DeiT-LT: Distillation Strikes Back for Vision Transformer Training on Long-Tailed Datasets

Harsh Rangwani*[1] Pradipto Mondal*[1,2] Mayank Mishra*[1]
Ashish Ramayee Asokan[1] R. Venkatesh Babu[1]

[1]Indian Institute of Science, Bangalore [2] Indian Institute of Technology, Kharagpur

Abstract

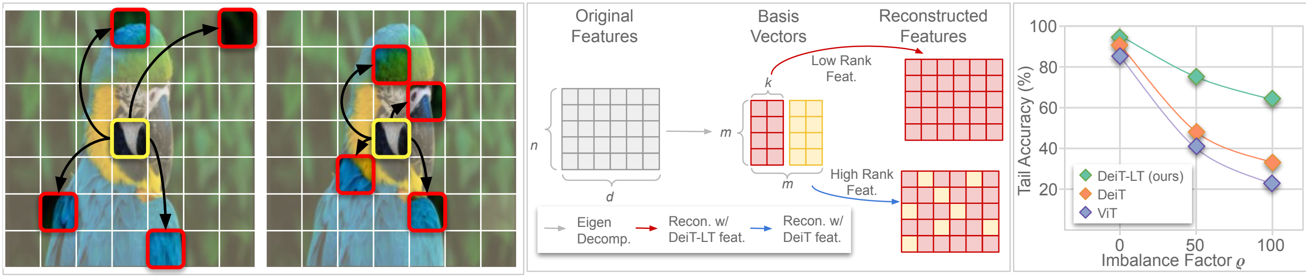
Vision Transformer (ViT) has emerged as a prominent architecture for various computer vision tasks. In ViT, we divide the input image into patch tokens and process them through a stack of self-attention blocks. However, unlike Convolutional Neural Network (CNN), ViT's simple architecture has no informative inductive bias (e.g., locality, etc.). Due to this, ViT requires a large amount of data for pre-training. Various data-efficient approaches (DeiT) have been proposed to train ViT on balanced datasets effectively. However, limited literature discusses the use of ViT for datasets with long-tailed imbalances. In this work, we introduce DeiT-LT to tackle the problem of training ViTs from scratch on long-tailed datasets. In DeiT-LT, we introduce an efficient and effective way of distillation from CNN via distillation `DIST` token by using out-of-distribution images and re-weighting the distillation loss to enhance focus on tail classes. This leads to the learning of local CNN-like features in early ViT blocks, improving generalization for tail classes. Further, to mitigate overfitting, we propose distilling from a flat CNN teacher, which leads to learning low-rank generalizable features for `DIST` tokens across all ViT blocks. With the proposed DeiT-LT scheme, the distillation `DIST` token becomes an expert on the tail classes, and the classifier `CLS` token becomes an expert on the head classes. The experts help to effectively learn features corresponding to both the majority and minority classes using a distinct set of tokens within the same ViT architecture. We show the effectiveness of DeiT-LT for training ViT from scratch on datasets ranging from small-scale CIFAR-10 LT to large-scale iNaturalist-2018. Project Page: https://rangwani-harsh.github.io/DeiT-LT.

1. Introduction

Visual Recognition has seen unprecedented success with the advent of deep neural networks trained on large datasets [10]. Consequently, efforts are being made to collect large datasets through crowd-sourcing to train deep neural networks for various applications across domains. As a result of crowd-sourcing, these datasets often exhibit long-tailed data distributions due to inherent natural statistics [14, 52], i.e., a large number of images belong to a small portion of (*majority*) classes, whereas other (*minority*) classes contain few image samples each. A lot of recent works [5, 9, 25, 32, 67] focus on training deep neural networks for recognition on such long-tailed datasets, such that networks perform reasonably well across all classes, including the minority classes. Loss manipulation-based techniques [5, 9, 23] enhance the network's focus toward learning tail classes by enforcing a large margin or increasing the weight for loss for these classes. As these techniques enhance the focus on the tail classes, they often lead to some performance degradation in the head (majority) classes. To mitigate this, State-of-the-Art (SotA) techniques currently train multiple expert networks [25, 56] that specialize in different portions of the data distribution. The predictions from these experts are then aggregated to produce the final output, which improves the performance over individual experts. However, all these efforts have been restricted to Convolutional Neural Networks (CNNs), particularly ResNets [15], with little attention to architectures such as Transformers [11, 53], MLP-Mixers [47] etc.

Recently, the transformer architecture adapted for computer vision, named as Vision Transformer (ViT) [12], has gained popularity due to its scalability and impressive performance on various computer vision tasks [6, 44]. One caveat behind its impressive performance is the requirement for pre-training on large datasets [11]. The data-efficient transformers (DeiT) [48] aimed to reduce this requirement for pre-training by distilling information from a pre-trained CNN. Subsequent efforts have further improved the data and compute efficiency [50, 51] of ViTs. However, all these improvements have been primarily based on increasing performance on the balanced ImageNet dataset. We find that these improvements are still insufficient for robust perfor-

* denotes equal contribution. Correspondence to harshr@iisc.ac.in.



(a) Comparison b/w attention maps of DeiT (left) and DeiT-LT (ours, right)

(b) DeiT-LT (ours) learns low rank feat. compared to the DeiT baseline

(c) Tail acc. for DeiT-LT (ours) vs baselines

Figure 1. We propose DeiT-LT (Fig. 2, a distillation scheme for Vision Transformer (ViT), tailored towards long-tailed data). In Deit-LT, **a)** we introduce OOD distillation from CNN, which leads to learning local generalizable features in early blocks. **b)** we propose to distill from teachers trained via SAM [13] which induces low-rank features across blocks in ViT to improve generalization. **c)** In comparison to other SotA ViT baselines, Deit-LT (ours) demonstrates significantly improved performance for minority classes, with increasing imbalance.

mance on long-tailed datasets (Fig. 1c).

In this work, we aim to investigate and improve the *training of Vision Transformers from scratch without the need for large-scale pre-training* on diverse long-tailed datasets, varying in image size and resolution. Recent works show improved performance for ViTs on long-tailed recognition tasks, but they often need expensive pre-training on large-scale datasets [7, 30]. The requirement of pre-training is computationally expensive and restricts their application to specialized domains such as medicine, satellite, speech, etc. Furthermore, the large-scale pre-trained datasets often contain biases that might be inadvertently induced with their usage [2, 34, 54]. To mitigate these shortcomings, we introduce ***Data-efficient Image Transformers for Long-Tailed Data (DeiT-LT)*** - a scheme for training ViTs from scratch on small and large-scale long tailed datasets. DeiT-LT is based on the following important design principles:

- DeiT-LT involves distilling knowledge from low-resolution teacher networks using out-of-distribution (OOD) images generated through strong augmentations. Notably, this method proves effective even if the CNN teacher wasn't originally trained on such augmentations. The outcome is the successful induction of CNN-like feature locality in the ViT student network, ultimately enhancing generalization performance, particularly for minority (tail) classes (Fig. 1a, 4a and Sec. 3.1).
- Further, to improve the generality of features, we propose to distill knowledge via flat CNN teachers trained through Sharpness Aware Minimization (SAM) [13]. This results in low-rank generalizable features for long-tailed setup across all ViT blocks (Fig. 1b and Sec. 3.2).
- In DeiT [48], the classification and distillation tokens produce similar predictions. However, in proposed DeiT-LT, we ensure their divergence such that the classification token becomes an expert on the majority classes. Whearas, the distillation token learns local low-rank features, becoming an expert on the minority. Hence, DeiT-LT can focus on both the majority and minority effectively, which is not possible with vanilla DeiT training (Fig. 5 and Sec. 3.1).

We demonstrate the effectiveness of DeiT-LT across diverse small-scale (CIFAR-10 LT, CIFAR-100 LT) as well as large-scale datasets (ImageNet-LT, iNaturalist-2018). We find that DeiT-LT effectively improves over the teacher CNN across all datasets and achieves performances superior to SotA CNN-based methods without requiring any pre-training.

2. Background

Long-Tailed Learning. With the increased scale of deep learning, large crowd-sourced long-tailed datasets have become common. A plethora of techniques are developed to learn machine learning models using such datasets, where the objective is improved performance, particularly on tail classes. The methods can be broadly divided into three categories: a) loss re-weighting b) decoupled classifier and representations and c) expert-based classifier training. In addition, there are some techniques based on the synthetic generation for long-tailed recognition [22, 37, 39, 40], which are orthogonal to this study. The loss re-weighting-based techniques include margin based techniques like LDAM [5], and Logit-Adj [32], which enforce a higher margin for tail classes. The other set (eg. CB-Loss [9], VS-Loss [23] etc.) introduce re-weighting factors in cross entropy loss based on the training set label distribution. The other set of techniques propose to decouple the learning of representations with classifier learning, as it's observed that margin based losses lead to sub-optimal representations [20]. The classifier is then learned using Learnable Weight Scaling (LWS), τ-normalization, which improves performance on the tail classes [20]. Further, after this follow-up works [55, 60] like MiSLAS [66] proposed Mixup [62] based improved representation learning and LADE [16] proposes improved classifier training by adapting to target label distribution. Further, contrastive methods, including PaCo [8] and BCL [41], have demonstrated improved performance with contrastive

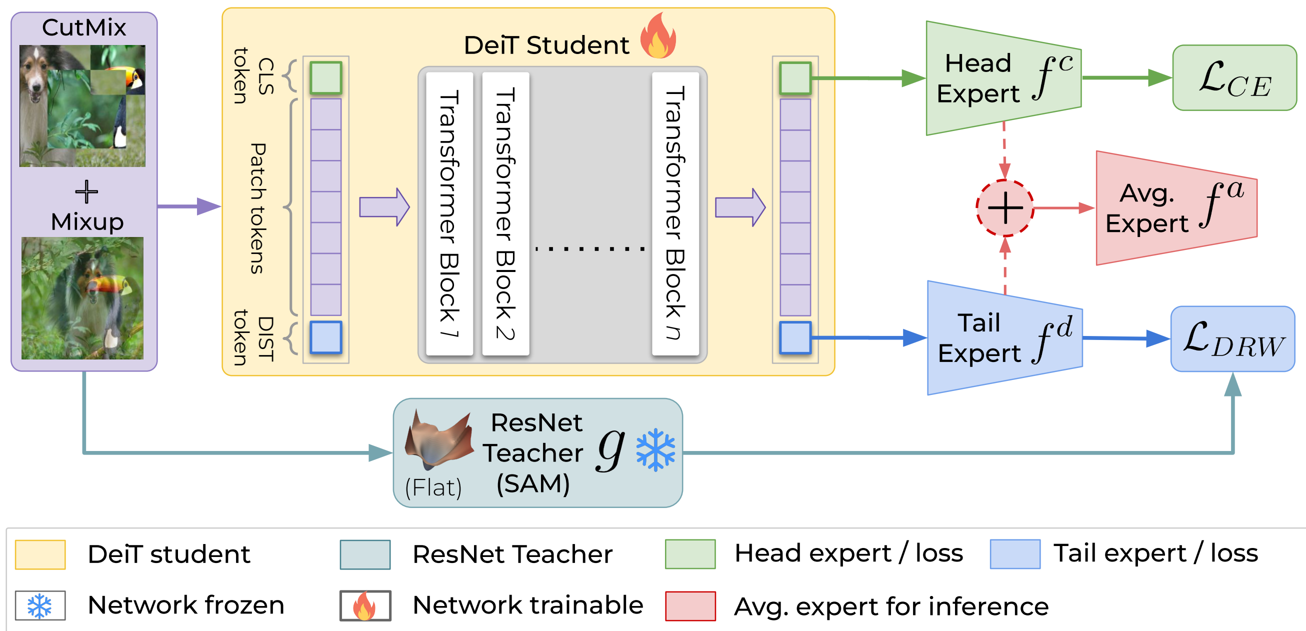


Figure 2. Overview of DeiT-LT. The Head Expert classifier trains using CE loss against ground truth, whereas the Tail Expert classifier trains using DRW loss against hard-distillation targets from the flat ResNet teacher trained via SAM [13]. The distillation is performed using out-of-distribution images created using strong augmentations and Mixup.

learning. However, all these methods lead to performance degradation on head classes to improve performance on tail classes. To mitigate this degradation, the techniques (like RIDE [56] etc.) learn different experts on different parts of the data distribution. These experts are learned in a way that makes them diverse in their predictions and can be combined efficiently to obtain improved predictions. However, these methods require additional computation to combine experts at the inference time. In our work, we can efficiently learn experts on majority and minority using a single ViT backbone, the predictions of which we average to prevent any additional inference overhead at the deployment time.

Vision Transformer. In recent literature, Vision Transformers [12] have emerged as strong competitors for ResNets as they are easier to scale and lead to improved generalization. DeiT [48] developed a data-efficient way to train these models by distilling through Convolutional Neural Networks. However, despite being data efficient, these models still produce sub-optimal performance on long-tailed data. RAC [30] utilizes pre-trained transformer for data-efficiency on long-tailed data. However, these pre-trained models are often domain specific and do not generalize well to other domains like medical, synthetic etc. In our work, we train Vision Transformers from scratch, even for small datasets like CIFAR-10 LT, CIFAR-100 LT, which makes them free from biases due to pre-training on large datasets [54].

Data Efficient Vision Transformers (DeiT). The Vision Transformer (ViT) architecture [12] consists of transformer architecture stacked with Multi-Headed Self-Attention blocks [53]. To provide input to the Vision Transformer architecture, we first convert the image into patches. These image patches are passed through a linear layer to convert them into tokens that are then passed to the attention blocks. The attention blocks learn the relationship between these tokens for performing a given task. In addition to this, the ViT architecture also contains one classifier (`CLS`) token that represents the features to be used for classification. In the Data Efficient Transformer (DeiT) [48], there is an additional distillation (`DIST`) token in the ViT backbone that learns via distillation from the teacher CNN. For the classification head and the distillation head, $\mathcal{L}_{CE}$ is used for training (Fig. 2). The final loss function for the network is:

$$\mathcal{L} = \mathcal{L}_{CE}(f^c(x), y) + \mathcal{L}_{CE}(f^d(x), y_t), y_t = \arg\max_i g(x)_i \tag{1}$$

Here $f^c(x)$ is output from the classifier of student `CLS` token, $f^d(x)$ is output from the classifier of student `DIST` token, $g(x)$ denotes the output of the teacher CNN network, $y \in [K]$ is the ground truth, y_t is the label produced by the teacher corresponding to the sample x, and N_i is the number of samples in class i. At the time of inference in DeiT, we obtain logit outputs from the two heads $f^d(x)$ and $f^c(x)$, and average them to produce the final prediction.

3. DeiT-LT (DeiT for Long-Tailed Data)

In this section, we introduce DeiT-LT - the Data-efficient Image Transformer that is specialized to be effective for

Table 1. **Effect of augmentations:** Comparison of teacher (*Tch*) and student (*Stu*) accuracy (%) and training time (in hours) on CIFAR-10 LT ($\rho = 100$) using various augmentation strategies with mixup (✓) and without mixup (✗). Despite low teacher training accuracy on the out-of-distribution images, the student (Stu.) performs better on the validation set.

Tch Model	Stu Augs.	Tch Augs.	Tch Acc.	Stu Acc.	Train Time
RegNetY 16GF	Strong (✓)	Strong (✓)	79.1	70.2	33.3
ResNet-32	Strong (✗)	Weak (✗)	97.2	54.2	17.8
	Strong (✗)	Strong (✗)	71.9	69.6	17.8
	Strong (✓)	Strong (✓)	56.6	79.4	19.0

Long-Tailed data. We start with a DeiT transformer-based architecture which, in addition to the classification (`CLS`) token, also contains a distillation (`DIST`) token (Fig. 2) that learns via distillation from a CNN. The DeiT-LT introduces three particular design components, which are: **a)** the effective distillation via out-of-distribution (OOD) images, which induces local features and leads to the creation of experts, **b)** training Tail Expert classifier using DRW loss and **c)** learning of low-rank generalizable features from flat teachers via distillation. In the following sections, we analyze our design choices in detail. We analyze CIFAR-10 LT using LDAM+DRW+SAM ResNet-32 [38] CNN teacher, to justify the rationale behind each design component.

3.1. Distillation via Out of Distribution Images

We now focus on how to distill knowledge from a CNN architecture to a ViT effectively. In the original DeiT work [48], the authors first train a large CNN, specifically RegNetY [35], with strong augmentations ($\mathcal{A}$) as used by a ViT for distillation. However, this incurs the additional expense of training a large CNN for subsequent training of the ViT through distillation. In contrast, we propose to train a small teacher CNN (ResNet-32) with the usual weak augmentations, but during distillation, we pass strongly augmented images to obtain predictions to be distilled.

These strongly augmented images are *out-of-distribution (OOD)* images for the ResNet-32 CNN as the model's accuracy on these training images is low, as seen in Table 1. However, despite the low accuracy, the strong augmentations lead to effective distillation in comparison to the weak augmentations on which the original ResNet was trained (Table 1). This works because the ViT student learns to mimic the incorrect predictions of the CNN teacher on the out-of-distribution images, which in turn enables the student to learn the inductive biases of the teacher.

$$f^d(X) \approx g(X), X \sim A(x) \tag{2}$$

Further, we find that creating additional out-of-distribution samples by mixing up images from two classes [61, 62]

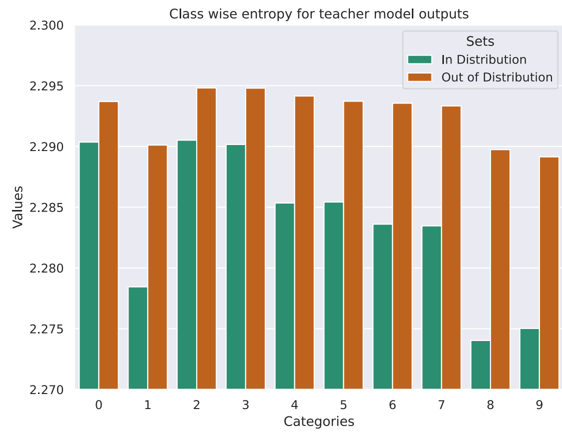


Figure 3. **Entropy of teacher outputs:** Comparison of the entropy of in-distribution samples and out-of-distribution samples with the ResNet-32 teacher on CIFAR-10 LT. We observe a higher accuracy in Table-1 corresponding to out-of-distribution samples.

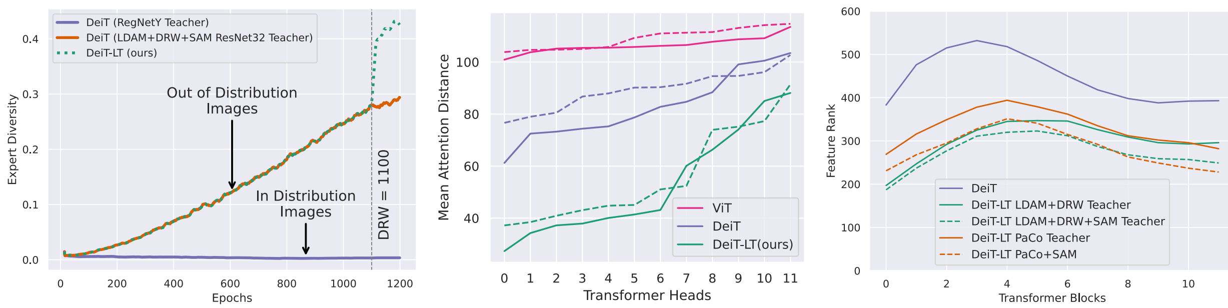
improves the distillation performance. This can also be seen from the entropy of predictions on teacher, which are high (*i.e.* more informative) for OOD samples (Fig. 3). *In general, we find that increasing diverse amount of out-of-distribution [33] data while distillation helps improve performance and leads to effective distillation from the CNN.* Details regarding the augmentations are in Suppl. Sec. A.4.

Due to distillation via out-of-distribution images, the teacher predictions y_t often differ from the ground truth y. Hence, the classification token (`CLS`) and distillation token (`DIST`) representations diverge while training. This phenomenon can be observed in Fig. 4a, where the cosine distance between the representation of the `CLS` and `DIST` tokens increases as the training progresses. This leads to the `CLS` token being an expert on head classes, while the `DIST` token specializes in tail class predictions. Our observation debunks the *myth that it is required for the* `CLS` *token predictions to be similar to* `DIST` for effective distillation in transformer, as observed by Touvron et al. [48].

Tail Expert with DRW loss. Further in this stage, we also introduce Deferred Re-Weighting (DRW) [5] for distillation loss, where we weigh the loss for each class using a factor $w_y = 1/\{1 + (e_y - 1)\mathbb{1}_{\text{epoch} \geq \text{K}}\}$, where $e_y = \frac{1-\beta^{N_y}}{1-\beta}$ is the effective number of samples in class y [9], after K number of epochs [5]. Hence the overall loss is given as:

$$\mathcal{L} = \frac{1}{2}\mathcal{L}_{CE}(f^c(x), y) + \frac{1}{2}\mathcal{L}_{DRW}(f^d(x), y_t),$$
$$\text{where } \mathcal{L}_{DRW} = -w_{y_t}\, log(f^d(x)_{y_t})$$

The DRW stage further enhances the focus of the distillation head (`DIST`) on the tail classes, leading to improved performance. This is also observed in Fig. 4a, where the diversity between the two tokens improves after the introduction of



(a) Diversity for `CLS` and `DIST` experts (b) Locality of Attention Heads (c) Rank of ViT from Distillation of CNN teachers.

Figure 4. Effect of Distillation in DeiT-LT. In **a)** we train DeiT-B with teachers trained on in-distribution images (RegNetY-16GF) and out-of-distribution images (ResNet32). The out-of-distribution distillation leads to diverse experts, which become more diverse with deferred re-weighting on the distillation token (DRW). In **b)** we plot the *Mean Attention Distance* for the patches across the early self attention block 1 (solid) and block 2 (dashed) for baselines, where we find that DeiT-LT leads to highly local and generalizable features. In **c)** we show the rank of features for `DIST` token, where we demonstrate that students trained with SAM are more low-rank in comparison to baselines

the DRW stage. This leads to the creation of diverse `CLS` and `DIST` tokens, which are experts on the majority and minority classes respectively.

Induction of Local Features: To gain insights into the generality and effectiveness of OOD Distillation, we take a closer look at the tail features produced by DeiT-LT. In Fig. 4b, we plot the mean attention distance for each patch across ViT heads [36] (Details in Suppl. Sec. F).

Insight 1: DeiT-LT contains heads that attend locally, like CNN, in the neighborhood of the patch in early blocks (1,2).

Due to this learning of local generalizable class agnostic features, we observe improved generalization on minority classes (Fig. 1c). Without the OOD distillation, we find that the vanilla DeiT-III and ViT baselines overfit only on the spurious global features (Fig. 4b) and do not generalize well for tail classes. Hence, this makes OOD distillation in DeiT-LT a well-suitable method for long-tailed scenarios.

3.2. Low-Rank Features via SAM teachers

To further improve the generalizability of the features, particularly for classes with less data, we propose to distill via *teacher CNN models trained via Sharpness Aware Minimization (SAM) objective* [13]. Models trained via SAM converge to flat minima [38] and lead to low-rank features [3]. For analyzing the rank of features for the ViT student in LT case, we calculate rank specifically for the features of tail classes [3]. We detail the procedure of our rank calculation in Suppl. Sec. G. We confirm our observations across diverse teacher models trained via LDAM and PaCo. We find the following insight for distillation via `DIST` token:

Insight 2. We observe that distilling into ViT via predictions made using SAM teacher leads to low-rank generalizable (`DIST`) token features across blocks of ViT (Fig. 4c).

This transfer of a CNN teacher's characteristic (low-rank) to the student, by just distilling via final logits, is a significant novel finding in the context of distillation for ViTs.

Training Time. In the original DeiT formulation, the authors [51] propose training a large CNN RegNetY-16GF at a high resolution (224×224) for distillation to the ViT. We find that competitive performance can be achieved even with training a smaller ResNet-32 CNN (32×32) at a lower resolution, as seen in Table 1. This significantly reduces compute requirement and overall training time by 13 hours, as the ResNet-32 model can be trained quickly (Table 1). Further, we find that with SAM teachers, the student converges much faster than vanilla teacher models, demonstrating the efficacy of SAM teachers for low-rank distillation (Suppl. Sec. G.1).

4. Experiments

4.1. Datasets

We analyze the performance of our proposed method on four datasets, namely **CIFAR-10 LT**, **CIFAR-100 LT**, **ImageNet-LT**, and **iNaturalist-2018**. We follow [5] to create long-tailed versions of CIFAR [24] datasets, where the number of samples is exponentially decayed using an imbalance factor $\rho = \frac{\max_i N_i}{\min_j N_j}$ (number of samples in the most frequent class by that in the least frequent class). For ImageNet-LT, we create an imbalanced version of the ImageNet [42] dataset as described in [29]. We also report performance on iNaturalist-2018 [52], a real-world long-tailed dataset. We divide the classes into three subcategories: **Head** (*Many*), **Mid** (*Medium*), and **Tail** (*Few*) classes. More details regarding the datasets can be found in Suppl. Sec. A.1.

4.2. Experimental Setup

We follow the setup mentioned in DeiT [48] to create the student backbone for our experiments. We use the DeiT-B student backbone architecture for all the datasets. We train our teacher models using re-weighting based LDAM-DRW-

Table 2. Results on CIFAR-10 LT and CIFAR-100 LT datasets with ρ=50 and ρ=100. We report the *overall* accuracy for available methods. (The teacher used to train the respective student (DeiT-LT) model can be identified by matching superscripts)

Method	CIFAR-10 LT		CIFAR-100 LT	
	ρ = 100	ρ = 50	ρ = 100	ρ = 50
ResNet32 Backbone				
CB Focal loss [9]	74.6	79.3	38.3	46.2
LDAM+DRW [5]	77.0	79.3	42.0	45.1
LDAM+DAP [19]	80.0	82.2	44.1	49.2
BBN [67]	79.8	82.2	39.4	47.0
CAM [64]	80.0	83.6	47.8	51.7
Log. Adj. [32]	77.7	-	43.9	-
RIDE [56]	-	-	49.1	-
MiSLAS [65]	82.1	85.7	47.0	52.3
Hybrid-SC [55]	81.4	85.4	46.7	51.9
SSD [27]	-	-	46.0	50.5
ACE [4]	81.4	84.9	49.6	51.9
GCL [26]	82.7	85.5	48.7	53.6
VS [23]	78.6	-	41.7	
VS+SAM [38]	82.4	-	46.6	-
[1]L-D-SAM [38]	81.9	84.8	45.4	49.4
[2]PaCo+SAM[8, 38]	86.8	88.6	52.8	56.6
ViT-B Backbone				
ViT [12]	62.6	70.1	35.0	39.0
ViT (cRT) [20]	68.9	74.5	38.9	42.2
DeiT [48]	70.2	77.5	31.3	39.1
DeiT-III [51]	59.1	68.2	38.1	44.1
[1]DeiT-LT(ours)	84.8	87.5	52.0	54.1
[2]DeiT-LT(ours)	**87.5**	**89.8**	**55.6**	**60.5**

SAM method [38] and the contrastive PaCo+SAM (training PaCo [8] with SAM [13] optimizer), employing ResNet-32 for small scale datasets (CIFAR-10 LT and CIFAR-100 LT) and ResNet-50 for large scale ImageNet-LT, and iNaturalist-2018. We train the head expert classifier with CE loss $\mathcal{L}_{CE}$ against the ground truth, while the tail expert classifier is trained with the CE+DRW loss $\mathcal{L}_{DRW}$ against the hard-distillation targets from the teacher network.

Small scale CIFAR-10 LT and CIFAR-100 LT. These models are trained for 1200 epochs, where DRW training for the Tail Expert Classifier starts from epoch 1100. Except for the DRW training (last 100 epochs), we use Mixup and Cutmix augmentation for the input images. These datasets are trained with a cosine learning rate schedule with a base LR of 5×10^{-4} using the AdamW [31] optimizer.

Large scale ImageNet-LT and iNaturalist-2018. These models are trained for 1400 and 1000 epochs, respectively, with the DRW training for the Tail Expert Classifier starting from 1200 and 900 epochs. We use Mixup and Cutmix

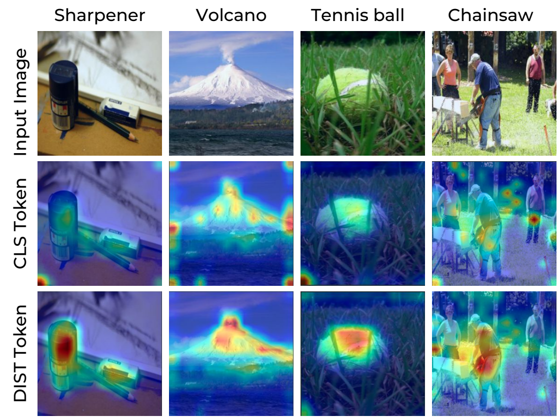


Figure 5. Visual comparison of the attention maps with respect to the `CLS` and `DIST` tokens for *tail* images from the ImageNet-LT dataset. The attention maps are computed by *Attention Rollout* [1].

throughout training. Both datasets follow a cosine learning rate schedule, with a base LR of 5×10^{-4}. More details on the experimental process can be found in Suppl. Sec A.

Baselines. We use the popular data-efficient baselines for ViT: **a) ViT:** The standard Vision Transformer (**ViT-B**)[12] architecture trained with CE Loss against the ground truth. For a fair comparison, we train ViT with the same augmentation strategy used for the DeiT-LT experiments. **b) DeiT [48]**: Vanilla DeiT model that uses RegNetY-16GF teacher trained with in-distribution images for distillation. **c) DeiT-III:** A recent improved version of DeiT ([51]) that focuses on improving the supervised learning of ViT on balanced datasets using three simple augmentations (GrayScale, Solarisation, and Gaussian Blur) and LayerScale [49], also demonstrating the redundancy of distillation in DeiTs. The long-tailed baseline of **d) ViT (cRT):** a decoupled approach of first training classifier (ViT) and then re-training the classifier for a small number of epochs with class-balanced sampling [20]. We further attempted training other baselines like LDAM, etc, on ViT. However, we found some optimization difficulties in training ViTs (details in Suppl. Sec. A.3).

We want to convey that we do not compare against baselines [30, 46, 58, 59], which use pre-training, usually on large datasets, to produce results on even CIFAR datasets (Ref. Suppl. Sec. C). Our goal is to develop a generic technique for training ViTs across domains and modalities on long-tailed data without requiring any external supervision.

5. Results

In this section, we present results for DeiT-LT across various datasets. We use re-weighting based LDAM+DRW+SAM (referred to as L-D-SAM in Table 2,3,4) and contrastive PaCo+SAM teachers for training DeiT-LT student models.

Results on Small Scale Datasets. Table 2 presents results for the CIFAR-10 LT and CIFAR-100 LT datasets, with varying imbalance factors (ρ = 100 and ρ = 50). We

Table 3. Results on ImageNet-LT. (The teacher used to train respective student (DeiT-LT) can be identified by matching superscripts)

Method	ImageNet-LT			
	Overall	Head	Mid	Tail
ResNet50 Backbone				
CB Focal loss [9]	33.2	39.6	32.7	16.8
LDAM [5]	49.8	60.4	46.9	30.7
c-RT [20]	49.6	61.8	46.2	27.3
τ-Norm [21]	49.4	59.1	46.9	30.7
Log. Adj. [32]	50.1	61.1	47.5	27.6
RIDE(3 exps) [56]	54.9	66.2	51.7	34.9
MiSLAS [65]	52.7	62.9	50.7	34.3
Disalign [63]	52.9	61.3	52.2	31.4
TSC [28]	52.4	63.5	49.7	30.4
GCL [26]	54.5	63.0	52.7	37.1
SAFA [17]	53.1	63.8	49.9	33.4
BCL [41]	57.1	67.9	54.2	36.6
ImbSAM [68]	55.3	63.2	53.7	38.3
CBD$_{ENS}$ [18]	55.6	68.5	52.7	29.2
^{1}L-D-SAM [38]	53.1	62.0	52.1	32.8
^{2}PaCo+SAM [8, 38]	57.5	62.1	58.8	39.3
ViT-B Backbone				
ViT [12]	37.5	56.9	30.4	10.3
DeiT-III [51]	48.4	**70.4**	40.9	12.8
^{1}DeiT-LT(ours)	55.6	65.2	54.0	37.1
^{2}DeiT-LT(ours)	**59.1**	66.6	**58.3**	**40.0**

Table 4. Results on iNaturalist-2018. (The teacher used to train student (DeiT-LT) can be identified by matching superscripts)

Method	iNaturalist-2018			
	Overall	Head	Mid	Tail
ResNet50 Backbone				
c-RT [20]	65.2	69.0	66.0	63.2
τ-Norm [21]	65.6	65.6	65.3	65.9
RIDE(3 exps) [56]	72.2	70.2	72.2	72.7
MiSLAS [65]	71.6	**73.2**	72.4	70.4
Disalign [63]	70.6	69.0	71.1	70.2
TSC [28]	69.7	72.6	70.6	67.8
GCL [26]	71.0	67.5	71.3	71.5
ImbSAM [68]	71.1	68.2	72.5	72.9
CBD$_{ENS}$ [18]	73.6	75.9	74.7	71.5
^{1}L-D-SAM [38]	70.1	64.1	70.5	71.2
^{2}PaCo+SAM [38]	73.4	66.3	73.6	75.2
ViT-B Backbone				
ViT [12]	54.2	64.3	53.9	52.1
DeiT-III [51]	61.0	72.9	62.8	55.8
^{1}DeiT-LT(ours)	72.9	69.0	73.3	73.3
^{2}DeiT-LT(ours)	**75.1**	70.3	**75.2**	**76.2**

primarily compare our results to the SotA methods, which train the networks from scratch. The other techniques utilize additional pre-training with extra data [7, 58], making the comparison unfair. Our proposed student network DeiT-LT outperformed the teachers used for their training by an average of 1.9% and 4.5% on CIFAR-10 LT and CIFAR-100 LT, respectively. This demonstrates the advantage of training the DeiT-LT transformer, which provides additional generalization improvements over the CNN teacher. Further, the DeiT-LT (PaCo+SAM) model significantly improves by 24.9% over the ViT baseline (which has the same augmentations as in DeiT-LT) and 28.4% over the data efficient DeiT-III transformer for CIFAR-10 LT dataset for $\rho = 100$. A similar improvement can also be observed for the CIFAR-100 LT dataset, where DeiT-LT (PaCo+SAM) fares better than ViT baseline and DeiT-III by 20.6% and 17.5%, respectively. This shows the effectiveness of the DeiT-LT distillation procedure via CNN teachers. Compared to CNN-based methods, we demonstrate that the transformer-based methods can achieve SotA performance when trained with DeiT-LT distillation procedure, combining both the scalability of transformers on head classes and utilizing inductive biases of CNN for tail classes. To the best of our knowledge, our proposed DeiT-LT for transformers is the *first work in literature that can achieve SotA performance for long-tailed data on small datasets when trained from scratch*. The other works [58] require transformer pre-training on large datasets, such as ImageNet, to achieve comparable performance on these small datasets.

Results on Large Scale Datasets. In this section, we present results attained by DeiT-LT on the large-scale long-tailed datasets of ImageNet-LT and iNaturalist-2018. We train all transformer-based methods for similar epochs for a particular dataset, to keep the comparison fair across all baselines (See Suppl. Sec. A.2). Table 3 presents the result on the ImageNet-LT dataset, where we find that when distilling using LDAM+DRW+SAM (L-D-SAM), our DeiT-LT significantly improves by 2.5% over the teacher network. Notably, it can be seen that our DeiT-LT method, when distilling from PaCo+SAM teacher, achieves a 1.6% performance gain over the already near SotA teacher network. Further, the distillation-based DeiT-LT method achieves a significant gain of 21.6% and 10.7% over the baseline transformer training methods, ViT and DeiT-III respectively. This demonstrates that improvement due to distillation scales well with an increase in the size of datasets. For iNaturalist-2018, we notice an improvement of close to 3% over the LDAM+DRW+SAM (L-D-SAM) teacher network and an improvement of 1.7% over the recent PaCo+SAM teacher. Additionally, we notice a significant improvement over the data-efficient transformer-based baselines. The data-efficient transformer-based methods struggle while modeling the tail

Table 5. Table showing ablations for various components in DeiT-LT for CIFAR-10 LT and CIFAR-100 LT.

OOD Distill	DRW	SAM	C10 LT	C100 LT
✗	✗	✗	70.2	31.3
✓	✗	✗	84.5	48.9
✓	✓	✗	87.3	54.5
✓	✓	✓	87.5	55.6

classes, which is supplemented via proposed Distillation loss in DeiT-LT. This enables DeiT-LT to work well across all the classes; the head classes benefit from enhanced learning capacity due to scalable Vision Transformer (ViT), and tail classes are learned well via distillation. Our results are superior for both datasets compared to the CNN-based SotA methods, demonstrating the advantage of DeiT-LT. (Refer Suppl. Sec. B for detailed results.)

6. Analysis and Discussion

Visualizations of Attentions. Our training methodology ensures that the `CLS` and `DIST` representations diverge while training. While the `CLS` token is trained against the ground truth, it cannot learn efficient representation for tail classes' images due to ViT's inability to train well on small amounts of data. Distilling from a teacher via out-of-distribution data and introducing re-weighting loss helps the `DIST` token to learn better representation for the images of minority classes as compared to the `CLS` token. We further corroborate this by comparing the attention visualization obtained through *Attention Rollout* [1], for the `CLS` and `DIST` token on tail images, for ImageNet-LT dataset (as CIFAR-10 is too small) using DeiT-LT. As can be seen in Fig. 5, the `CLS` and the `DIST` token focus on different parts of the image. The `DIST` token is able to identify the patches of interest (high red intensity) for images of tail classes, while the `CLS` token fails to do so. The diversity in localized regions demonstrates the complementary information present across the `CLS` and `DIST` experts, which is in contrast with DeiT, where both the tokens `CLS` and `DIST` are quite similar. We compare visualization with different methods in Suppl. Sec. D.

Ablation Analysis Across DeiT-LT components. We analyze the influence of three key components of our DeiT-LT method, namely OOD distillation, training the Tail Expert classifier with DRW loss, and using SAM teacher for distillation. As can be seen in Table 5, using OOD distillation brings around 14% and 18% improvement over DeiT [48] for CIFAR-10 LT and CIFAR-100 LT, respectively, followed by the other two components, which further improve the accuracy by around 3% and 6.7% for CIFAR-10 LT and CIFAR-100 LT, respectively.

Analysis across Transformer Variants. In this section, we aim to analyze the performance of DeiT-LT across transformer variants having different capacities. For this, we fix the teacher network and training schedules while varying the network sizes. We experiment with the ViT-Ti, ViT-S, and ViT-B architectures, as introduced in the original ViT work [12]. In Table 6, we observe that the proposed DeiT-LT method scales well with the increased capacity of the Transformer network, and leads to performance improvements.

Table 6. Analysis across transformer capacity for CIFAR-10 LT and CIFAR-100 LT for DeiT-LT student($\rho = 100$) with PaCo teacher.

Model	Overall	Head	Mid	Tail
CIFAR-10 LT ($\rho = 100$)				
DeiT-LT Tiny (Ti)	80.8	89.7	75.1	79.4
DeiT-LT Small (S)	85.5	92.7	81.5	83.7
DeiT-LT Base (B)	87.5	94.5	84.1	85.0
CIFAR-100 LT ($\rho = 100$)				
DeiT-LT Tiny (Ti)	49.3	66.3	50.0	27.3
DeiT-LT Small (S)	54.3	72.6	54.8	31.1
DeiT-LT Base (B)	55.6	73.1	56.9	32.1

Limitations. One limitation of our framework is that the learning for tail classes is done mostly through distillation. Hence, the performance on tail classes remains similar (Table 3 and 4) to that of the CNN classifier. Future works can aim to develop adaptive methods that can shift their focus from CNN to ground truth labels, as the CNN feedback saturates.

7. Conclusion

In this work, we introduce DeiT-LT, a training scheme to train ViTs from scratch on real-world long-tailed datasets efficiently. We reintroduce the idea of knowledge distillation into ViT students via teacher CNN, as it enables effective learning on the tail classes. This distillation component was found to be redundant and removed from the latest DeiT-III. Further, in DeiT-LT, we introduce out-of-distribution (OOD) distillation via the teacher, in which we pass strongly augmented images to teachers originally trained via mild augmentations for distillation. The distillation loss is re-weighted to enhance the focus on learning from tail classes. This helps make the classification token an expert on the head classes and the distillation token an expert on the tail classes. To improve generality in minority classes, we induce low-rank features in ViT by distilling from teachers trained from Sharpness Aware Minimization (SAM). The proposed DeiT-LT scheme allows ViTs to be trained from scratch as CNNs and achieve performance competitive to SotA without requiring any pre-training on large-datasets.

Acknowledgements. Harsh Rangwani is supported by the PMRF Fellowship. We thank Sumukh for the discussions on the draft. This work is supported by the SERB-STAR Project (STR/2020/000128) and KIAC Grant.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. 6, 8, 5

[2] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021. 2

[3] Maksym Andriushchenko, Dara Bahri, Hossein Mobahi, and Nicolas Flammarion. Sharpness-aware minimization leads to low-rank features. *arXiv preprint arXiv:2305.16292*, 2023. 5

[4] Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *ICCV*, 2021. 6

[5] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019. 1, 2, 4, 5, 6, 7

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1

[7] Jun Chen, Aniket Agarwal, Sherif Abdelkarim, Deyao Zhu, and Mohamed Elhoseiny. Reltransformer: A transformer-based long-tail visual relationship recognition. In *CVPR*, 2022. 2, 7

[8] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *ICCV*, 2021. 2, 6, 7, 1, 3

[9] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 1, 2, 4, 6, 7

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[11] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE TPAMI*, 2015. 1

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 3, 6, 7, 8

[13] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. 2, 3, 5, 6, 7

[14] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[16] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *CVPR*, 2021. 2

[17] Yan Hong, Jianfu Zhang, Zhongyi Sun, and Ke Yan. Safa:sample-adaptive feature augmentation for long-tailed image classification. In *ECCV*, 2022. 7

[18] Ahmet Iscen, André Araujo, Boqing Gong, and Cordelia Schmid. Class-balanced distillation for long-tailed visual recognition. 2021. 7

[19] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *CVPR*, 2020. 6

[20] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019. 2, 6, 7

[21] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2020. 7

[22] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *CVPR*, 2020. 2

[23] Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. In *Advances in Neural Information Processing Systems*, pages 18970–18983. Curran Associates, Inc., 2021. 1, 2, 6

[24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[25] Jun Li, Zichang Tan, Jun Wan, Zhen Lei, and Guodong Guo. Nested collaborative learning for long-tailed visual recognition. In *CVPR*, pages 6949–6958, 2022. 1

[26] Mengke Li, Yiu-ming Cheung, and Yang Lu. Long tail visual recognition via gaussian clouded logit adjustment. In *CVPR*, 2022. 6, 7

[27] Tianhao Li, Limin Wang, and Gangshan Wu. Self supervision to distillation for long-tailed visual recognition. In *ICCV*, 2021. 6

[28] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6918–6928, 2022. 7

[29] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019. 5

[30] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. Retrieval augmented classification for long-tail visual recognition. In *CVPR*, 2022. 2, 3, 6

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[32] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar.

Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020. 1, 2, 6, 7

[33] Gaurav Kumar Nayak, Konda Reddy Mopuri, and Anirban Chakraborty. Effectiveness of arbitrary transfer sets for data-free knowledge distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1430–1438, 2021. 4

[34] Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4262–4274, 2021. 2

[35] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020. 4

[36] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021. 5, 4

[37] Harsh Rangwani, Konda Reddy Mopuri, and R Venkatesh Babu. Class balancing gan with a classifier in the loop. In *Conference on Uncertainty in Artificial Intelligence (**UAI**)*, 2021. 2

[38] Harsh Rangwani, Sumukh K Aithal, Mayank Mishra, and Venkatesh Babu R. Escaping saddle points for effective generalization on class-imbalanced data. In *Advances in Neural Information Processing Systems*, pages 22791–22805. Curran Associates, Inc., 2022. 4, 5, 6, 7, 1, 2

[39] Harsh Rangwani, Naman Jaswani, Tejan Karmali, Varun Jampani, and R. Venkatesh Babu. Improving gans for long-tailed data through group spectral regularization. In *European Conference on Computer Vision (**ECCV**)*, 2022. 2

[40] Harsh Rangwani*, Lavish Bansal*, Kartik Sharma, Tejan Karmali, Varun Jampani, and R. Venkatesh Babu. Noisytwins: Class-consistent and diverse image generation through style-GANs. In *Conference on Computer Vision and Pattern Recognition (**CVPR**)*, 2023. 2

[41] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. *arXiv preprint arXiv:2007.10740*, 2020. 2, 7

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 5, 1

[43] Jiang-Xin Shi, Tong Wei, Zhi Zhou, Xin-Yan Han, Jie-Jing Shao, and Yu-Feng Li. Parameter-efficient long-tailed recognition. *arXiv preprint arXiv:2309.10019*, 2023. 4

[44] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021. 1

[45] Zichang Tan, Yang Yang, Jun Wan, Hanyuan Hang, Guodong Guo, and Stan Z Li. Attention-based pedestrian attribute analysis. *TIP*, 28(12):6126–6140, 2019. 4

[46] Changyao Tian, Wenhai Wang, Xizhou Zhu, Jifeng Dai, and Yu Qiao. Vl-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition. In *European Conference on Computer Vision*, pages 73–91. Springer, 2022. 6, 2, 4

[47] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021. 1

[48] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers amp; distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357, 2021. 1, 2, 3, 4, 5, 6, 8

[49] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 32–42, 2021. 6

[50] Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Jakob Verbeek, and Herve Jegou. Three things everyone should know about vision transformers. *arXiv preprint arXiv:2203.09795*, 2022. 1

[51] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 516–533. Springer, 2022. 1, 5, 6, 7

[52] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018. 1, 5

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 3

[54] Angelina Wang and Olga Russakovsky. Overwriting pretrained bias with finetuning data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3957–3968, 2023. 2, 3

[55] Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. Contrastive learning based hybrid networks for long-tailed image classification. In *CVPR*, 2021. 2, 6

[56] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *ICLR*, 2021. 1, 3, 6, 7

[57] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023. 4

[58] Zhengzhuo Xu, Ruikang Liu, Shuo Yang, Zenghao Chai, and Chun Yuan. Learning imbalanced data with vision transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6, 7, 2

[59] Zhengzhuo Xu, Shuo Yang, Xingjun Wang, and Chun Yuan. Rethink long-tailed recognition with vision transforms. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 6, 2

[60] Han-Jia Ye, Hong-You Chen, De-Chuan Zhan, and Wei-Lun Chao. Identifying and compensating for feature deviation in imbalanced deep learning. *arXiv preprint arXiv:2001.01385*, 2020. 2

[61] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 4

[62] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 2, 4

[63] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *CVPR*, 2021. 7

[64] Yongshun Zhang, Xiu-Shen Wei, Boyan Zhou, and Jianxin Wu. Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. In *AAAI*, 2021. 6

[65] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *CVPR*, 2021. 6, 7

[66] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE TPAMI*, 2017. 2

[67] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, 2020. 1, 6

[68] Yixuan Zhou, Yi Qu, Xing Xu, and Hengtao Shen. Imbsam: A closer look at sharpness-aware minimization in class-imbalanced recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11345–11355, 2023. 7