A Simple Episodic Linear Probe Improves Visual Recognition in the Wild

Yuanzhi Liang*[1,2], Linchao Zhu[2], Xiaohan Wang[3], and Yi Yang[3]

[1]Baidu Research
[2]ReLER Lab, AAII, University of Technology Sydney
[3]Zhejiang University
liangyzh18@outlook.com linchao.zhu@uts.edu.au xiaohan.wang@zju.edu.cn yangyics@zju.edu.cn

Abstract

Understanding network generalization and feature discrimination is an open research problem in visual recognition. Many studies have been conducted to assess the quality of feature representations. One of the simple strategies is to utilize a linear probing classifier to quantitatively evaluate the class accuracy under the obtained features. The typical linear probe is only applied as a proxy at the inference time, but its efficacy in measuring features' suitability for linear classification is largely neglected in training. In this paper, we propose an episodic linear probing (ELP) classifier to reflect the generalization of visual representations in an online manner. ELP is trained with detached features from the network and re-initialized episodically. It demonstrates the discriminability of the visual representations in training. Then, an ELP-suitable Regularization term (ELP-SR) is introduced to reflect the distances of probability distributions between the ELP classifier and the main classifier. ELP-SR leverages a re-scaling factor to regularize each sample in training, which modulates the loss function adaptively and encourages the features to be discriminative and generalized. We observe significant improvements in three real-world visual recognition tasks: fine-grained visual classification, long-tailed visual recognition, and generic object recognition. The performance gains show the effectiveness of our method in improving network generalization and feature discrimination.

1. Introduction

Deep neural networks have achieved impressive improvements in visual recognition. The neural networks trained on large-scale visual recognition datasets, e.g., ImageNet [30], OpenImages [27], demonstrate remarkable generalization capabilities. The learned visual representations are compact and enjoy strong discriminability. Many works have been conducted to theoretically explain the rationale behind deep networks' generalization [60], but this problem is still largely unsolved and remains to be investigated.

There are a few analytical tools to probe deep neural networks' learning and generalization capabilities. Early works utilize visualization tools to understand the optimized parameters or employ dimensionality reduction techniques to visualize the quality of learned representations [42, 51, 59]. Though helpful, such visualization techniques only provide qualitative inspections on deep networks [8]. Some works develop geometric probes to analyze the geometric properties of object manifold and connect object category manifolds' linear separability with the underlying geometric properties [46]. These methods reveal the structure of memorization from different layers in deep networks but only probe layer capacity at the inference time, as shown in Fig. 1 (a).

Another simple strategy is to perform linear probing. One can use linear probes to evaluate the feature's quality quantitatively. Since the discrimination capability of linear classifiers is low, linear classifiers heavily rely on the quality of the input representation to obtain good classification accuracy [3]. Alain *et al*. [1] use linear probes to examine the dynamics of intermediate layers. The linear probe is a linear classifier taking layer activations as inputs and measuring the discriminability of the networks. This linear probe does not affect the training procedure of the model. Recently, linear probes [3] have been used to evaluate feature generalization in self-supervised visual representation learning. After representation pre-training on pretext tasks [3], the learned feature extractor is kept fixed. The linear probe classifier is trained on top of the pre-trained feature representations. Though conceptually straightforward, linear probes are effective and have been widely used

*This work was performed at Baidu Research.

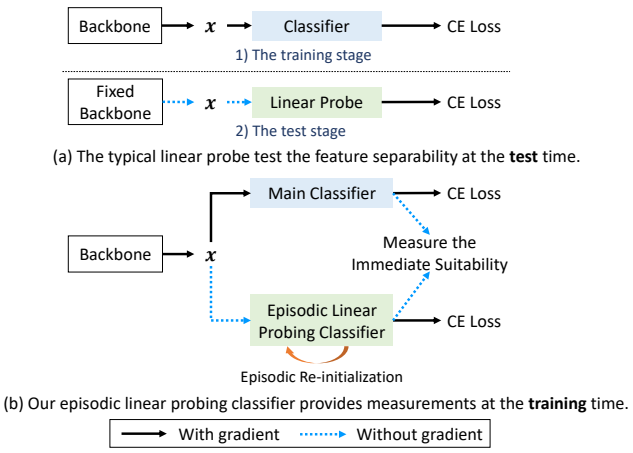

Figure 1. The typical linear probe in testing (a) and our ELP in training (b). Our ELP is episodically re-initialized to maintain simplicity. It effectively measures the discrimination of visual representations in an online manner.

in measuring the discriminability of visual representations. Noticeably, the linear probing classifier is only used in testing. A natural question arises: can we utilize linear probes during training and bring the signal from the linear probes to regularize the model training?

In this paper, we introduce a simple strategy to regularize the network to be immediately plausible for an episodic linear probing classifier. Our simple framework (Fig. 1 (b)) consists of a main classifier, an episodic linear probing classifier, and a regularization term. The regularization term considers the relation between the main classifier and the episodic linear probing classifier, which effectively penalizes examples that are not immediately plausible for episodic linear probes.

First, we propose an episodic linear probing (ELP) classifier to estimate the discrimination of visual representation in an online way. Similar to the existing linear probes [1], ELP is applied on top of the last layer of a deep network. ELP classifier is trained to classify the detached features into the same label space as a regular classifier. Different from [1], ELP is applied during model training. It is episodically re-initialized at each epoch. This maintains its simplicity, avoids classifier overfitting, and prevents the classifier from memorizing features. ELP implicitly reflects the feature discriminability and separability [40,41]. If the ELP classifier can quickly classify the feature points, it indicates that the given features are easily separable and would potentially be more generalizable.

Second, we introduce a penalization for less suitable examples for an episodic linear probe. Intuitively, given a training example, if the episodic linear probe and the main classifier contradict each other, *e.g.*, the episodic linear probe receives a *low* prediction score while the main classifier produces a *high* prediction score, it indicates that the main network exhibits overfitting on the given instance and a larger penalty should be enforced for proper regularization. Thus we design an ELP-suitable Regularization term (ELP-SR) to mitigate the intrinsic model bias and improve the linear separability of the learned features. ELP-SR sets a re-scaling factor to each instance and adaptively modulates the cross-entropy loss to avoid overfitting. The re-scaling factor considers the deviation between an example's predictive score from the main classifier and ELP classifier, which, to a certain extent, assesses the example's suitability for linear classification.

Without bells and whistles, our method achieves significant improvements for visual recognition tasks in the wild, providing consistent gains for fine-grained, long-tailed, and generic visual recognition. The fine-grained visual recognition datasets often contain high inter-class similarities. The long-tailed visual recognition datasets exhibit long-tailed data distribution, which is realistic in real-world recognition problems. We extensively evaluate the generalization performance on six standard datasets. The results indicate that our strategy empowers various deep networks with better discrimination and mitigates the model bias.

2. Related Work

Various works have been proposed to learn visual representation based on deep learning. In diverse recognition tasks in the wild, deep neural networks possess the powerful ability to learn and represent images to high-dimensional features. With the high-quality features, some simple classifiers [29, 56] are components to recognize the samples. Further, the quality of features is influenced by many factors. We roughly divided the factors into three aspects: data processing, network design, and training manner. Though the exact effect of representation learning [60] remains to be investigated, numerous researchers keep exploring and propose many valuable solutions.

For data processing, large-scale datasets provide considerable network samples and are the most straightforward way to improve representation. Benefiting from the powerful ability of networks, taking large-scale datasets as inputs lead the network to learn various samples and memorize plenty of properties for discriminating. Some diverse and hard examples may be difficult in a limited data scale [2,35]. Under the view of larger scales of collections, it is always possible for the network to mine particular patterns. Besides directly collecting real data, pre-processing [11,64] or generating data [63] are also equivalent. Various augmentations [43, 50] enforce the networks to solve problems with higher requirements and urge the network to be generalized to different conditions.

Moreover, well-designed network structures also dramatically boost representation and become the hottest di-

rection in recent years. Diverse methods constantly emerge like skip-connection [19, 22], fusing channels [48], attention strategies [4, 37], architecture searching [5], transformers [52, 54], etc. With the same inputs, these methods explore different directions to boost the network's capacity. Meanwhile, almost all kinds of visual tasks [30, 33] develop further with better networks.

Furthermore, besides data processing and network designs, the training manner is also crucial for visual representation. It contains various aspects like the optimizer [20, 39], regularization [31, 32], learning manner [25, 44], etc. In this direction, regularization plays an important role. It can be reflected in the loss function [9, 32], training strategies [18], etc., and is general to various networks and datasets. A proper regularization can leverage the network to learn better visual representation, for example, avoiding overfitting [32], explicit attention to the target [9], better diversity [14], etc. Vikash et al. [41] propose an interesting margin to describe the separability of features. Rather than focusing on the accuracy of the classifier, the quality of features can be reflected through immediate suitability. The more discriminative features are considered more than memorable by the classifier.

In our work, going further with the immediate suitability, we propose an episodic linear probing (ELP) classifier to reflect the generalization of visual representation online. ELP can be applied as a novel regularization to encourage the network to produce more discriminative features. Rather than re-weighting according to samples' easiness [25] or a meta set with iterative learning [44], we design an ELP-suitable regularization (ELP-SR) and leverage the ELP-SR to the regular loss function. Experimental results show that ELP-SR generally improves the performances of networks in three different benchmarks.

3. Method

In this work, we introduce an auxiliary episodic linear probing classifier to provide additional regularization for better representation learning. As illustrated in Fig. 2, our framework consists of three components, i.e., a deep neural network, a main linear classifier, and an episodic linear probing classifier. We illustrate our episodic linear probing classifier in Section 3.1. The details of the ELP-suitable regularization are introduced in Section 3.2. In Section 3.3, we describe the training and inference strategies of the model.

3.1. Episodic Linear Probing Classifier

3.1.1 Review of The Typical Linear Probes

Training the Feature Extractor. Given a training sample $\boldsymbol{x}$, a neural network (F) extracts its feature $\boldsymbol{h}$. A linear classifier (Cls) projects the feature to a probability distribution $\boldsymbol{p}$. The cross-entropy (CE) loss calculates the cross-entropy

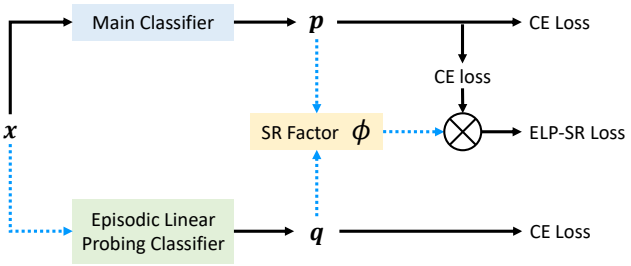

Figure 2. The training flow of our framework. Black lines indicate that the gradient can be back-propagated, while the blue dotted lines indicate that the gradient back-propagation is stopped.

between $\boldsymbol{p}$ and the ground-truth distribution $\boldsymbol{y}$. Formally, we denote the typical training procedure below:

$$\boldsymbol{h} = F(\boldsymbol{x}), \tag{1}$$

$$\boldsymbol{p} = Cls(\boldsymbol{h}), \tag{2}$$

$$\ell_{ce}(\boldsymbol{p}, \boldsymbol{y}) = -\sum_{j=1}^{C} y^j \log(p^j), \tag{3}$$

where C is the number of categories. $y^j = 1$ if j is the ground-truth label. Otherwise, $y^j = 0$. p^j is the prediction score of class j. The feature extractor and the classifier are jointly optimized end-to-end using back-propagation.

Test-time Linear Probing. Linear probing is usually built to assess the quality of deep representations after the neural network is sufficiently trained [1]. That amounts to training an auxiliary linear classifier on top of the pre-trained features. The parameters of the linear probe are randomly initialized, while the original classifier layer is neglected. The pre-trained backbone is frozen and not trained during linear probing. Since the complexity of the auxiliary classifier is not sufficient to provide additional discrimination, the classification performance heavily depends on the quality of the feature representations. Thus, predictive scores of the auxiliary linear classifier can probe the discrimination of the input features. During implementation, a linear probe can be extended to a Multi-Layer Perceptron (MLP) probe where the linear layer is replaced with a MLP [21].

The existing probes are mainly used during inference time, either providing quantitative evaluation on pre-trained features or interpreting intermediate layers [15]. This drives us to incorporate a linear probe during training and borrow the simple nature of the linear probe for network regularization.

3.1.2 Episodic Linear Probing Classifier

Motivated by the efficacy of test-time linear probe in assessing representation quality, we aim to design a linear probing classifier in training to measure the discrimination of a

neural network and further leverage the probing signal to empower representation learning. We introduce an episodic linear probing (ELP) classifier and discuss its weight update scheme in training.

Detached Linear Probing Classifier in Training. When incorporating a linear probing classifier in training, we need to maintain its independence from the main classifier. While keeping the main classifier and the backbone network unchanged, we build a new episodic linear probing classifier on top of the feature extractor. We stop the linear probe classifier's gradient to back-propagate to the backbone network. This helps the linear probe not be biased by the main classifier and produce a neutral evaluation of the discrimination of the feature representations.

Formally, the episodic linear probing classifier is trained to classify the features into C categories using the same labels assigned to the main classifier,

$$\boldsymbol{p} = Cls_{\text{main}}(\boldsymbol{h}), \tag{4}$$

$$\boldsymbol{q} = Cls_{\text{elp}}(\text{stop-grad}(\boldsymbol{h})), \tag{5}$$

$$\ell_{\text{main}}(\boldsymbol{x}, \boldsymbol{y}) = \ell_{ce}(\boldsymbol{p}, \boldsymbol{y}), \tag{6}$$

$$\ell_{\text{elp}}(\boldsymbol{x}, \boldsymbol{y}) = \ell_{ce}(\boldsymbol{q}, \boldsymbol{y}). \tag{7}$$

Cls_{main} is the main classifier, and it produces a probability prediction of $\boldsymbol{p}$. Cls_{elp} is the linear probe classifier, and it generates a probability prediction of $\boldsymbol{q}$. Cls_{elp} is trained in an online manner, but its optimization is detached from the main branch. "stop-grad" indicates that feature $\boldsymbol{h}$ is detached to train Cls_{elp}. The gradients from the ELP classifier are unavailable to the backbone and main classifier, and vice versa. The main difference between the detached linear classifier and the test-time linear probe is that the features of the detached linear classifier are adaptively changed by the network, while the features of the test-time linear classifier are always fixed.

Episodic weight re-initialization overcomes overfitting. Training the detached linear classifier with the same number of epochs as the main classifier would lead to the detached linear classifier overfitting the features. This overfitting should be avoided because the simple linear probe is supposed to reflect the discrimination of the features. If the ELP classifier memorizes all samples, it would not be competent to evaluate the features effectively. To prevent the ELP classifier from overfitting the training data, we re-initialize its parameters episodically every $\mathcal{I}$ epochs where $\mathcal{I}$ indicates episodic re-initialization interval. Specifically, given a linear classifier parameterized with W and b, where W is the projection matrix, and b is the bias, both W and b are randomly re-initialized at the interval of $\mathcal{I}$ epochs.

The episodic linear probe enables us to measure and understand the feature discriminability throughout the training process. A larger value of $\mathcal{I}$ enforces the ELP classifier to be better trained, but it makes the ELP classifier more likely to be overfitted. In contrast, the ELP classifier is underfitted, if $\mathcal{I}$ is too small. An under-fitted ELP classifier may not well describe the generalization capabilities of the features. In practice, we set $\mathcal{I}$ as a hyper-parameter. Empirically, $\mathcal{I} = 2$ achieves consistent good probing performances across datasets.

3.2. The ELP-Suitable Regularization

ELP-Suitable Regularization through loss modulation. ELP assesses the features' separability in an online way. The standalone ELP is detached from the backbone and does not influence the main network. In this paper, we aim to utilize the prediction from the auxiliary ELP classifier to effectively improve the discriminability of the main branch. However, the design of this regularization is not straightforward. Considering the episodic nature of the ELP classifier, ELP's prediction is periodic and not as confident as the main classifier. If the regularization is not well constructed, the performance of the main branch would be severely impaired.

In this paper, we introduce a simple formulation that modulates the cross-entropy loss with an adaptive factor ϕ,

$$\mathcal{L}_{ELP-SR} = \sum_{i=1}^{B} \text{stop-grad}(\phi_i) * \ell_{ce}(\boldsymbol{p}_i, \boldsymbol{y}_i), \tag{8}$$

where $\boldsymbol{p}_i$ is the prediction probability from the main classifier, B is the batch size. The scalar factor ϕ_i is assigned to each instance to modulate its cross-entropy loss adaptively. ϕ measures the main network's suitability for an ELP classifier. If an instance is not *suitable* for the ELP classifier, *e.g.*, the instance may be not discriminative, or an out-of-distribution data point, ϕ imposes a relatively large value so that the network would pay more attention to this instance. Our ELP-Suitable Regularization (ELP-SR) effectively mitigates the intrinsic model bias and regularizes the network towards better linear separability.

We detach the gradients from ϕ so that the factor only influences the magnitude of the loss gradients, but the gradient orientation is not altered. This makes the optimization progress relatively easy and stable. The strategy works surprisingly well in practice.

The instantiation of the ELP-SR factor. As aforementioned, ϕ aims to measure the main network's suitability for an ELP classifier. Given an instance $\boldsymbol{x}$ with the label c, we instantiate the ELP-SR factor by considering the prediction score of the main classifier (p^c) and the prediction of the ELP classifier (q^c). We utilize two elements when we construct the regularization factor ϕ.

First, the distance metric (D) between the prediction of the ELP classifier and the prediction of the main classifier should be concerned. The distance should reflect the main classifier's confidence gap compared to the ELP classifier.

If the distance is minimized, the main classifier is pushed to act like a less-trained linear classifier. Relatively, The features would be remarkably discriminative if a less-trained classifier is already sufficient for recognizing, Therefore, this metric encourages the main classifier to become simpler, promoting the features to be more discriminative. We instantiate D by simply computing the ℓ_1 distance between p^c and q^c, *i.e.*, $D = |p^c - q^c|$.

Second, we incorporate a normalization metric (R) to reveal the discriminability of both the ELP classifier and the main classifier. The distance metric (D) measures the relative confidence gap, but we should also consider the absolute values of the confidence scores. If the distance between p^c and q^c is small, but both absolute scores are low, the network has not been well optimized to classify the instance. Thus, we should normalize the distance with a normalization metric. For simplicity, we set R as the average of p^c and q^c, *i.e.*, $R = (p^c + q^c)/2$.

We formulate the ELP-SR factor ϕ as,

$$\phi = (\frac{D}{R})^\gamma = (\frac{2|p^c - q^c|}{p^c + q^c})^\gamma, \tag{9}$$

where γ smoothly adjusts the rate between D and R. We empirically study other ELP-SR factor variants in the experiment section.

3.3. Training and inference

In the training phase, we calculate the softmax cross-entropy loss for both the main classifier and the ELP classifier. Our ELP-SR loss is summed with these losses. The overall training objective is below,

$$\mathcal{L} = \sum_{i=1}^{B} \ell_{\text{main}}(\boldsymbol{p}_i, \boldsymbol{y}_i) + \ell_{\text{elp}}(\boldsymbol{q}_i, \boldsymbol{y}_i) + \phi_i * \ell_{ce}(\boldsymbol{p}_i, \boldsymbol{y}_i) \tag{10}$$

In the test phase, we remove the auxiliary ELP classifier and only keep the main classifier. The final prediction is obtained only from the main classifier. Our framework does not introduce any additional overhead during testing.

4. Experiments

In the challenges of diverse objects of images in the wild, our method shows significant superiority for generalization. We evaluate three classification tasks, *i.e.*, fine-grained visual recognition, long-tailed recognition, and generic object recognition. First, since the classes in fine-grained recognition are similar, and samples are difficult to be recognized even by humans, the fine-grained recognition task brings extra challenges to learning discriminative features. Second, long-tailed recognition involves the extremely imbalanced distributions of data samples. This requests the method to possess generalization ability and recognize the tailed classes with limited samples. The evaluations of these tasks reveal the advantages of our method in improving visual representations.

We further evaluate our method on ImageNet-1K to study the generalization ability of ELP-SR. Besides the classification accuracy metric, we also report the results of a k-nearest-neighbor (KNN) classifier on the test set. This further manifests the effectiveness of our method in improving the discriminability of feature representations. Moreover, we provide ablation studies to compare different γ, $\mathcal{I}$, and formulations of the ELP-SR factor. To further demonstrate the ability of the ELP classifier, we present a comparison of the linear classifier's accuracy. The results reflect that the network with ELP-SR produces more discriminative and generalized features.

To be noticed, for all the tasks, we did **NOT** introduce any additional annotations nor incorporate extra parameters at the inference time. During testing, **only the backbone networks** are used to produce predictions.

4.1. Fine-grained Visual Recognition

Classes in fine-grained recognition are similar. They are difficult to distinguish, even for a human. Meanwhile, samples in every class are diverse [2]. Objects may be shown in various angles, illuminations, occlusions, backgrounds, etc. These induce fine-grained categories to show large intra-class variances, but small inter-class variances [2]. Samples in fine-grained classification are hard to be generalized and discriminated, which brings difficulties for learning discriminative features by networks.

Dataset and Implementation Details. To show the efficacy, we compare the performances on three standard benchmarks: CUB-200-2011 (CUB) [53], Stanford Cars (CAR) [28], and FGVC-Aircarft (AIR) [36].

Following the same training procedure in [10], we adapt ResNet-50 [19] pre-trained by ImageNet [30] as the backbone model. As the regular augmentations [10, 16, 65] in this task, resizing, random crops, rotations, and horizontal flips are applied. After operating these standard transformations, the final inputs become 448×448 resolutions. Similar to the ResNet50 baseline [10, 65], we train our method for 240 epochs and optimize the loss function by SGD. In our method, we report the results of $\gamma = 3$ for all three datasets with $D = p^c - q^c$ and $R = (p^c + q^c)/2$. For CUB, CAR, and AIR, we set $\mathcal{I} = 2$, 2, and 1, respectively. These are the best settings for parameters and will be discussed in the ablation section 4.4.

Experimental Results. As in Table 1, our method achieves significant improvements based on the ResNet50 baseline. Without bells and whistles, our results are competitive or even outperform many recent methods with complicated network designs [24], additional augmentations [10, 16], or

Method	Dataset		
	CUB	CAR	AIR
B-CNN [34]	84.1	91.3	84.1
HIHCA [6]	85.3	91.7	88.3
RA-CNN [17]	85.3	92.5	88.2
OPAM [38]	85.8	92.2	-
Kernel-Pooling [13]	84.7	91.1	85.7
MA-CNN [62]	86.5	92.8	89.9
MAMC [47]	86.5	93.0	-
HBP [58]	87.1	93.7	90.3
DFL-CNN [55]	87.4	93.1	91.7
NTS-Net [57]	87.5	93.9	91.4
DCL [10]	87.8	94.5	93.0
PMG [16]	88.9	95.0	92.8
ACNet [24]	88.1	94.6	92.5
LIO [65]	88.0	94.5	92.7
ResNet50 Baseline	85.5	92.7	90.3
ResNet50 Baseline + ELP-SR	88.8	94.2	92.7

Table 1. Comparison of three benchmarks of fine-grained classification. Without additional augmentations or network designs, our method achieves significant improvements.

Method	CIFAR-10			CIFAR-100		
Imbalance ratio	100	50	10	100	50	10
Focal Loss [32]	70.4	76.7	86.7	38.3	43.9	55.7
CB Focal [12]	74.6	79.3	87.1	39.6	45.2	58.0
Meta-weight [44]	75.2	80.0	87.8	42.0	46.7	58.4
CDB-CE [45]	-	-	-	42.5	46.7	58.7
Mixup [61]	73.1	77.8	88.3	39.6	45.0	58.2
ERM [7]	70.4	74.8	86.4	38.3	43.9	55.7
ERM [7] + ELP-SR	77.4	81.2	87.9	39.1	44.7	57.9
ERM [7] + ELP-SR ($\tau = 1$)	77.5	81.5	88.4	42.4	48.3	58.9
ERM [7] + ELP-SR ($\tau*$)	**78.0**	**81.5**	**88.7**	**42.4**	**48.3**	**59.1**
LDAM [7]	77.0	81.0	88.2	42.0	46.6	58.7
LDAM [7] + ELP-SR	**78.2**	**82.3**	**88.1**	**43.9**	**48.2**	**59.1**

Table 2. Comparison of top-1 validation accuracy of different methods on imbalanced CIFAR-10 and CIFAR-100 datasets. All results are implemented based on ResNet-32. $\tau = 1$ indicates applying τ-normalization [26] with $\tau = 1$. $\tau*$ stands for results with the best settings of τ.

multi-scale features [16, 65]. Merely utilizing naive backbone with ELP-SR in training, the simple backbone networks boost 3.3%, 1.5%, and 2.4% respectively in three datasets which are significant improvements in this task. Boosts in this task reveal that our method effectively improves the networks' ability to discriminate and generalize samples. To further manifest the superiority of our method, more discussions will be presented in 4.4.

4.2. Long-tailed Visual Recognition

In long-tail recognition, the data distributions of different classes show extreme imbalance. As the long-tailed distribution, a handful of 'head' classes contain considerable samples, but a large number of 'tail' classes only include limited samples. The networks are biased toward 'head' classes, and the samples in 'tail' classes are hard to be generalized. In this section, we also evaluate the performances of our method under the challenging long-tailed distribution.

Dataset and Implementation Details. The experiments are operated based on long-tailed CIFAR-10 and CIFAR-100 datasets [29]. We first produce several versions of long-tailed datasets following [7] under different imbalance ratios, which denotes the ratio between the largest and smallest numbers of samples in classes. We report the results in three kinds of imbalance ratios which are 100, 50, and 10, respectively. To perform fair comparisons, we evaluate our method based on the ResNet-32 baseline from [7].

Experimental Results. As shown in Table 2, ELP-SR dramatically improves the performances of the baseline method in all the settings and datasets. The improvements in CIFAR-10 of imbalance ratio 100 and 50 are even larger than LDAM [7]. Moreover, after adapting the normalization from [26], the results of our method show more competitiveness in this task. All results in different settings outperform LDAM.

Besides, we further investigate our method based on the LDAM [7]. By minimizing the margin-based boundary considering the generalization [7], LDAM is well-designed for long-tailed recognition and boosts the performances dramatically. Meanwhile, our method can achieve higher performances on the foundation of LDAM. Though without specific consideration for the long-tailed distribution, ELP-SR offers general improvements to this task. These results demonstrate that our method helps the network generalize and produce discriminative features against the challenging distributions.

4.3. Generic Visual Recognition on ImageNet

To reveal the generalization of ELP-SR, we further investigate our method in generic object recognition on the standard benchmark for visual representation.

Dataset and Implementation Details. We evaluate ELP-SR on ImageNet-1K [30], containing 1.28 million images with 1000 categories. To show the effectiveness and generalization, we apply ELP-SR on different backbone networks, which are ResNet-50 [19], ResNet-101 [19], ResNet-152 [19], BN-Inception [23], Inception-V3 [49], and Inception-ResNet-V2 [48]. According to the standard implementations of these works, we adapt SGD with momentum 0.9 as the optimizer. All the networks are trained with the augmentations of random crops and horizontal flips. For ResNet-50, ResNet-101, ResNet-152, and BN-inception, we first resize the images to 256×256 resolutions and then randomly crop them to 224×224. For Inception-V3 and Inception-ResNet-V2, we resize to 320×320 and

Backbone	Top-1 Accuracy		Top-5 Accuracy	
	Baseline	ELP-SR	Baseline	ELP-SR
ResNet50	76.13	76.82	92.86	93.32
ResNet101	77.37	77.86	93.54	94.06
ResNet152	78.31	78.77	94.04	94.42
BN-Inception	73.52†	74.05	91.56†	91.74
Inception-V3	77.45	78.12	93.56	94.04
Inception-ResNet-V2	79.63†	80.22	94.79†	95.24
SE-ResNet50	77.05	77.45	93.48	93.88
SE-ResNet101	77.62	77.94	93.93	94.38
SE-ResNet152	78.43	78.61	94.27	94.53

Table 3. Comparison of single-crop accuracy (%) on the ImageNet-1K validation set. Different backbones with our method show significant improvements. To perform a fair comparison, † indicates the results implemented and re-trained by ours.

randomly crop to 299×299 as the corresponding implementations in their works [48, 49]. As in Table 3, we report top-1 and top-5 accuracy respectively and compare all the backbones with ELP-SR.

Experimental Results. As in Table 3, with ELP-SR, all backbone networks achieve performance gains. The results reveal that our method is valuable to various backbone models and generally ameliorates the representations of networks. Almost all the backbones obtain about a 0.5% percent increase in top-1 accuracy.

Furthermore, to verify the general improvements introduced by our method, we explore the performances of our method with SE-block [22]. As shown in Table 3, though SE-block already promotes the performances, our method leads to further boosts on the fundamental of SE-block [22].

k-nearest neighbors accuracy. To reveal the effectiveness of our method, we provide an additional evaluation with the KNN classifier [56]. For feature vector h, we select the top k nearest neighbors by the weights $\exp(h \cdot h'/t)$ corresponding to the labels, where h' indicates features from the training set and t is a temperature term. We apply $t = 0.1$ in our experiments.

As shown in Table 4, the results with 20 and 200 nearest neighbors are displayed. With the KNN classifier, our method outperforms the backbone network. This reflects that the features after training with ELP-SR become more discriminative.

In all, the general improvements in all the backbones, methods, and tasks reflect that ELP-SR is not sensitive to particular networks, designs, or visual challenges. It provides a valuable regularization for visual representation learning.

4.4. Ablation Studies

4.4.1 Ablation on Hyper-parameters

Episodic interval $\mathcal{I}$. The number of periodical intervals prevents the ELP from overfitting the features. We experiment with the different values of $\mathcal{I}$ in the CUB dataset. As shown in Table 5, the performances are influenced by $\mathcal{I}$. The larger $\mathcal{I}$ induces the degradation of performances. With plenty of training iterations, the ELP classifier tends to be overfitting and cannot measure generalization effectively.

Method	20	200
ResNet50	75.04	73.21
ResNet50 + ELP-SR	**75.48**	**73.88**

Table 4. KNN accuracy on ImageNet-1K. Results of accuracy with 20 and 200 nearest neighbors are presented.

Besides, we also operate comparisons on the ImageNet dataset. The model achieves 76.13, 76.82, and 76.30 when $\mathcal{I}$ equals to 1, 2, and 3, respectively. The proper value of $\mathcal{I}$ can better empower the advantages of ELP. Minor $\mathcal{I}$ may not be sufficient for the construction of ELP. The more significant $\mathcal{I}$ may induce degradation of the ability of the ELP classifier to indicate features' discrimination. Thus, we apply $\mathcal{I} = 2$ in our experiments as this condition generally shows improvements in several datasets.

γ in the SR Factor. The parameter γ is responsible for adjusting the intensity of regularization. Since $\frac{D}{R}$ is always lower than 1, the larger γ leverages smaller regularization for the inputs. As shown in Table 5, we compare multiple conditions of γ in fine-grained classification. The variances of γ slightly influence the performances. A proper γ leads to better performances but is not deterministic for fine-grained classification. Moreover, we evaluate different γ values under the condition of $\mathcal{I} = 2$ on ImangeNet-1K. The recognition accuracies are 76.23, 76.82, and 76.30 when γ is set to 1, 2, and 3, respectively.

The Variations of SR Factor. We further investigate our ELP-SR in different forms, as shown in Table 6. First, for regularization, the confidences of the ELP classifier reflect the discriminability of features. Since the main classifier tends to be overfitting, p^c is relatively higher and close to 1. Thus, a similar effect may occur for $1 - q^c$ and $p^c - q^c$. As shown in Table 6, both formulations enable regularizing the networks to perform better while the model with $p^c - q^c$ achieves a higher result. This is because $p^c - q^c$ provides a more precise measurement of the deviation between the main classifier and the ELP classifier.

Second, to formulate the normalization term, we require both confidences of the ELP classifier and the main classifier to become higher. The higher confidence of the main classifier indicates that the sample can be correctly recognized. This is a primary requirement for better representation of the feature. If the features are hard to recognize even for the main classifier, this may indicate that the visual representation quality is relatively low. It is a primary criterion that the network should provide at least recognizable features. As shown in Table 6, higher performances are

shown if applying the normalization terms. Both p^c+q^c and $p^c * q^c$ are valid to normalize our ELP-SR. Third, only the regularization of higher q^c can also boost the performances. Without the normalization term, the impact of ELP-SR also guides the networks to be more generalized. However, lacking normalization, the improvements are relatively lower. Besides, simple normalization is also valuable. Since $\frac{2}{p^c+q^c}$ and $\frac{2}{p^c*q^c}$ also expect higher confidences of ELP, a similar influence may occur through leveraging the normalization term only. These results demonstrate that regularization and normalization are valuable in ELP-SR. Simultaneously, the combinations of both sides introduce a further increase in performances.

Finally, we also operate ablations for the distillation of the probability of two classifiers. Remarkable decreases are shown in Table 6 of both conditions for L1 and L2 regressions. The network should not be optimized to solve features' discriminability directly. Distilling can lead the main classifier to perform similarly to the ELP classifier but does not encourage the network to be more generalized. If the main classifier is optimal according to the ELP classifier, the network can 'pretend' to achieve discriminative features. However, in testing, this 'cheating' is useless. Additionally, we replace the ELP classifier with a memory bank and update the memory by a momentum-based moving average. When the momentum is 0.9 and 0.1, the results are $86.1\%$and 86.5%, respectively. The results show that the moving average operation helps fine-grained recognition, but it provides a weaker regularization than the episodically initialized ELP classifier.

4.4.2 Visualization

To demonstrate the efficacy of our ELP, we present a visualization for the testing accuracy of our ELP based on CUB. In detail, we train the baseline method, take the features from the backbone to train ELP, but do not leverage ELP-SR for network training. Meanwhile, we take our method training with ELP-SR as the comparison. This is similar to applying linear probing for every epoch. Since ELP is re-initialized every two epochs for CUB, to better reveal the capacity of ELP under different conditions, we plot the accuracy every two epochs. As shown in Fig. 3, unseen features in the testing set are remarkably more recognizable. This indicates that the network with ELP-SR is more generalized and produces more discriminative features. Even for the simple classifier, the unseen samples represented by the network are easier to be classified.

5. Conclusion

In this paper, we propose episodic linear probing (ELP) to estimate the generalization and discriminability of features online. By ELP, we propose an ELP-suitable Regularization term (ELP-SR) to regularize the models. Our insights are two-fold. 1). Since the main classifier may be overfitting and its confidence may not indicate the discrimination of features, the ELP classifier provides additional regularization for more discriminative features. 2). Immediate suitability is effective in measuring the discrimination of features. An intuitive hypothesis is that if the features are highly discriminative, they should be recognizable by an easily learned linear classifier. Our ELP is episodically re-initialized, effectively mitigating overfitting and regularizing the network towards better linear separability.

Parameter	$\mathcal{I}=1$	$\mathcal{I}=2$	$\mathcal{I}=3$	$\mathcal{I}=4$	$\mathcal{I}=5$
$\gamma=1$	88.0	88.2	88.2	88.0	87.8
$\gamma=2$	88.0	88.5	88.2	88.0	87.8
$\gamma=3$	87.6	**88.8**	88.0	88.0	87.6
$\gamma=4$	87.5	88.0	87.8	87.8	87.5

Table 5. Results for different values of $\mathcal{I}$ and γ on CUB. $\mathcal{I}$ prevents the ELP from overfitting, and γ adjusts the intensity of regularization.

Formulation	D	R	Top-1 Accuracy
$\frac{D}{R}$	p^c-q^c	p^c+q^c	**76.82**
	p^c-q^c	p^c*q^c	76.75
	$1-q^c$	p^c+q^c	76.78
	$1-q^c$	p^c*q^c	76.70
D	p^c-q^c	-	76.71
	$1-q^c$	-	76.60
$\frac{1}{R}$	-	p^c+q^c	76.25
	-	p^c*q^c	76.23
Distillation	L1		76.12
	L2		76.18

Table 6. Comparison for variations of SR Factor on ImageNet-1K. Various conditions are presented, including different formulations of D and R, with or without D and R, and direct distillation of the main and ELP classifier.

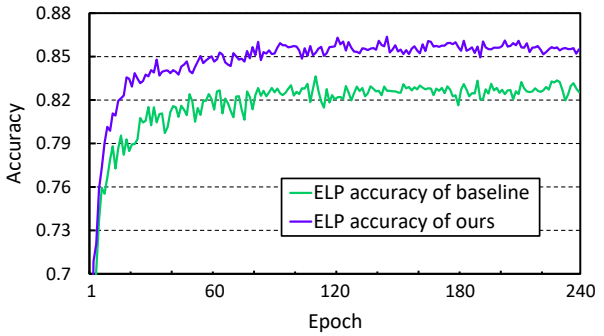

Figure 3. Curves of testing accuracy only with ELP classifier on CUB. Compared with our method, We utilize the baseline method that extracts the features from the backbone, trains ELP with features individually but does not leverage ELP-SR for the backbone training. Features trained with ELP-SR are more discriminative than the baseline and easier to be classified by simple ELP.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. 1, 2, 3

[2] Connor Anderson, Matt Gwilliam, Adam Teuscher, Andrew Merrill, and Ryan Farrell. Facing the hard problems in fgvc. *arXiv preprint arXiv:2006.13190*, 2020. 2, 5

[3] Yuki M Asano, Christian Rupprecht, and Andrea Vedaldi. A critical analysis of self-supervision, or what we can learn from a single image. *arXiv preprint arXiv:1904.13132*, 2019. 1

[4] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, 2012. 3

[5] Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. *arXiv preprint arXiv:1812.00332*, 2018. 3

[6] Sijia Cai, Wangmeng Zuo, and Lei Zhang. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 511–520, 2017. 6

[7] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, 2019. 6

[8] Arantxa Casanova, Michal Drozdzal, and Adriana Romero-Soriano. Generating unseen complex scenes: are we there yet? *arXiv preprint arXiv:2012.04027*, 2020. 1

[9] D. Chang, Y. Ding, J. Xie, A. K. Bhunia, X. Li, Z. Ma, M. Wu, J. Guo, and Y. Song. The devil is in the channels: Mutual-channel loss for fine-grained image classification. *IEEE Transactions on Image Processing*, 29:4683–4695, 2020. 3

[10] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 5, 6

[11] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. 2

[12] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 6

[13] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge Belongie. Kernel pooling for convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2930, 2017. 6

[14] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2970–2979, 2017. 3

[15] Amit Dhurandhar, Karthikeyan Shanmugam, Ronny Luss, and Peder Olsen. Improving simple models with confidence profiles. *arXiv preprint arXiv:1807.07506*, 2018. 3

[16] Ruoyi Du, Dongliang Chang, Ayan Kumar Bhunia, Jiyang Xie, Yi-Zhe Song, Zhanyu Ma, and Jun Guo. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *European Conference on Computer Vision*, 2020. 5, 6

[17] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4438–4446, 2017. 6

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 3

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016. 3, 5, 6

[20] Robert Hecht-Nielsen. Theory of the backpropagation neural network. In *Neural networks for perception*, pages 65–93. Elsevier, 1992. 3

[21] Evan Hernandez and Jacob Andreas. The low-dimensional linear geometry of contextualized word representations. *arXiv preprint arXiv:2105.07109*, 2021. 3

[22] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3, 7

[23] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 6

[24] Ruyi Ji, Longyin Wen, Libo Zhang, Dawei Du, Yanjun Wu, Chen Zhao, Xianglong Liu, and Feiyue Huang. Attention convolutional binary neural tree for fine-grained visual categorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10468–10477, 2020. 5, 6

[25] Lu Jiang, Deyu Meng, Shoou-I Yu, Zhenzhong Lan, Shiguang Shan, and Alexander Hauptmann. Self-paced learning with diversity. *Advances in Neural Information Processing Systems*, 27:2078–2086, 2014. 3

[26] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020. 6

[27] Ivan Krasin, Tom Duerig, Neil Alldrin, Andreas Veit, Sami Abu-El-Haija, Serge Belongie, David Cai, Zheyun Feng, Vittorio Ferrari, Victor Gomes, Abhinav Gupta, Dhyanesh Narayanan, Chen Sun, Gal Chechik, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2016. 1

[28] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 5

[29] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2, 6

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 1, 3, 5, 6

[31] Buyu Li, Yu Liu, and Xiaogang Wang. Gradient harmonized single-stage detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8577–8584, 2019. 3

[32] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3, 6

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[34] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015. 6

[35] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[36] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 5

[37] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014. 3

[38] Yuxin Peng, Xiangteng He, and Junjie Zhao. Object-part attention model for fine-grained image classification. *IEEE Transactions on Image Processing*, 27(3):1487–1500, 2017. 6

[39] David E Rumelhart, Richard Durbin, Richard Golden, and Yves Chauvin. Backpropagation: The basic theory. *Backpropagation: Theory, architectures and applications*, pages 1–34, 1995. 3

[40] Mohammad Sabokrou, Mohammad Khalooei, and Ehsan Adeli. Self-supervised representation learning via neighborhood-relational encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8010–8019, 2019. 2

[41] Vikash Sehwag, Mung Chiang, and Prateek Mittal. On separability of self-supervised representations. *ICML workshop on Uncertainty and Robustness in Deep Learning (UDL)*, 2020. 2, 3

[42] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1

[43] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019. 2

[44] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. *arXiv preprint arXiv:1902.07379*, 2019. 3, 6

[45] Saptarshi Sinha, Hiroki Ohashi, and Katsuyuki Nakamura. Class-wise difficulty-balanced loss for solving class-imbalance. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 6

[46] Cory Stephenson, Suchismita Padhy, Abhinav Ganesh, Yue Hui, Hanlin Tang, and SueYeon Chung. On the geometry of generalization and memorization in deep neural networks. *arXiv preprint arXiv:2105.14602*, 2021. 1

[47] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 805–821, 2018. 6

[48] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016. 3, 6, 7

[49] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 6, 7

[50] Martin A Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540, 1987. 2

[51] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 1

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[53] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 5

[54] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. *arXiv preprint arXiv:2112.01529*, 2021. 3

[55] Yaming Wang, Vlad I Morariu, and Larry S Davis. Learning a discriminative filter bank within a cnn for fine-grained recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4148–4157, 2018. 6

[56] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 2, 7

[57] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 420–435, 2018. 6

[58] Chaojian Yu, Xinyi Zhao, Qi Zheng, Peng Zhang, and Xinge You. Hierarchical bilinear pooling for fine-grained visual recognition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 574–589, 2018. 6

[59] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 1

[60] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016. 1, 2

[61] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 6

[62] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 5209–5217, 2017. 6

[63] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE international conference on computer vision*, pages 3754–3762, 2017. 2

[64] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020. 2

[65] Mohan Zhou, Yalong Bai, Wei Zhang, Tiejun Zhao, and Tao Mei. Look-into-object: Self-supervised structure modeling for object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 5, 6