $\pm$ 0.2	77.6 $\pm$ 0.2	<b>78.5</b> $\pm$ 0.1	78.0 $\pm$ 0.1	77.5 $\pm$ 0.1	77.9 $\pm$ 0.3	77.8 $\pm$ 0.2	78.2 $\pm$ 0.1	78.3 $\pm$ 0.1

1% when employing ResNet-18/50 architectures. Hence, the proposed adaptive DA framework is more effective in boosting model performance.

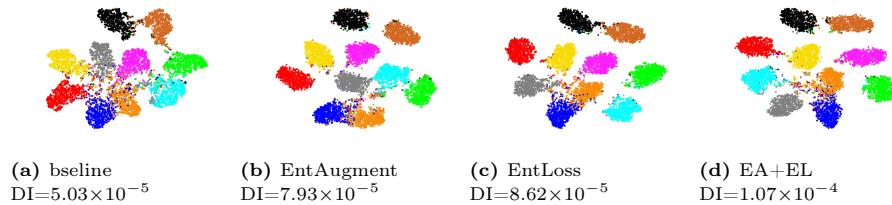
Meanwhile, it is worth noting that such substantial performance improvements are achieved without incurring noticeable computational costs, underscoring its superior efficiency. Consequently, EntAugment and EntLoss can serve as highly efficient plug-and-play techniques for model training.

## 4.2 Transfer Learning

In the realm of data augmentation, transfer learning is often utilized to assess the transferability of DA methods [16, 31, 52]. Thus, we pre-train models on the CIFAR-100 dataset using various augmentations and fine-tune these models on CIFAR-10 using ResNet-50. Theoretically, models trained using more effective DA methods should yield stronger transferability. As shown in Figure 3, while the discrepancies in transferred accuracy may appear subtle, it can be observed that EntAugment outperforms other SOTA methods in terms of transferred test accuracy. Moreover, EntAugment+EntLoss demonstrates even more significant performance improvements.

## 4.3 Results on Large-scale ImageNet

We also evaluate our framework on the large-scale ImageNet [20] dataset. Specifically, we train ResNet-50 models on ImageNet using various DA methods, closely following the experimental setup in [6, 28]. In Table 2, it can be observed that while prior methods present similar and modest improvements compared to the baseline (e.g., less than 0.9%), our methods demonstrate a substantial superiority by achieving improvements exceeding 1.2%. While EA is slightly worse



**Fig. 4: t-SNE Visualization of CIFAR-10 dataset.** DI: Dunn index. EA+EL: EntAugment+EntLoss.

than MADAug, ours achieves a 2x speed increase over MADAug. Meanwhile, the training of our proposed methods achieves a 4x faster than TeachAugment. This highlights the superior effectiveness of our approaches on large-scale datasets.

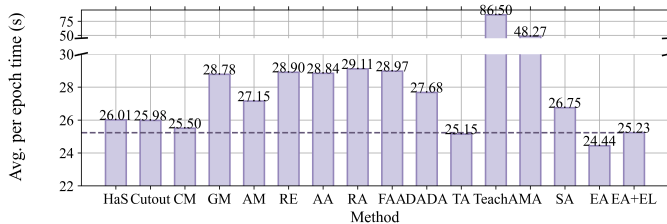
Notably, most well-performed methods (e.g., AdvMask [43], AutoAugment [6], RandAugment [7], TeachAugment [36], MADAugment [17] etc.) entail significant additional training costs. Conversely, the proposed methods obtain nearly equivalent computational overhead compared to the baseline. Consequently, our proposed methods achieve better results while incurring nearly negligible additional computational overhead. More results on Tiny-Imagenet and large-scale long-tailed ImageNet-LT [26] and Places-LT [26] are provided in Appendix D and E.

#### 4.4 t-SNE Visualizations

To conduct a comparative analysis of model performance with and without utilizing the proposed methods, we visualize the deep features of the CIFAR-10 dataset using t-SNE algorithm [27]. Specifically, we employ EntAugment, EntLoss, and EntAugment+EntLoss to train ResNet-18 models. Subsequently, we extract deep features using these trained models and leverage the t-SNE algorithm to analyze the model performance.

In Figure 4, it can be observed intuitively that applying EntAugment or EntLoss independently results in a pronounced delineation in cluster distribution, i.e., **enhanced inter-cluster separation and intra-cluster compactness**. This phenomenon is further enhanced when EntAugment and EntLoss are utilized jointly, as shown in Figure 4(d), where the discriminative characteristics of learned features are effectively enhanced. Therefore, the proposed methods can bolster the generalization capabilities of the models, thereby facilitating the extraction of more discriminative features.

Moreover, we utilize the Dunn index (DI) [30] to quantitatively analyze the clustering results, which is  $DI = \min_{1 \leq i \neq j \leq m} \delta(C_i, C_j) / \max_{1 \leq j \leq m} \Delta_j$ , where separation  $\delta(C_i, C_j)$  is the inter-cluster distance metric between clusters  $C_i$  and  $C_j$ , and compactness  $\Delta_j$  calculates the mean distance between all pairs in each cluster. Hence, a higher DI means better clustering. As shown in Figure 4, using EntAugment and EntLoss individually can bring higher DI values than the



**Fig. 5: Time consumption.** Comparison of training cost of various DA methods.

baseline. More precisely, the DI values achieved by EntAugment and EntLoss are found to be 57.7% and 71.4% higher than those of the baseline, respectively. Furthermore, for EntAugment + EntLoss, the DI value experiences a substantial increase, effectively doubling the DI value of the baseline. Consequently, EntAugment and EntLoss can be utilized separately or jointly to enhance the model performance. This investigation not only underscores the efficacy of these methods but also contributes to the interpretability of the proposed techniques.

#### 4.5 Comparison of Training Efficiency

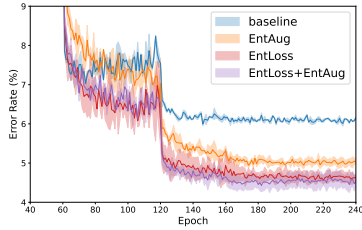
To demonstrate the efficiency of our proposed methods, we present a comparison of the training cost associated with employing various DA methods. All experiments are conducted on 2 NVIDIA RTX2080TI GPUs with batch size 128 and 8 parallel workers. The experiments are repeated across five independent runs. The average time cost per epoch is presented in Figure 5. Consistent with the theoretical analysis in Section 3.2, the time cost of our proposed methods remains at the lowest level, highlighting the efficiency. Although the proposed methods obtain similar time costs with CutMix and TrivialAugment, ours consistently achieves better performance, highlighting the practical effectiveness.

#### 4.6 Convergence Analysis

To more clearly present the dynamic evolution of test errors throughout the training process, we train ResNet-110 models [15] on CIFAR-10 using a multi-step learning rate decay schedule. The learning rate is initialized as 0.1 and multiplied by 0.2 at epochs 60, 120, 160, 220, and 280. It is worth noting that all other experimental settings remain unchanged. The convergence trajectory is shown in Figure 6. It can be observed that both EntAugment and EntLoss achieve a significant improvement after the second learning rate drop. Meanwhile, when EntAugment is combined with EntLoss, it presents lower error rates, suggesting more efficient convergence.

#### 4.7 Ablation Study

**The effect of EntAugment** To validate the efficacy of EntAugment, we present the results of merely using EntAugment across benchmark datasets in



**Fig. 6: Convergence analysis.** Test error on CIFAR with ResNet-110.

**Table 3: Effect of EntLoss.** Comparative analysis between the test accuracy with and without employing EntLoss. CELoss: cross-entropy loss.

model	CELoss	CELoss+EntLoss	CELoss	CELoss+EntLoss
	CIFAR-10 [19]		CIFAR-100 [19]	
ResNet18 [15]	95.28%	<b>95.41%</b> $\uparrow 0.13$	77.54%	<b>79.17%</b> $\uparrow 1.63$
ResNet44 [15]	94.10%	<b>94.73%</b> $\uparrow 0.63$	71.75%	<b>73.20%</b> $\uparrow 1.45$
ResNet50 [15]	95.66%	<b>95.88%</b> $\uparrow 0.22$	77.41%	<b>80.29%</b> $\uparrow 2.88$
WRN-28-10 [46]	95.52%	<b>96.28%</b> $\uparrow 0.76$	78.96%	<b>80.78%</b> $\uparrow 1.82$
Shake-26-32 [10]	94.90%	<b>96.55%</b> $\uparrow 0.65$	76.65%	<b>79.15%</b> $\uparrow 2.50$

Table 1 and 2, as well as results in Appendix D, F, and H. The results demonstrate that the EntAugment framework outperforms prior SOTA DA methods in most cases, highlighting the effectiveness of the adaptive DA scheme.

**The effect of EntLoss** In this section, we will demonstrate that the utilization of EntLoss yields substantial improvements compared to conventional classification model training. Comparative analyses between models trained with and without utilizing EntLoss are presented in Table 3. It can be observed that EntLoss brings a notable improvement in model performance when combined with vanilla CE loss (consistent with the theoretical analysis in Section 3.2). Especially on CIFAR-100, EntLoss can improve test accuracy by nearly 2%. It is also noteworthy that such significant improvements are achieved without introducing any noticeable computational overhead.

## 5 Conclusion

In this paper, we propose a novel, tuning-free, and adaptive DA framework, EntAugment. EntAugment operates without the need for manual tuning and dynamically adjusts the magnitudes of DA applied to each training data during online training based on sample difficulty and model training progression. Moreover, we also introduce an entropy regularization loss, EntLoss, to enhance the effectiveness of EntAugment for better generalization. Through theoretical analysis, we show that EntLoss brings more significant benefits of achieving closer alignment between the model’s distribution and the dataset’s inherent distribution. Notably, without introducing auxiliary models or extra training overhead, both EntAugment and EntLoss introduce minimal computational cost to the task model training process, ensuring ease of integration and practical feasibility. Experimental results on several benchmark datasets show that the proposed methods achieve state-of-the-art performance while showcasing superior efficiency. Meanwhile, models trained with EntAugment can exhibit enhanced transferability and generalization capabilities compared to other augmentations. In the future, we will explore the application of the proposed methods on other popular tasks, e.g., self-supervised learning and object detection.

## Acknowledgements

This work was supported in part by the STI 2030-Major Projects of China under Grant 2021ZD0201300 and by the National Natural Science Foundation of China under Grant 62276127.

## References

1. Ahmed, K., Wang, E., Chang, K.W., Van den Broeck, G.: Neuro-symbolic entropy regularization. In: Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI). PMLR (2022)
2. Ahmed, Z., Le Roux, N., Norouzi, M., Schuurmans, D.: Understanding the impact of entropy on policy optimization. In: International conference on machine learning. pp. 151–160. PMLR (2019)
3. Chen, P., Liu, S., Zhao, H., Jia, J.: Gridmask data augmentation. arXiv preprint arXiv:2001.04086 (Jan 2020)
4. Chong, S.S.: Loss function entropy regularization for diverse decision boundaries. In: 2022 7th International Conference on Big Data Analytics (ICBDA). pp. 123–129. IEEE (2022)
5. Chrabaszcz, P., Loshchilov, I., Hutter, F.: A downsampled variant of imagenet as an alternative to the cifar datasets. arXiv preprint arXiv:1707.08819 (Aug 2017)
6. Cubuk, E.D., Zoph, B., Mané, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation strategies from data. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 113–123 (2019)
7. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.: Randaugment: Practical automated data augmentation with a reduced search space. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Proc. Adv. Neural Inf. Process. Syst. vol. 33, pp. 18613–18624 (2020)
8. Della Vecchia, R., Shilova, A., Preux, P., Akrou, R.: Entropy regularized reinforcement learning with cascading networks. arXiv preprint arXiv:2210.08503 (2022)
9. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (Nov 2017)
10. Gastaldi, X.: Shake-shake regularization. arXiv preprint arXiv:1705.07485 (May 2017)
11. Ghandi, S., Tavakol, M.: Entropy-regularized model-based offline reinforcement learning (2023), <https://openreview.net/forum?id=bBBA-8ELXcF>
12. Gong, C., Wang, D., Li, M., Chandra, V., Liu, Q.: Keepaugment: A simple information-preserving data augmentation approach. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 1055–1064 (2021)
13. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. Advances in neural information processing systems **17** (2004)
14. Hacoheh, G., Weinshall, D.: On the power of curriculum learning in training deep networks. In: International conference on machine learning. pp. 2535–2544. PMLR (2019)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 770–778 (2016)
16. Hou, C., Zhang, J., Zhou, T.: When to learn what: Model-adaptive data augmentation curriculum. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV). pp. 1717–1728 (2023)

17. Hou, C., Zhang, J., Zhou, T.: When to learn what: Model-adaptive data augmentation curriculum. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1717–1728 (2023)
18. Iglesias, G., Talavera, E., González-Prieto, Á., Mozo, A., Gómez-Canaval, S.: Data augmentation techniques in time series domain: a survey and taxonomy. *Neural Computing and Applications* **35**(14), 10123–10145 (2023)
19. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Communications of the ACM* **60**(6), 84–90 (2017)
21. Li, Y., Hu, G., Wang, Y., Hospedales, T.M., Robertson, N.M., Yang, Y.: DADA: differentiable automatic data augmentation. In: European Conference on Computer Vision. p. 580–595. Springer (2020)
22. Li, Y., Kim, Y., Park, H., Geller, T., Panda, P.: Neuromorphic data augmentation for training spiking neural networks. In: European Conference on Computer Vision. pp. 631–649. Springer (2022)
23. Lim, S., Kim, I., Kim, T., Kim, C., Kim, S.: Fast autoaugment. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), <https://proceedings.neurips.cc/paper/2019/file/6add07cf50424b14fdf649da87843d01-Paper.pdf>
24. Lin, S., Zhang, Z., Li, X., Chen, Z.: Selectaugment: hierarchical deterministic sample selection for data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1604–1612 (2023)
25. Liu, Y., Yan, S., Leal-Taixé, L., Hays, J., Ramanan, D.: Soft augmentation for image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16241–16250 (2023)
26. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2537–2546 (2019)
27. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. *J. Machine Learning Research* **9**(11) (2008)
28. Müller, S.G., Hutter, F.: Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 774–782 (2021)
29. Naveed, H., Anwar, S., Hayat, M., Javed, K., Mian, A.: Survey: Image mixing and deleting for data augmentation. *Engineering Applications of Artificial Intelligence* **131**, 107791 (2024)
30. Ncir, C.E.B., Hamza, A., Bouaguel, W.: Parallel and scalable dunn index for the validation of big data clusters. *Parallel Computing* **102**, 102751 (2021)
31. Pan, S.J., Yang, Q.: A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**(10), 1345–1359 (May 2010). <https://doi.org/10.1109/TKDE.2009.191>
32. Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., Hinton, G.: Regularizing neural networks by penalizing confident output distributions. arXiv preprint arXiv:1701.06548 (2017)
33. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. *Journal of big data* **6**(1), 1–48 (2019)
34. Singh, K.K., Lee, Y.J.: Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV). pp. 3544–3553. IEEE (Oct 2017)
35. Soviany, P., Ionescu, R.T., Rota, P., Sebe, N.: Curriculum learning: A survey. *International Journal of Computer Vision* **130**(6), 1526–1565 (2022)

36. Suzuki, T.: Techaugment: Data augmentation optimization using teacher knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10904–10914 (2022)
37. Tian, K., Lin, C., Sun, M., Zhou, L., Yan, J., Ouyang, W.: Improving auto-augment via augmentation-wise weight sharing. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Proc. Adv. Neural Inf. Process. Syst. vol. 33, pp. 19088–19098. Curran Associates, Inc. (2020), <https://proceedings.neurips.cc/paper/2020/file/dc49dfebb0b00fd44aef5c60cc1f825-Paper.pdf>
38. Wang, J., Ma, Z., Nie, F., Li, X.: Entropy regularization for unsupervised clustering with adaptive neighbors. Pattern Recognition **125**, 108517 (2022)
39. Wang, X., Chen, Y., Zhu, W.: A survey on curriculum learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(9), 4555–4576 (2021)
40. Wang, Y., Yue, Y., Lu, R., Liu, T., Zhong, Z., Song, S., Huang, G.: Efficienttrain: Exploring generalized curriculum learning for training visual backbones. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5852–5864 (October 2023)
41. Wu, X., Zhang, S., Zhou, Q., Yang, Z., Zhao, C., Latecki, L.J.: Entropy minimization versus diversity maximization for domain adaptation. IEEE Transactions on Neural Networks and Learning Systems **34**(6), 2896–2907 (2023). <https://doi.org/10.1109/TNNLS.2021.3110109>
42. Yang, S., Guo, S., Zhao, J., Shen, F.: Investigating the effectiveness of data augmentation from similarity and diversity: An empirical study. Pattern Recognition **148**, 110204 (2024). <https://doi.org/https://doi.org/10.1016/j.patcog.2023.110204>
43. Yang, S., Li, J., Zhang, T., Zhao, J., Shen, F.: Advmask: A sparse adversarial attack-based data augmentation method for image classification. Pattern Recognition **144**, 109847 (2023)
44. Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., Shen, F.: Image data augmentation for deep learning: A survey. arXiv preprint arXiv:2204.08610 (2022)
45. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6023–6032 (2019)
46. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: Richard C. Wilson, E.R.H., Smith, W.A.P. (eds.) Proceedings of the British Machine Vision Conference (BMVC). pp. 87.1–87.12. BMVA Press (September 2016). <https://doi.org/10.5244/C.30.87>, <https://dx.doi.org/10.5244/C.30.87>
47. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (2018), <https://openreview.net/forum?id=r1Ddp1-Rb>
48. Zhang, X., Wang, Q., Zhang, J., Zhong, Z.: Adversarial autoaugment. In: International Conference on Learning Representations (2020), <https://openreview.net/forum?id=ByxdUySKvS>
49. Zhang, Y., Zhu, H., Yu, S.: Adaptive data augmentation for contrastive learning. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
50. Zheng, Y., Zhang, Z., Yan, S., Zhang, M.: Deep autoaugmentation. In: ICLR (2022)
51. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proc. AAAI. vol. 34, pp. 13001–13008 (2020)
52. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning. Proc. IEEE **109**(1), 43–76 (2021). <https://doi.org/10.1109/JPROC.2020.3004555>