Semantic Transfer from Head to Tail: Enlarging Tail Margin for Long-Tailed Visual Recognition

Shan Zhang[*,1], Yao Ni[*,1], Jinhao Du[2], Yanxia Liu[†,3], Piotr Koniusz[†,4,1]
[1]Australian National University, [2]Peking University, [3]Beijing Union University, [4]Data61♥CSIRO
[1]firstname.lastname@anu.edu.au, [2]dujinhao02@gmail.com, [3]yanxia.liu@163.com

Abstract

Deep neural networks excel in visual recognition tasks, but their success hinges on access to balanced datasets. Yet, real-world datasets often exhibit a long-tailed distribution, compromising network efficiency and hampering generalization on unseen data. To enhance the model's generalization in long-tailed scenarios, we present a novel feature augmentation approach termed SeMAntic tRansfer from head to Tail (SMART), which enriches the feature patterns for tail samples by transferring semantic covariance from the head classes to the tail classes along semantically correlating dimensions. This strategy boosts the model's generalization ability by implicitly and adaptively weighting the logits, thereby widening the classification margin of tail classes. Inspired by the success of this weighting, we further incorporate a semantic-aware weighting strategy for the loss tied to tail samples. This amplifies the effect of enlarging the margin for tail classes. We are the first to provide theoretical analysis that demonstrates a large semantic diversity in tail samples can increase class margins during the training stage, leading to improved generalization. Empirical observations support our theory. Notably, with no need for extra data or learnable parameters, SMART achieves state-of-the-art results on five long-tailed benchmark datasets: CIFAR-10/100-LT, Places-LT, ImageNet-LT, and iNaturalist 2018.

1. Introduction

The breakthroughs in deep neural networks [26] are rooted in the utilization of abundant training data from diverse balanced categories. These meticulously curated balanced datasets have empowered deep models to excel across computer vision tasks, including image recognition [14,57], video analysis [48,52], object detection [8,9], image generation [42–44] and self-supervised learning [54,72,73], *etc*.

Despite impressive achievements of deep models, their success hinges on data that is evenly distributed across categories. Yet, most real-world datasets exhibit a long-tailed distribution, with few dominant (head) classes with numerous instances and most classes (tail classes) with sparse samples. This imbalance hinders the efficacy of deep models, causing poor generalization to unseen data, especially for the tail classes [19].

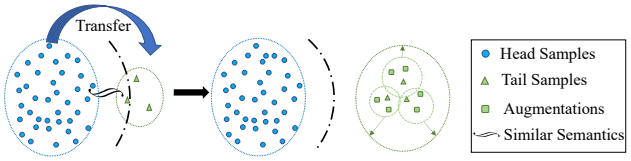


Figure 1. Semantic data augmentation for long-tailed problems increases the diversity of samples of tail classes and enlarges the classification margin boundary.

To tackle this challenge, recent research has focused on three main approaches. The first involves re-sampling the imbalanced dataset [4, 7, 19, 35, 45, 47] either by 'over-sampling' tail classes or 'under-sampling' the head classes. Yet, simplistic re-sampling has the potential to overfit to tail classes, as the repeatedly selected samples often share similar image contexts [49]. The second approach emphasizes re-weighting the loss of tail samples, employing a factor inversely proportional to the sample frequency of each class. This ensures that instances from tail classes receive greater importance compared to those from head classes [6,13,46,74]. The goal is to achieve balanced training gradients. However, this method often overlooks the inherent diversity of samples within tail classes, leading to limited within-class variations.

The third branch emphasizes the enrichment of tail samples through data augmentation, targeting either the raw-pixel space [45] of an image or the feature space [35]. However, methods that operate on raw pixel space or the variance vector of features may inadvertently erode the underlying semantics. Our SMART adheres to this family of methods, yet sets itself apart by leveraging a technique based on research [57], that highlights the role of the covariance matrix in preserving semantic diversity. Accordingly, as illustrated in Figure 1, we channel the diversity of samples

[*]Equal contribution. [†]Corresponding author.

of the head classes, captured by their covariance, into the feature space of tail classes. Intuitively, diversifying the tail features via such an augmentation will increase the semantic overlap with the head ones. However, as we optimize the classifier to distinguish between them, it naturally separates the tail features further from the head features, creating a larger margin for these tail classes.

Our theoretical analysis confirms the enhanced effect of enlarging the tail margin. Delving deeper, we discovered that the success of this augmentation is due to its implicit and adaptive adjustment of sample logits. Drawing inspiration from this, we further introduce a semantic-aware reweighting strategy for the loss on tail samples. This strategy utilizes the second-order co-occurrences of classes to adaptively enhance tail 'hard' points [1] with 'push-back soft weights', giving them greater emphasis and ensuring they are not overshadowed by the dominant classes.

Our extensive experiments validate the margin-enlarging effect of SMART and affirm its design rationale, highlighting its effectiveness in enhancing the model's generalization ability. Our contributions can be summarized as follows:

i. We are the first to devise a semantic covariance transformation matrix that ensures augmented features align closely with desired head features. We are also the first to theoretically prove that this semantic transfer from head to tail classes has the effect of enlarging margin for tail classes, enhancing the model's generalization.

ii. The proposed SeMAntic tRansfer from head to Tail, abbreviated as SMART, borrows semantic information from head classes to augment tail classes without requiring extra data or learnable parameters. This approach not only enriches the tail classes but also calibrates the feature distribution distorted by limited tail data.

iii. The proposed method gains significant improvement over the state-of-the-art on five popular long-tailed benchmarks, including CIFAR-10/100-LT, Places-LT, ImageNet-LT and iNaturalist 2018.

2. Related Works

Traditional strategies to counter an imbalanced distribution fall into three branches: re-sampling methods [4, 7, 19, 62–65], re-weighting strategies [6, 13, 31, 40, 46, 46, 58, 61, 74] and data augmentation [7, 24, 35, 45, 60].

2.1. Re-sampling methods

Numerous studies [4, 7, 47, 62–65] have centered on addressing class imbalances in training using re-sampling, which includes under-sampling and over-sampling. Undersampling mainly reduces the majority of samples. However, in cases of significant class disparity, it is often impractical and can cause training instability, especially under extreme imbalance [47]. Over-sampling entails increasing the number minority class samples to achieve balance but it might fail to capture the full diversity of the majority classes, risking overfitting to the minority classes.

2.2. Re-weighting strategies

Re-weighting strategies primarily involve two approaches: loss function re-weighting and logit adjustment. Loss function re-weighting allocates different weights to individual training samples, rebalancing their contribution to the overall loss function during training. Approaches [6, 13, 31, 46, 74] assign higher weights to instances from tail classes than those from head classes. Alternatively, logit adjustment methods recalibrate the logit values to maintain a balanced gradient. For example, techniques from [40, 61] enhance the accuracy of model via adjusting these logits post-training. However, many re-weighting methods hinge on heuristic designs. In contrast, our semantic-aware weighting strategy prioritizes second-order co-occurrences within classes, avoiding the risk of semantic eroding of the conventional first-order frequency methods.

2.3. Data augmentation

Data augmentation in the context of long-tailed learning can be broadly classified into two categories: image-level augmentation and feature-level augmentation.

Image-level augmentation. Image-level augmentation techniques, such as works [15, 34, 51] have shown promise in computer vision tasks, and their potential extends to addressing imbalanced data scenarios. For example, CMO [45] selects images across different distributions based on specific characteristics of long-tailed distributions. Image generation based method SMOTE [7] enriches tail classes by interpolating minority samples with their neighboring majority counterparts. Major-to-minor Translation (M2m) [22] transfers knowledge from dominant classes with the aid of a pre-trained classifier. BLT [24] employs gradient-driven image generation. Method [60] directly uses pure noisy images as tail class samples. Despite merits, these methods largely use raw-pixel image space which lacks deep semantic meaning. Moreover, the integration of deep generative models inflates the computational demands. In contrast, our SMART operates in a semantically rich space, bypassing the pitfalls of raw-pixel based methods and eliminating the need for extra learning parameters.

Feature-level augmentation. Complementing traditional image-level augmentation, feature-level augmentation in CNNs helps prevent overfitting. This is grounded in the observation that CNNs prowess to capture high-level semantic details, where altering deep features leads to meaningful semantic changes in the image space [3]. To address

[1] Tail hard points are samples within tail classes that most closely resemble head classes in terms of semantic characteristics.

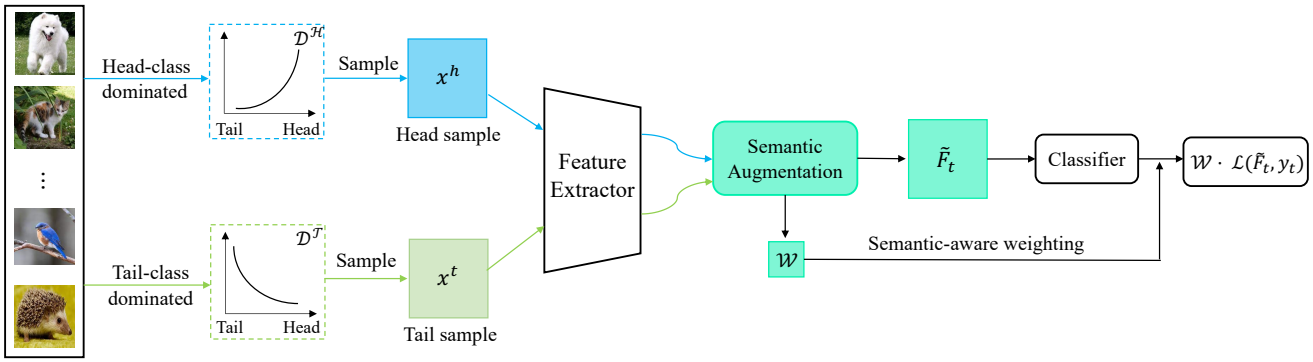


Figure 2. The pipeline of SeMAntic tRansfer from head to Tail (SMART). The head class data and tail class data are fed into the feature extractor to obtain the deep features. The augmented features are produced by transferring the diversity of the head sample to the tail sample along the semantically meaningful directions, as measured by the covariance matrix. Refer to the text for details and the pseudo-code 1 for the training procedure. The augmentation process is adopted in the training phase, without incurring inference overheads.

class imbalances, LEAP [35] leverages the angle variance of head classes to aid tail classes. Similarly, ISDA [57] employs class-conditional statistics for semantic augmentation, though it fails to measure covariance in underrepresented classes. MetaSAug [30] opts for an intricate meta-learning approach, necessitating extensive iterations and additional validation samples from tail classes. In contrast, our method applies Exponential Moving Average (EMA) [23] to update the covariance matrix across iterations based on label frequencies. This facilitates the transfer of rich semantic content from dominant to minority classes, enhancing intra-class variability without the need for extra data or demanding computation. While concurrent research [29] also endorses augmenting tail classes with head class semantics, the semantics are represented by feature points that differ in their co-occurrences of covariance matrix. Moreover, our approach is underpinned by robust theoretical motivation and a proof, offering solid foundations.

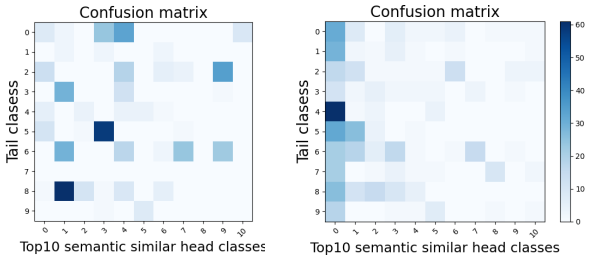
3. Method

Starting with the long-tailed setting, we identify two principles that improve its performance from theoretical and experimental standpoints. Using such insights, we develop our method, grounded in theory and empirical evidence.

3.1. Preliminaries

In the long-tailed learning, the goal is to train a deep neural network model using an imbalanced training dataset, represented as $\mathcal{D}=\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ where n is the number of training samples. Each training sample $\boldsymbol{x}_i$ is associated with label y_i from a set of K classes, denoted as $y_i \in \{1, \ldots, K\}$. Within each class k, n_k indicates the number of samples, and $\mathcal{D}_k$ represents the set of samples in class k.

A neural network employs a feature extractor, described by $f(\cdot; \boldsymbol{\Theta}_f) : \boldsymbol{x} \mapsto \boldsymbol{a}$, which is parameterized with $\boldsymbol{\Theta}_f$ and is comprised of several convolutional layers. This extractor transforms the input into a feature space. Following the feature extraction, a linear classifier, represented as $\phi(\cdot; \boldsymbol{W}, \boldsymbol{b}) : \boldsymbol{a} \mapsto \boldsymbol{z}$, produces the output logit $\boldsymbol{z}$, and is characterized by its weight matrix $\boldsymbol{W} = [\boldsymbol{w}_1, ..., \boldsymbol{w}_K]^T \in \mathbb{R}^{K \times C}$ where C defines the feature dimension, and a bias vector $\boldsymbol{b} = [b_1, ..., b_K]^T \in \mathbb{R}^K$. All network parameters are encapsulated within $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_f, \boldsymbol{W}, \boldsymbol{b}\}$.



(a) Baseline (b) Ours

Figure 3. Comparison of confusion matrices: Our method *vs*. the baseline on CIFAR100-LT shows our approach dramatically reduces misclassification among tail classes and their top 10 similar head classes, with '0' on the x-axis indicating the tail class.

3.2. Motivations

As we are interested in improving the performance of the model on an imbalanced dataset, we present the following lemma that bounds the generalization error in the long-tailed setting.

Lemma 1. ***(Theorem 2 of [46])*** *Let $\gamma_k = t - \max_{(\boldsymbol{x},y)\in\mathcal{D}_k} l_k$ be the margin for class k under threshold $t \geq 0$ where l_k is the standard negative log-likelihood with Softmax. Denote err_{bal} as the 0-1 error on the balanced test dataset and $\mathcal{C}$ as some complexity measure related to the Rademacher complexity [2]. With probability $1-\delta$, for all $\gamma_k > 0$, we have the balanced-class generalization error bounded by:*

$$err_{\text{bal}}(t) \lesssim \frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{\gamma_k} \sqrt{\frac{\mathcal{C}}{n_k}} + \frac{\log n}{\sqrt{n_k}} \right), \qquad (1)$$

where we use $\lesssim$ *to hide constant factors. The optimal* err_{bal} *is obtained at* $\gamma_k^* = \frac{\beta}{\sqrt[4]{n_k}\sum_{j=1}^{K} n_j^{-1/4}}$ *with a constraint on* $\beta = \sum_{j=1}^{K} \gamma_j$.

Lemma 1 reveals that to optimize the generalization capability in long-tailed settings, one can: 1) increase the number of training samples per class, which means more samples help achieve better generalization; 2) enlarge the margin for each class relative to the decision boundaries. The optimal γ_k^* further underscores that tail classes with fewer samples require a larger margin for better generalization.

In deep learning, empirical observations consistently indicate that misclassification predominantly occurs between semantically related categories. Addressing this requires researchers to use more balanced datasets, ensuring models can distinguish between such closely related categories. We hypothesize this challenge is likely more pronounced in long-tail learning, where the significant imbalance between the well-represented head categories and the underrepresented tail ones heightens the misclassification risk.

To test our hypothesis about tail-class misclassification, we employed WordNet [41] to identify semantically similar head classes for tail classes on CIFAR100-LT. Leveraging Lin Similarity [33], we quantified semantic similarities between classes. For instance, in Figure 3, using Lin Similarity, the top 10 semantically similar head classes (indexed from 1-10) to the tail class 'train' (indexed 0) are: bus, bicycle, bridge, motorcycle, road, pickup truck, camel, elephant, lawn mower, and mountain. In Figure 3a, the confusion matrix reveals that tail classes are often misclassified into their top-10 semantically similar head classes rather than their true categories. This consistent mislabeling matches our hypothesis and highlights that the feature space of tail classes, despite being limited, overlaps significantly with the more expansive features of their analogous head classes.

Based on the above analysis and observations, to enhance performance in the long-tailed setting, it is essential to expand the margin of tail classes, with the focus on increasing distances to decision boundaries from their semantically similar head classes.

3.3. Semantic augmentation for the tail classes

To improve the classification of tail classes in relation to their semantically similar head classes, we propose to leverage the diversity inherent in head classes. Our method augments tail class features by integrating the covariance matrix of head classes, while preserving the content of tail class samples.

We initiate this semantic augmentation by extracting features from f, and subsequently, we update the mean and covariance for each class. Denote the flattened feature extracted from f for sample $\boldsymbol{x}$ as $\boldsymbol{F} \in \mathbb{R}^{HW \times C}$ where HW is the spatial dimension. We update the mean $\boldsymbol{\mu}_k \in \mathbb{R}^{C\times 1}$ and covariance $\boldsymbol{\Sigma}_k \in \mathbb{S}_{++}^C$ (or $\boldsymbol{\Sigma}_k \in \mathbb{S}_{+}^C$ if rank deficient) for each class k using the exponential moving average as (EMA):

$$\boldsymbol{\mu}_k^{(t)} = (1-\alpha_k^{(t)})\boldsymbol{\mu}_k^{(t-1)} + \alpha_k^{(t)}\boldsymbol{\mu}_k^B, \tag{2}$$

$$\boldsymbol{\Sigma}_k^{(t)} = (1-\alpha_k^{(t)})\boldsymbol{\Sigma}_k^{(t-1)} + \alpha_k^{(t)}\boldsymbol{\Sigma}_k^B, \tag{3}$$

where $\boldsymbol{\mu}_k^B$ and $\boldsymbol{\Sigma}_k^B$ are estimations from the current batch for class k, with $\alpha_k^{(t)} = M_k^B / N_k^{(t)}$ and $N_k^{(t)} = N_k^{(t-1)} + M_k^B$ where N_k is the number of samples for class k the feature extractor has seen, and M_k^B is the number of samples for class k in current batch.

Leveraging $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, in pursuit of enhancing feature diversity without compromising the content of the original features, we randomly sample a head class k_h for a given tail class k_t. After augmentation, the features of class k_t, originally denoted as $\{\boldsymbol{F}_t\}$, are transformed to $\{\tilde{\boldsymbol{F}}_t\}$. These augmented features $\{\tilde{\boldsymbol{F}}_t\}$ are expected to preserve the tail class mean $\boldsymbol{\mu}_t$ and achieve a covariance of $\boldsymbol{\Sigma}_{th}$ that is akin to the covariance of head class $\boldsymbol{\Sigma}_h$. The computation of $\boldsymbol{\Sigma}_{th}$ is detailed as follows:

$$\tilde{\boldsymbol{\Sigma}}_{th} = \boldsymbol{\Sigma}_t^T \boldsymbol{\Sigma}_h,$$

$$\boldsymbol{\Gamma} = \text{Softmax}(\tilde{\boldsymbol{\Sigma}}_{th}; \tau), \tag{4}$$

$$\boldsymbol{\Sigma}_{th} = \boldsymbol{\Gamma} \odot \tilde{\boldsymbol{\Sigma}}_{th}. \tag{5}$$

Here, $\boldsymbol{\Sigma}_t$ is the covariance of tail class, Softmax incorporates a temperature parameter $\tau > 0$ and operates over the final dimension of the matrix $\tilde{\boldsymbol{\Sigma}}_{th}$. A larger τ implies a greater emphasis on the most similar features, and $\odot$ denotes the element-wise multiplication. The validity for this approach is theoretically elaborated in Appendix §A.4 and experimentally ablated in Sec. 4.3.

We acquire the augmented feature as follows:

$$\tilde{\boldsymbol{F}}_t = \frac{\boldsymbol{F} - \boldsymbol{\mu}_t}{\sqrt{\boldsymbol{\sigma}_t^2 + \varepsilon}} \boldsymbol{\Sigma}_{th} + \boldsymbol{\mu}_t. \tag{6}$$

Here, $\varepsilon = 1e{-}8$ prevents division by 0 and $\boldsymbol{\sigma}_t^2 = \text{diag}(\boldsymbol{\Sigma}_t) \in \mathbb{R}^{C\times 1}$ is used specifically to sidestep matrix inversion operations that are not friendly for GPU processing.

3.4. Analysis for semantic augmentation

Based on the theoretical analysis provided below, we affirm the effectiveness of our semantic augmentation. This augmentation implicitly strengthens the effect of enlarging the margin for tail classes, leading to improved generalization as evidenced in Lemma 1 for long-tailed scenarios. The detailed proofs are available in Appendix §A.

Assumption 1. *Assume that after augmentation, the feature* $\tilde{\boldsymbol{a}}_i^t$ *presented to the classifier, which comes from a tail sample* $\boldsymbol{x}_i^t$*, can be approximately represented by a distribution* $\tilde{\boldsymbol{a}}_i^t \sim \mathcal{N}(\boldsymbol{a}_i^t, \Delta\boldsymbol{\Sigma}_{th}^i)$*. Here,* $\boldsymbol{a}_i^t$ *is the feature obtained without augmentation, and* $\Delta\boldsymbol{\Sigma}_{th}^i$ *is a positive definite covariance matrix.*

Lemma 2. *Given the negative log softmax function, the loss L_k for samples of class k without augmentation can be derived as:*

$$L_k = \frac{1}{n_k}\sum_{i=1}^{n_k} -\log\frac{e^{\boldsymbol{w}_k^T\boldsymbol{a}_i+b_k}}{\sum_{j=1}^{K}e^{\boldsymbol{w}_j^T\boldsymbol{a}_i+b_j}} = \frac{1}{n_k}\sum_{i=1}^{n_k}\log\Big(1+\sum_{j\neq k}e^{d_i\|\boldsymbol{w}_j-\boldsymbol{w}_k\|_2\cdot\mathrm{sign}(\cos\theta_{i,jk})}\Big). \quad (7)$$

Drawing upon [37], the decision boundary between class j and class k can be formulated as: $(\boldsymbol{w}_j-\boldsymbol{w}_k)^T\boldsymbol{a}+(b_j-b_k)=0$, d_i is the distance from point $\boldsymbol{a}_i$ to the decision boundary, $\theta_{i,jk}$ denotes the angle between $\boldsymbol{w}_j-\boldsymbol{w}_k$ and $\boldsymbol{a}_i$.

Lemma 2 (the proof is in Appendix §A.2) indicates that an easy sample $\boldsymbol{a}_i$ of class k, which aligns with $\boldsymbol{w}_k$, typically has an opposite direction with $\boldsymbol{w}_j-\boldsymbol{w}_k$, resulting in the condition $\mathrm{sign}(\cos\theta_{i,jk})=-1$. Minimizing L_k under this condition enlarges d_i, pushes $\boldsymbol{a}_i$ away from the decision boundary, thereby expanding the margin of class k. For hard samples on the wrong side with $\mathrm{sign}(\cos\theta_{i,jk})=1$, the minimization of L_k works to correct their position. Building on this intuition, the loss function has the intrinsic effect of increasing the margin of easy samples and correcting hard samples. However, in long-tailed situations, this effect weakens for tail classes as head classes overshadow it.

Theorem 1. *Assert that Assumption 1 holds when using our augmentation. The the loss function L_k^t for tail class k is:*

$$L_k^t = \frac{1}{n_k}\sum_{i=1}^{n_k}\mathbb{E}_{\tilde{\boldsymbol{a}}_i^t}\Big[-\log\frac{e^{\boldsymbol{w}_k^T\tilde{\boldsymbol{a}}_i^t+b_k}}{\sum_{j=1}^{K}e^{\boldsymbol{w}_j^T\tilde{\boldsymbol{a}}_i^t+b_j}}\Big] \le \frac{1}{n_k}\sum_{i=1}^{n_k}\log\Big(1+\sum_{j\neq k}\beta_{jk}^i e^{d_i\|\boldsymbol{w}_j-\boldsymbol{w}_k\|_2\cdot\mathrm{sign}(\cos\theta_{i,jk})}\Big), \quad (8)$$

where $\beta_{jk}^i = e^{\frac{1}{2}(\boldsymbol{w}_j-\boldsymbol{w}_k)^T\Delta\boldsymbol{\Sigma}_{th}^i(\boldsymbol{w}_j-\boldsymbol{w}_k)}$. Furthermore,

$$\beta_{jk}^i = \exp\Big(\frac{1}{2}{\boldsymbol{v}_{jk}^i}^T\boldsymbol{\Lambda}^i\boldsymbol{v}_{jk}^i\Big) > 1, \quad (9)$$

where $\boldsymbol{V}^i\boldsymbol{\Lambda}^i{\boldsymbol{V}^i}^T=\Delta\boldsymbol{\Sigma}_{th}^i$ and $\boldsymbol{v}_{jk}^i={\boldsymbol{V}^i}^T(\boldsymbol{w}_j-\boldsymbol{w}_k)$.

The proof can be found in Appendix §A.2. Comparing Eq. 8 in Theorem 1 with Lemma 2, it becomes evident that the augmentation implicitly strengths the effect of enlarging the margin and rectifying hard samples for tail classes using the factor β_{jk}^i. Eq. 9 in Theorem 1 further demonstrates that this strengthening factor β_{jk}^i is not only larger than the strength (equal to 1) in Lemma 2, but is also adaptively adjusted by $\Delta\boldsymbol{\Sigma}_{th}^i$. Notably, the covariance matrix borrowed from a head class with greater diversity and semantic similarity to the tail class will have larger values in $\boldsymbol{\Lambda}^i$, resulting in a larger β_{jk}^i. With a larger β_{jk}^i, the effect of expanding the margin for tail classes will be amplified, leading to improved generalization in Lemma 1.

3.5. Semantic-aware weighted loss function

The insights drawn from Theorem 1 suggest that the augmentation process indirectly assigns weights to the elements within the log function, favorably influencing generalization. To further harness this potential, we introduce a more explicit approach, leveraging the covariance matrix to create "push-back soft weights" to counter the overshadowing effect of the dominant classes. This method prioritizes and applies greater penalties to hard tail samples, ensuring our model undergoes a semantic-aware optimization. Consequently, this results in a semantic-aware weighted loss:

$$L_{\mathrm{SMART}} = \frac{1}{n}\sum_{i=1}^{n}\omega_i\ell\big((\tilde{\boldsymbol{F}}_t^i, y_i);\boldsymbol{\Theta}\big), \quad (10)$$

where $\ell(\cdot;\boldsymbol{\Theta})$ is the loss function for a sample and for each $\omega_i\in\mathcal{W}$:

$$\omega_i = \begin{cases} 1 & \text{for } (\boldsymbol{x}_i, y_i)\in\mathcal{D}^{\mathcal{H}}, \\ \dfrac{1}{C}\displaystyle\sum_{j=1}^{C}\sum_{k=1}^{C}\boldsymbol{\Sigma}_{th}[j,k] & \text{for } (\boldsymbol{x}_i, y_i)\in\mathcal{D}^{\mathcal{T}}, \end{cases} \quad (11)$$

where $\boldsymbol{\Sigma}_{th}[j,k]$ is the (j,k)-th entry of $\boldsymbol{\Sigma}_{th}$. Algorithm 1 and the pipeline in Figure 2 illustrate our method.

Algorithm 1 SeMAntic tRansfer from head to Tail

Input: Training dataset $\mathcal{D}$, learned parameters $\boldsymbol{\Theta}$, loss function $L(\cdot)$;
Output: Trained model;

Sample tail-class dominated dataset $\mathcal{D}^{\mathcal{T}}\sim T$
Sample head-class dominated dataset $\mathcal{D}^{\mathcal{H}}\sim H$
for epoch $=1,\dots,E$ **do**
  for batch $j=1,\dots,B$ **do**
    Sample a batch $(\boldsymbol{x}^t, y^t)^{|B|}$ from $\mathcal{D}^{\mathcal{T}}$
    Sample a batch $(\boldsymbol{x}^h, y^h)^{|B|}$ from $\mathcal{D}^{\mathcal{H}}$
    Update $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ in Eq. 2 & 3
    Estimate covariance $\boldsymbol{\Sigma}_{th}$ in Eq. 5
    Generate augmented features $\tilde{F}_t$ in Eq. 6
    Computer the weights $\mathcal{W}^{|B|}$ in Eq. 10
    $\boldsymbol{\Theta}\leftarrow\boldsymbol{\Theta}-\eta\nabla\mathcal{W}^{|B|}L\big((\tilde{\boldsymbol{F}}_t^{|B|}, (y^t)^{|B|});\boldsymbol{\Theta}\big)$
  end for
end for

4. Experiments

4.1. Experiment setup

Datasets. We test our method on the most commonly used long-tailed recognition benchmark datasets, including CIFAR10-LT [13], CIFAR100-LT [13], ImageNet-LT [38], Places-LT [70], and iNaturalist 2018 [53]. **CIFAR10-LT and CIFAR100-LT** [13] are derived from the balanced CIFAR datasets [25] and are sampled using an exponential decay across classes. The CIFAR-10 dataset consists of 50,000 training images and 10,000 validation images, with a

resolution of 32×32 and 10 classes. The CIFAR-100 maintains the same number of images and resolution as CIFAR-10, but spans 100 classes. We utilized the long-tailed versions of these datasets, introducing imbalance factors of 100 and 50, respectively. The degree of data imbalance was controlled using an Imbalance Factor (IF), calculated as the ratio between number of instances in the most frequent category and the least frequent category. **ImageNet-LT** [38] is a long-tailed version of vanilla ImageNet. It was constructed by sampling a subset of images following the Pareto distribution with a power value of $\alpha = 0.6$. ImageNet-LT comprises 115.8K images across 1,000 categories. **Places-LT** is created from the large-scale dataset Places [70]. It consists of 184.5K images distributed across 365 categories, with the number of instances per class ranging from 5 to 4,980. **iNaturalist 2018** [53] stands as the largest dataset for long-tailed visual recognition. It encompasses 437.5K images belonging to 8,142 classes. This dataset exhibits extreme imbalance with an imbalance factor of 512.

Implementation Details. Following works [13, 21], performance is mainly reported as the overall top-1 accuracy on three disjoint subsets: many (more than 100 images), medium (20 to 100 images), and few (less than 20 images). For the CIFAR datasets, we use ResNet-32 [16] as our backbone network. The network is trained for 200 epochs, and the initial learning rate is set to 0.1 and decreases by 0.1 at epoch 220 and 260, respectively, following the training strategy in work [45]. For ImageNet-LT, we utilize ResNet-50 [16], and train the network for 100 epochs using an initial learning rate of 0.1. The learning rate is decayed at the 60^{th} and 80^{th} epochs by 0.1. In addition, as stated in works [12, 21, 38], ResNet-50 is used for iNaturalist 2018, while for Places-LT, we utilize the ResNet-152 pre-trained on imageNet. We train the networks for 200 epochs using an initial learning rate of 0.1, and decay the learning rate at epochs 75 and 160 by 0.1. We use Stochastic Gradient Descent (SGD) with a momentum of 0.9 and weight decay of 2×10^{-4} as the optimizer to train all models.

4.2. Comparison with the state of the arts

We compare SMART with the state-of-the-art long-tail recognition methods, *e.g.*, ACE [5], GCL [28], CMO [45], GCL+CR [39] and PC [50] on CIFAR10/10-LT and CIFAR10/100-LT. Table 1 shows that SMART consistently outperforms all prior methods across all imbalance ratios in both datasets. This demonstrates that SMART can be applied to different datasets and imbalance ratios. Our designed loss function is build upon the basic cross-entropy (CE) loss with pairwise semantic bias. Replacing CE with balanced softmax cross-entropy (BSCE) [46] can further boost performance. The gap between classical CE and class-balanced BSCE is smaller than that between image-level augmentation CMO [45], indicating that our augmentation strategy is more effective for long-tailed recognition tasks. Moreover, our proposed method even surpasses the performance of previous state-of-the-art models that use an ensemble of multiple experts. TLC [27] employs four experts for prediction, which results in higher computational complexity. In contrast, our method achieves better results with only a single model for evaluation, without the need for any extra computation.

Table 1. Comparisons on CIFAR100-LT and CIFAR10-LT datasets with the IF of 100 and 50. † indicates the ensemble performance is reported.

Method	Ref.	CIFAR100-LT		CIFAR10-LT	
		100	50	100	50
CB Focal loss [13]	CVPR'19	38.7	46.2	74.6	79.3
LDAM+DRW [6]	NeurIPS'19	42.0	45.1	77.0	79.3
LDAM+DAP [20]	CVPR'20	44.1	49.2	80.0	82.2
BBN [69]	CVPR'20	39.4	47.0	79.8	82.2
LFME [59]	ECCV'20	42.3	–	–	–
CAM [66]	AAAI'21	47.8	51.7	80.0	83.6
Logit Adj. [40]	ICLR'21	43.9	–	77.7	–
LDAM+M2m [22]	CVPR'21	43.5	–	79.1	–
MiSLAS [68]	CVPR'21	47.0	52.3	82.1	85.7
LADE [18]	CVPR'21	45.4	50.5	–	–
Hybrid-SC [55]	CVPR'21	46.7	51.9	81.4	85.4
CE+MetaSAug [30]	CVPR'21	46.9	51.9	80.1	84.0
DiVE [17]	ICCV'21	45.4	51.3	–	–
SSD [32]	ICCV'21	46.0	50.5	–	–
PaCo [12]	ICCV'21	52.0	56.0	–	–
xERM [71]	AAAI'22	46.9	52.8	–	–
RISDA [10]	AAAI'22	50.2	53.8	79.9	84.2
GCL [28]	CVPR'22	48.7	53.6	82.7	85.5
BCL [74]	CVPR'22	51.9	56.6	84.3	87.2
CE+CMO [45]	CVPR'22	43.9	48.3	–	–
BSCE+CMO [45]	CVPR'22	46.6	51.4	–	–
RIDE (3 experts)† [56]	ICLR'21	49.1	–	–	–
ACE (4 experts)† [5]	ICCV'21	49.6	51.9	81.4	84.9
TLC (4 experts)† [27]	CVPR'22	49.8	–	80.4	–
RIDE+CMO+CR [39]	CVPR'23	50.7	54.3	–	–
GCL+CR [39]	CVPR'23	–	–	83.5	86.8
PC [50]	IJCAI'23	53.4	57.8	–	–
SMART	–	55.4	60.4	84.0	87.2
BSCE+SMART	–	**56.1**	**61.2**	**84.5**	**87.9**

Table 2 presents comparisons on ImageNet-LT and Places-LT. The significant improvements demonstrate that our semantic augmentation for tail classes mitigates overfitting and enhances the generalization ability of the long-tailed learner, even on large-scale imbalanced datasets. Furthermore, when combined with BSCE, it consistently improves performance, demonstrating the versatility of our method and potential to seamlessly combine it with state-of-the-art long-tailed recognition methods. Lastly, applying SMART to ACE further boosts performance, outperforming the results of ACE with three experts. On the naturally-skewed dataset, as shown in Table 3, applying SMART to the simple training scheme surpasses the state of the arts as our performance improvement is not dependent on multi-stage training [21, 32] or post-processing [40, 61].

Table 2. Comparisons on ImageNet-LT and Places-LT datasets. † indicates the ensemble performance is reported.

Method	Ref.	ImageNet-LT Res50	ImageNet-LT ResX50	Places-LT Res152
OLTR [38]	CVPR'19	–	–	35.9
BBN [69]	CVPR'20	48.3	49.3	–
NCM [21]	ICLR'20	44.3	47.3	36.4
cRT [21]	ICLR'20	47.3	49.6	36.7
τ-norm [21]	ICLR'20	46.7	49.4	37.9
LWS [21]	ICLR'20	47.7	49.9	37.6
BSCE [46]	NeurIPS'20	–	–	38.7
DisAlign [61]	CVPR'21	52.9	–	–
DiVE [17]	ICCV'21	53.1	–	–
SSD [32]	ICCV'21	–	56.0	–
PaCo [12]	ICCV'21	57.0	58.2	41.2
ALA Loss [67]	AAAI'22	52.4	53.3	40.1
xERM [71]	AAAI'22	–	54.1	39.3
RISDA [10]	AAAI'22	50.7	–	–
MBJ [36]	AAAI'22	–	52.1	38.1
WD [1]	CVPR'22	53.9	–	–
GCL [28]	CVPR'22	54.9	–	40.6
BCL [74]	CVPR'22	56.0	57.1	–
CE+CMO [45]	CVPR'22	49.1	–	–
BS+CMO [45]	CVPR'22	52.3	–	–
ACE (3 experts)† [5]	ICCV'21	54.7	56.6	–
RIDE (3 experts)† [56]	ICLR'21	55.4	56.8	–
MBJ+RIDE (4 experts)† [36]	AAAI'22	–	57.7	–
TLC (4 experts)† [27]	CVPR'22	55.1	–	–
RIDE (4 experts)†+CR [39]	CVPR'23	–	57.8	–
PC [50]	IJCAI'23	54.9	–	–
SMART	–	57.8	58.2	41.0
BSCE+SMART	–	58.1	58.6	41.3
ACE (3 experts)†+SMART	–	**59.2**	**59.4**	**42.6**

Table 3. Comparisons on iNaturalist 2018 dataset with ResNet-50. † indicates the ensemble performance is reported.

Method	Ref.	Accuracy
OLTR [38]	CVPR'19	63.9
BBN [69]	CVPR'20	66.3
DAP [20]	CVPR'20	67.6
cRT [21]	ICLR'20	65.2
τ-norm [21]	ICLR'20	65.6
LWS [21]	ICLR'20	65.9
LDAM+DRW [6]	NeurIPS'19	68.0
Logit Adj. [40]	ICLR'21	66.4
CAM [66]	AAAI'21	70.9
PaCo [12]	ICCV'21	73.2
ALA Loss [67]	AAAI'22	70.7
xERM [71]	AAAI'22	67.3
RISDA [10]	AAAI'22	69.1
MBJ [36]	AAAI'22	70.0
WD [1]	CVPR'22	70.2
GCL [28]	CVPR'22	72.0
BCL [74]	CVPR'22	71.8
CE+CMO [45]	CVPR'22	68.9
BSCE+CMO [45]	CVPR'22	70.9
RIDE (3 experts)† [56]	ICLR'21	72.6
ACE (3 experts)† [5]	ICCV'21	72.9
RIDE (4 experts)†+CR [39]	CVPR'23	73.5
PC [50]	IJCAI'23	70.6
SMART	–	72.7
BSCE+SMART	–	73.1
ACE (3 experts)†+SMART	–	**74.3**

Overall, SMART improves previous methods on CIFAR10-LT , CIFAR100-LT, ImageNet-LT, Places-LT and iNaturalist2018 with accuracies of 84.5% (IF of 100), 56.1% (IF of 100), 59.2% (with ResNet-50), 42.6% and 74.3%, respectively. The significant performance over the state of the art underscores the efficacy of our SMART.

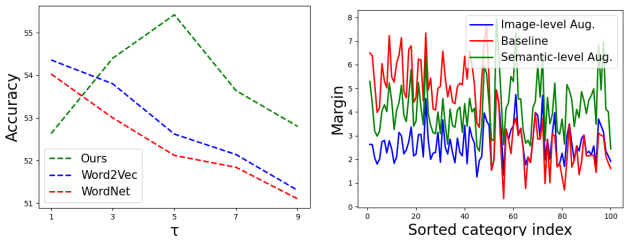
4.3. Ablation Study

In order to gain an understanding of the effectiveness of various components of SMART, various experiments are conducted on the CIFAR100-LT dataset with an IF of 100, as shown in Table 4. Specifically, several variants are considered: (1) 'Baseline model' trained with the CE loss, (2) 'Minority Aug.' adapting the semantic data augmentation on minority samples, (3) '$\mathcal{W}$' weighting imbalanced samples using the semantic-aware bias, and (4) combining with the class-aware loss 'BSCE'.

Impact of semantic data augmentation. The impact of semantic data augmentation is evident in the superior performance of the 'Minority Aug.' compared to the baseline model, achieving an accuracy of 38.62% *vs*. 52.12%, respectively. This improvement can be attributed to two key factors: 1) our method diversifies data points around decision boundaries. Consequently, it bolsters the generalizability, particularly for the minority classes, and 2) by explicitly transferring information from head samples to tail samples, thus inducing semantic similarity, the risk of misclassification during training is reduced. This effectively avoids overfitting in the tail classes. Furthermore, our data augmentation technique effectively reinstates the accurate data distribution of minority classes, even without utilizing any specialized balanced loss function. The results show a 12.18% accuracy gap between the use of BSCE against CE loss. However, our augmentation approach significantly reduces this gap to just 2.69%.

Impact of semantic-aware weighting. BSCE helps address a severe class imbalance by adjusting losses based on prior class distributions, thereby providing a balancing cue. We propose an alternative solution based on pairwise semantic relationships. Our strategy assigns higher weights to hard-tail samples that exhibit the highest semantic correlation with head features. This approach subsequently leads to heightened penalization during the training phase. In essence, our model undergoes training through a semantic-aware optimization mechanism. The comparison results are presented in the third panel of Table 4, where the performance scores are similar with each other (50.80% *vs*. 50.73%). Re-weighing BSCE according to semantic cues further boosts accuracy, demonstrating their synergy (50.80% *vs*. 53.89%).

The meaningful semantic directions. Building on the discussion in §3.2, we employ WordNet as an auxiliary tool



(a) (b)

Figure 4. The variation in scaling hyperparameter τ, contrasting adaptive online *vs*. fixed offline semantic sampling scheme (4a), and visualization on the classification margin (4b).

to define the semantic relationships between classes. Moreover, we also turn to Word2Vec [11], generating word embeddings for each class name. Using a cosine function, we assess pairwise similarities based on these embeddings. Leveraging the ranked class similarity, we randomly pick one from the top-5 head classes that closely aligns with each tail class. The head classes provide valuable semantic guidance and facilitate diversity transfer to the tail classes. Referencing Figure 4a, these offline methods require a minor scaling hyperparameter τ. In contrast, our online strategy demands a greater τ to yield a sharper distribution from the softmax. This constrains search directions while also ensures flexibility, allowing for the discovery of meaningful semantic directions.

Visualization of the margin. We empirically show that SMART achieves larger margins in majority classes compared to both the baseline and image-level augmentation [45]. We plot the average margin of each class. As depicted in Figure 4b, the baseline margins exhibit an imbalance: tail classes have notably smaller margins than head classes. While the image-level augmentation approach seems to decrease the margins of head classes in an attempt to balance them out, our method distinctively enlarge the margins of tail classes, which in turn leads to improved performance and generalization ability.

Types of transformation matrix. The formulation in Eq. 5 and Appendix §A.4 suggest that the transfer matrix, which is derived from the covariance of both tail and head samples, can guarantee optimal transformation. To validate this assertion, we examined scenarios with covariances calculated from either head instances, tail instances, or both combined. The resulting scores for these scenarios were 53.14%, 50.71%, and 55.42%, respectively, highlighting the advantage of our combined approach.

Impact of different layers on covariance. Each layer of feature extractor is responsible for capturing and processing specific feature patterns inherent in the input data. As data processes through these layers, the covariance will change across layers. This means the choice of layers plays a pivotal role in determining the effectiveness of the covariance. Table 5 shows that the 4^{th} layer yields the best result.

Table 4. Ablation results w.r.t. minority augmentation (Minority Aug.), semantic-aware weights ($\mathcal{W}$), the basic cross entropy loss (CE) and the balanced softmax cross-entropy loss (BSCE).

Minority Aug.	$\mathcal{W}$	CE	BSCE	Accuracy
		✓		38.62
✓		✓		52.12
✓			✓	54.81
			✓	50.80
	✓	✓		50.73
	✓		✓	53.89
✓	✓	✓		55.42
✓	✓		✓	56.13

Table 5. Impact on the different layers of feature extractor and the EMA momentum (γ) on covariance estimation.

#Layer	Accuracy	γ	Accuracy
1	48.57	0.01	36.21
2	51.72	0.05	35.32
3	52.98	0.10	35.62
4	55.42	Dynamic	55.42

Impact of EMA momentum. In situations where a class is sparsely represented, accurately capturing its data covariance and identifying the correct semantic direction can be challenging. To counter this, EMA is often employed to stabilize the estimation of covariance matrices across iterations. With EMA, the covariance estimates can be updated gradually, mitigating the effect of severely data imbalances or noise. EMA improves the robustness of covariance estimates and reduces their sensitivity to short-term variations, provided the momentum parameter γ is appropriately chosen. Table 5 shows that the fixed preset value for γ can be biased by long-tailed distributions, while dynamic adjustment based on sample frequency performs best.

5. Conclusions

We have introduced SeMAntic tRansfer from head to Tail (SMART), a novel data augmentation method that enhances the model generalization in long-tailed scenarios. Our thorough theoretical analysis and comprehensive experimental ablations, attest that SMART enjoys reasonable design and confirm its efficacy in enlarging the margin for tail classes. This highlights the effectiveness of SMART in improving the generalization ability in imbalanced settings. SMART performs well across five popular benchmarks and consistently outperforms existing methods. Thus, SMART presents a valuable approach to countering the pitfalls of learning with long-tailed distribution.

Acknowledgments

This work is supported by CSIRO's Science Digital, and by Academic Research Projects of Beijing Union University (No.ZK 20202302).

References

[1] Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, and Shu Kong. Long-tailed recognition via weight balancing. In *CVPR*, 2022. 7

[2] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3(Nov):463–482, 2002. 3

[3] Yoshua Bengio, Grégoire Mesnil, Yann Dauphin, and Salah Rifai. Better mixing via deep representations. In *ICCV*, pages 552–560. PMLR, 2013. 2

[4] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 2018. 1, 2

[5] Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *ICCV*, 2021. 6, 7

[6] 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, 6, 7

[7] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *JAIR*, 16:321–357, 2002. 1, 2

[8] Qiang Chen, Xiaokang Chen, Jian Wang, Shan Zhang, Haocheng Feng, Junyu Han, Errui Ding, Gang Zeng, and Jingdong Wang. Group detr: Fast detr training with group-wise one-to-many assignment. 2022. 1

[9] Qiang Chen, Jian Wang, Chuchu Han, Shan Zhang, Zexian Li, Xiaokang Chen, Jiahui Chen, Xiaodi Wang, Shuming Han, Gang Zhang, et al. Group detr v2: Strong object detector with encoder-decoder pretraining. *arXiv preprint arXiv:2211.03594*, 2022. 1

[10] Xiaohua Chen, Yucan Zhou, Dayan Wu, Wanqian Zhang, Yu Zhou, Bo Li, and Weiping Wang. Imagine by reasoning: A reasoning-based implicit semantic data augmentation for long-tailed classification. In *AAAI*, 2022. 6, 7

[11] Kenneth Ward Church. Word2vec. *Natural Language Engineering*, 23(1):155–162, 2017. 8

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

[13] 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, 5, 6

[14] Lixi Deng, Jingjing Chen, Qianru Sun, Xiangnan He, Sheng Tang, Zhaoyan Ming, Yongdong Zhang, and Tat Seng Chua. Mixed-dish recognition with contextual relation networks. In *ACM MM*, pages 112–120, 2019. 1

[15] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[17] Yin-Yin He, Jianxin Wu, and Xiu-Shen Wei. Distilling virtual examples for long-tailed recognition. In *ICCV*, 2021. 6, 7

[18] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *CVPR*, 2021. 6

[19] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *CVPR*, 2016. 1, 2

[20] 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, 7

[21] 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. 6, 7

[22] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *CVPR*, 2020. 2, 6

[23] Frank Klinker. Exponential moving average versus moving exponential average. *Mathematische Semesterberichte*, 58:97–107, 2011. 3

[24] Jedrzej Kozerawski, Victor Fragoso, Nikolaos Karianakis, Gaurav Mittal, Matthew Turk, and Mei Chen. Blt: Balancing long-tailed datasets with adversarially-perturbed images. In *ACCV*, 2020. 2

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

[26] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. 1

[27] Bolian Li, Zongbo Han, Haining Li, Huazhu Fu, and Changqing Zhang. Trustworthy long-tailed classification. In *CVPR*, 2021. 6, 7

[28] Mengke Li, Yiu-ming Cheung, and Yang Lu. Long-tailed visual recognition via gaussian clouded logit adjustment. In *ICCV*, pages 6929–6938, 2022. 6, 7

[29] Mengke Li, Zhikai Hu, Yang Lu, Weichao Lan, Yiu-ming Cheung, and Hui Huang. Feature fusion from head to tail: an extreme augmenting strategy for long-tailed visual recognition. *arXiv preprint arXiv:2306.06963*, 2023. 3

[30] Shuang Li, Kaixiong Gong, Chi Harold Liu, Yulin Wang, Feng Qiao, and Xinjing Cheng. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In *CVPR*, pages 5212–5221, 2021. 3, 6

[31] 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 *CVPR*, pages 6918–6928, 2022. 2

[32] Tianhao Li, Limin Wang, and Gangshan Wu. Self supervision to distillation for long-tailed visual recognition. In *ICCV*, 2021. 6, 7

[33] Dekang Lin et al. An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304, 1998. 4

[34] 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 *ECCV*, 2014. 2

[35] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *CVPR*, pages 2970–2979, 2020. 1, 2, 3

[36] Jialun Liu, Jingwei Zhang, Wenhui Li, Chi Zhang, and Yifan Sun. Memory-based jitter: Improving visual recognition on long-tailed data with diversity in memory. In *AAAI*, 2022. 7

[37] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, pages 212–220, 2017. 5, 12

[38] 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, 6, 7

[39] Yanbiao Ma, Licheng Jiao, Fang Liu, Shuyuan Yang, Xu Liu, and Lingling Li. Curvature-balanced feature manifold learning for long-tailed classification. In *CVPR*, pages 15824–15835, 2023. 6, 7

[40] 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. 2, 6, 7

[41] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 4

[42] Yao Ni and Piotr Koniusz. NICE: NoIse-modulated Consistency rEgularization for Data-Efficient GANs. In *NeurIPS*, 2023. 1

[43] Yao Ni, Piotr Koniusz, Richard Hartley, and Richard Nock. Manifold learning benefits GANs. In *CVPR*, pages 11265–11274, 2022. 1

[44] Yao Ni, Dandan Song, Xi Zhang, Hao Wu, and Lejian Liao. Cagan: consistent adversarial training enhanced gans. In *IJCAI*, pages 2588–2594, 2018. 1

[45] Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Young Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *CVPR*, pages 6887–6896, 2022. 1, 2, 6, 7, 8

[46] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *NeurIPS*, 33:4175–4186, 2020. 1, 2, 3, 6, 7

[47] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *ICML*, 2018. 1, 2

[48] Michael S Ryoo, AJ Piergiovanni, Mingxing Tan, and Anelia Angelova. Assemblenet: Searching for multi-stream neural connectivity in video architectures. *arXiv preprint arXiv:1905.13209*, 2019. 1

[49] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation. In *ECCV*, pages 680–697, 2018. 1

[50] Saurabh Sharma, Yongqin Xian, Ning Yu, and Ambuj Singh. Learning prototype classifiers for long-tailed recognition. *arXiv preprint arXiv:2302.00491*, 2023. 6, 7

[51] Cecilia Summers and Michael J Dinneen. Improved mixed-example data augmentation. In *WACV*, pages 1262–1270. IEEE, 2019. 2

[52] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 1

[53] 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. 5, 6

[54] Lei Wang and Piotr Koniusz. Self-supervising action recognition by statistical moment and subspace descriptors. In *ACM MM*, MM '21, page 4324–4333, New York, NY, USA, 2021. Association for Computing Machinery. 1

[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. 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. 6, 7

[57] Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, and Cheng Wu. Implicit semantic data augmentation for deep networks. *NeurIPS*, 32, 2019. 1, 3

[58] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *NeurIPS*, 2017. 2

[59] Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *ECCV*, 2020. 6

[60] Shiran Zada, Itay Benou, and Michal Irani. Pure noise to the rescue of insufficient data: Improving imbalanced classification by training on random noise images. In *ICML*, pages 25817–25833. PMLR, 2022. 2

[61] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *CVPR*, 2021. 2, 6, 7

[62] Shan Zhang, Dawei Luo, Lei Wang, and Piotr Koniusz. Few-shot object detection by second-order pooling. In *ACCV*, 2020. 2

[63] Shan Zhang, Naila Murray, Lei Wang, and Piotr Koniusz. Time-reversed diffusion tensor transformer: A new tenet of few-shot object detection. In *ECCV*, pages 310–328. Springer, 2022. 2

[64] Shan Zhang, Lei Wang, Naila Murray, and Piotr Koniusz. Kernelized few-shot object detection with efficient integral aggregation. In *CVPR*, pages 19207–19216, 2022. 2

[65] Shan Zhang, Tianyi Wu, Sitong Wu, and Guodong Guo. Catrans: context and affinity transformer for few-shot segmentation. 2022. 2

[66] 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, 7

[67] Yan Zhao, Weicong Chen, Xu Tan, Kai Huang, and Jihong Zhu. Adaptive logit adjustment loss for long-tailed visual recognition. In *AAAI*, 2022. 7

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

[69] 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. 6, 7

[70] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE TPAMI*, 2017. 5, 6

[71] Beier Zhu, Yulei Niu, Xian-Sheng Hua, and Hanwang Zhang. Cross-domain empirical risk minimization for unbiased long-tailed classification. In *AAAI*, 2022. 6, 7

[72] Hao Zhu and Piotr Koniusz. EASE: Unsupervised discriminant subspace learning for transductive few-shot learning. In *CVPR*, pages 9078–9088, June 2022. 1

[73] Hao Zhu, Ke Sun, and Peter Koniusz. Contrastive laplacian eigenmaps. *NeurIPS*, 34:5682–5695, 2021. 1

[74] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *CVPR*, pages 6908–6917, 2022. 1, 2, 6, 7