Long-Tailed Multi-Label Visual Recognition by Collaborative Training on Uniform and Re-balanced Samplings

Hao Guo, Song Wang
University of South Carolina, Columbia, SC 29201, US
hguo@email.sc.edu; songwang@cec.sc.edu

Abstract

Long-tailed data distribution is common in many multi-label visual recognition tasks and the direct use of these data for training usually leads to relatively low performance on tail classes. While re-balanced data sampling can improve the performance on tail classes, it may also hurt the performance on head classes in training due to label co-occurrence. In this paper, we propose a new approach to train on both uniform and re-balanced samplings in a collaborative way, resulting in performance improvement on both head and tail classes. More specifically, we design a visual recognition network with two branches: one takes the uniform sampling as input while the other takes the re-balanced sampling as the input. For each branch, we conduct visual recognition using a binary-cross-entropy-based classification loss with learnable logit compensation. We further define a new cross-branch loss to enforce the consistency when the same input image goes through the two branches. We conduct extensive experiments on VOC-LT and COCO-LT datasets. The results show that the proposed method significantly outperforms previous state-of-the-art methods on long-tailed multi-label visual recognition.

1. Introduction

By classifying an image into multiple classes, multi-label visual recognition is an important task in computer vision and the state-of-the-art approaches [45, 54, 44, 2, 20, 21, 12, 40, 52, 5] are to train deep networks on a set of training data with ground-truth labels. However, as in many single-label recognition tasks [25, 3, 6, 41, 24, 15, 16, 53], the training data of multi-label recognition may exhibit a long-tailed distribution [39] in terms of class labels – *head classes* have many samples while *tail classes* have very few samples. Direct training on such data (with uniform sampling) usually produces relatively low performance on the tail classes. In this paper, we focus on solving the problem of long-tailed multi-label visual recognition (LTML).

Re-balanced data sampling [4, 32, 1, 10] is a proven effective approach for addressing the long-tailed visual recognition. It achieves class-wise balance by either down-sampling the head-class data or up-sampling the tail-class data. However, repeating/dropping a tail-class/head-class image may also duplicate/remove head-class/tail-class samples due to label co-occurrence in multi-label recognition [47]. Thus, while re-balanced sampling can improve the recognition performance of tail classes, it may simultaneously decrease the performance of some head classes for LTML. Since performance of different classes, either head or tail ones, is usually considered to be equally important in multi-label visual recognition, in this paper, we develop a new method that can combine different data samplings for improving the performance of both head and tail classes.

We consider the uniform and re-balanced samplings. Given a long-tailed training set for multi-label recognition, the uniform sampling leads to the original long-tailed distribution, while the re-balanced sampling expects to achieve a balanced distribution, but yields another biased distribution due to label-occurrence. Our basic idea is to use each of them to train a branch of a two-branch network, where two branches follow the same architecture. We further define a loss that enforces the consistency across the two branches for the same input to achieve a collaborative training, inspired by the previous mutual learning [51] and co-regularization [27]. The cross-branch consistency compromises two distributions to achieve an effect equivalent to learning the proposed network from a balanced implicit distribution somewhere between two biased distributions from different samplings.

More specifically, as shown in Fig. 1(b), the two branches have the same architecture but different parameters to reflect the different distributions of their respective inputs. For each branch, a binary-cross-entropy-based multi-label classification loss with learnable logit compensation is defined for LTML. For combining two branches, we introduce another loss to collaboratively enforce the prediction consistency across the two branches when the same input image is fed to the two branches. Finally, this two-

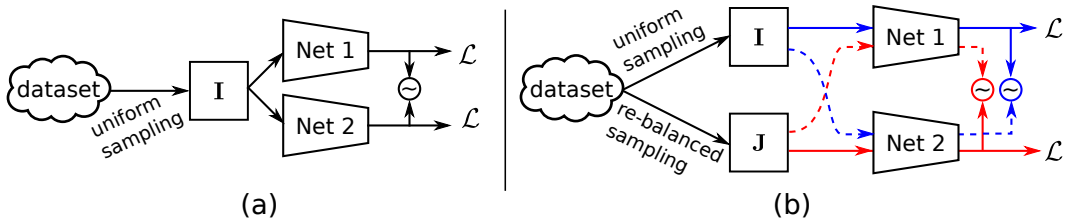

Figure 1. An illustration of the difference between (a) the previous mutual learning [51]/co-regularization [27] networks, where the input from the same distribution is always fed to the two branches, and (b) the proposed network where different inputs, from different samplings, are fed to the two branches. We only use the same input for the two branches for computing the consistency loss. **I** and **J** are mini-batch images, $\sim$ indicates the consistency measurement, and $\mathcal{L}$ is the classification loss.

branch network is trained in an end-to-end manner by minimizing both classification and consistency losses. During the test phase, each test image is fed to both branches without considering cross-branch paths and the average of predictions from the two branches is taken as the final prediction.

Different from previous mutual learning methods [51, 27], where the two branches always take the input from a single distribution, as shown in Fig. 1(a), our proposed method learns two branches from different inputs generated by different samplings and the same input for two branches is only used for computing the consistency loss.

To summarize, the main contributions of this work are:

1 We propose the use of both uniform and re-balanced samplings of the same training set for long-tailed multi-label visual recognition.

2 We develop a two-branch network, as well as a cross-branch loss to enforce the consistency between two branches, for collaborative learning on both uniform and re-balanced samplings.

3 We conduct extensive experiments on VOC-LT and COCO-LT datasets to verify that the proposed method can simultaneously improve the performance of both head and tail classes.

2. Related Work

2.1. Multi-label Visual Recognition

In many traditional methods, multi-label visual recognition is reduced to multiple binary image classifications [38, 50] or finding k-nearest neighbors [49]. As CNNs [17, 34, 35, 11, 14] become a standard component in vision systems, many deep-learning based methods have been developed for multi-label visual recognition and they can be generally categorized into two main groups: label-localization methods and label-correlation methods. Label-localization methods [45, 54, 44, 8] attempt to localize the label-related image regions using either supervised learning on manual annotations or weakly supervised learning on class labels. Label-correlation methods [2, 20, 21, 12, 40, 52, 5] improve multi-label visual recognition by exploiting and leveraging the co-occurrence of different labels in the same image. For examples, CNN-RNN [40] combines RNNs with CNNs to learn the correlations between different labels. ML-GCN [5] adopts Graph Convolutional Networks (GCN) to embed the label correlations to the classifier learning. When the training set is long-tailed, head classes usually dominate the network training, resulting in inaccurate label localization and label correlations for tail classes, which severely hurts the recognition performance on tail classes.

2.2. Re-balancing Long-Tailed Visual Recognition

Data re-balancing is a widely used strategy for handling long-tailed visual recognition, by emphasizing tail classes more in the network learning, and it has achieved improved results on many long-tailed recognition tasks. Re-balanced sampling [4, 32, 1, 10, 53] and cost sensitive re-weighting [3, 6, 13, 43, 29, 19, 36] are the two typical kinds of data re-balancing methods. The former improves the class balance by either up-sampling the tail classes or down-sampling the head classes, while the latter improves the class balance by weighting more on tail classes in the loss functions. However, all these methods are for single-label recognition, i.e., each image only has one label. Wu et al. [47] extend re-balanced sampling and cost-sensitive re-weighting methods to handle long-tailed multi-label visual recognition and propose an optimized DB Focal method, which does improve the recognition performance of tail classes. However, because of label co-occurrence in multi-label recognition, emphasizing the tail classes may impair the head-class training. The re-balanced sampling may simultaneously decrease the performance of some head classes [47]. In this paper, we propose to collaboratively train on uniform and re-balanced samplings to improve the performance on both head and tail classes.

2.3. Network Consistency

In this paper, we use the consistency between two branches to collaboratively train the model for the multi-label visual recognition. Different kinds of network consistency have been considered for improving network training

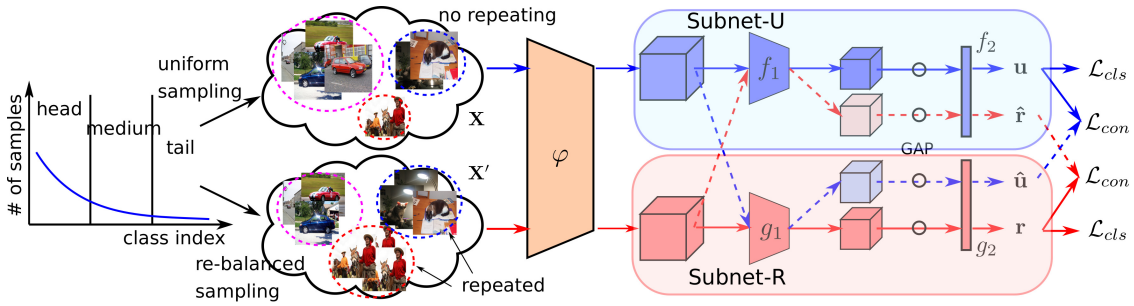

Figure 2. An illustration of the proposed network for long-tailed multi-label visual recognition. GAP denotes the global average pooling.

in different tasks. *Perturbation-based consistency* requires a trained network to produce same prediction after applying a small perturbation to the input image [33, 31, 26, 46, 8, 42] and it has been widely used for data augmentation [33]. *Model-based consistency* [18, 48, 51, 28, 27] is usually formulated and applied between networks. It enforces the two different networks to produce the same results when the same image is taken as the input, as shown in Fig. 1(a). Examples include Π-model [18] and Mean Teacher [37] used for semi-supervised learning, deep mutual learning [51] and co-regularization [27] for training two networks collaboratively, and co-teaching [9] for handling noisy labels. However, by taking the input from the same distribution, two branches trained in [51, 27] may collapse to each other if their network parameters are not carefully initialized with substantial difference. In [28], an adversarial scheme is introduced to address this issue, while in this paper, we use images of different samplings as the inputs of the two branches, which adds diversity to each branch training.

3. Proposed Method

Let the training set for the long-tailed multi-label visual recognition (LTML) be $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ are the N training images and $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N\}$ are their respective ground-truth class labels. Specifically, each $\mathbf{y}_i = [y_{i1}, y_{i2}, \cdots, y_{iK}], i = 1, 2, \cdots, N$ is a binary K-dimensional vector where $y_{ik} = 1$ indicates the presence of label k in image i and $y_{ik} = 0$ otherwise, with $k = 1, 2, \cdots, K$. K is the total number of labels for the visual recognition. There may be multiple elements of value 1 in each $\mathbf{y}_i$ for multi-label visual recognition.

3.1. Framework Overview

Given that $(\mathbf{X}, \mathbf{Y})$ follows a long-tail distribution in terms of class labels, we use both uniform and re-balanced samplings in preparing the inputs for network training. For the uniform sampling, each image $\mathbf{x}_i \in \mathbf{X}$ is sampled with an instance-level probability of $1/N$. For the re-balanced sampling [32, 16, 47], images of each class are sampled with a class-level probability of $\frac{1}{K}$, and thus, each image $\mathbf{x}_i$ is sampled with a probability of $\frac{1}{K}\sum_{k=1}^{K}\frac{y_{ik}}{N_k}$, where N_k is the number of images with class label k in the training set. By sampling the original training set M times, the re-balanced sampling actually provides us a new relatively class-balanced training set $(\mathbf{X}', \mathbf{Y}')$, with M samples, but not real balanced due to label co-occurrence.

As shown in Fig. 2, the two branches of the proposed network share the same bottom network φ, followed by another CNN module, denoted as 'Subnet-U' in the branch for the uniform sampling and 'Subnet-R' in the branch for the re-balanced sampling. Subnet-U and Subnet-R have the same architecture but trained with different parameters, as shown in Fig. 2. To be specific, the shared bottom network is the conventional ResNet [11] excluding the last stage. For Subnet-U and Subnet-R, we first include an identical copy of the last stage of ResNet, as shown by f_1 and g_1 in Fig. 2. After that, a linear classifier in the form of a fully connected layer is added to each branch, as shown by f_2 and g_2 in Fig. 2 for multi-label recognition. When feeding images $\mathbf{x}_i^u \in \mathbf{X}$ and $\mathbf{x}_j^r \in \mathbf{X}'$ to the two branches respectively, we obtain K-dimensional logits for the two branches as

$$\begin{cases} \mathbf{u}_i = f_2(f_1(\varphi(\mathbf{x}_i^u))), \\ \mathbf{r}_j = g_2(g_1(\varphi(\mathbf{x}_j^r))). \end{cases} \tag{1}$$

By formulating the task as multiple binary image classifications, we apply logistic linear regression on logits $\mathbf{u}_i \in \mathbb{R}^K$ and $\mathbf{r}_j \in \mathbb{R}^K$ to learn the two branches, respectively. The solid arrows in blue and red in Fig.2 indicate the classification paths for the two branches, respectively. The binary-cross-entropy-based classification losses $\mathcal{L}_{cls}(\mathbf{u}_i, \mathbf{y}_i^u)$ and $\mathcal{L}_{cls}(\mathbf{r}_j, \mathbf{y}_j^r)$ are adopted for respective branch optimization, where $(\mathbf{u}_i, \mathbf{y}_i^u)$ and $(\mathbf{r}_j, \mathbf{y}_j^r)$ represent the pair of predicted logits and ground-truth labels for the i-th image in $\mathbf{X}$ and the j-th image in $\mathbf{X}'$, respectively.

We further cross the inputs of two branches and estimate the logits, indicated by the blue/red dashed arrows in Fig. 2 and obtain

$$\begin{cases} \hat{\mathbf{u}}_i = g_2(g_1(\varphi(\mathbf{x}_i^u))), \\ \hat{\mathbf{r}}_j = f_2(f_1(\varphi(\mathbf{x}_j^r))). \end{cases} \tag{2}$$

To enforce the two branches to make consistent predictions from the same input, we introduce a mean-square-

error based consistency loss $\mathcal{L}_{con}(\mathbf{u}_i, \hat{\mathbf{u}}_i)$ and $\mathcal{L}_{con}(\mathbf{r}_j, \hat{\mathbf{r}}_j)$ between the logits from different branches, indicate by the same color arrows (one dashed and one solid) in Fig. 2.

Finally, the network is learned by jointly minimizing the loss function

$$\begin{aligned}\mathcal{L}(\mathbf{x}_i^u, \mathbf{x}_j^r; \mathbf{y}_i^u, \mathbf{y}_j^r) =& \mathcal{L}_{cls}(\mathbf{u}_i, \mathbf{y}_i^u) + \mathcal{L}_{cls}(\mathbf{r}_j, \mathbf{y}_j^r) + \\ & \lambda(\mathcal{L}_{con}(\mathbf{u}_i, \hat{\mathbf{u}}_i) + \mathcal{L}_{con}(\mathbf{r}_j, \hat{\mathbf{r}}_j)),\end{aligned} \tag{3}$$

where $(\mathbf{x}_i^u, \mathbf{y}_i^u) \in (\mathbf{X}, \mathbf{Y})$, $(\mathbf{x}_j^r, \mathbf{y}_j^r) \in (\mathbf{X}', \mathbf{Y}')$, and λ is a hyper-parameter to balance the two kinds of loss functions.

3.2. Conventional Classification Loss

Conventionally, the weighted sigmoid cross entropy loss [19, 8, 36] is used for multi-label visual recognition, in the form of multiple binary image classifications. Taking the branch for the uniform sampling as an example, this loss is

$$\begin{aligned}\mathcal{L}_{cls}(\mathbf{u}_i, \mathbf{y}_i^u) = -\frac{1}{K}\sum_{k=1}^{K} \omega_k \left(y_{ik}^u \log(\varsigma(u_{ik})) + \right. \\ \left. (1 - y_{ik}^u) \log(1 - \varsigma(u_{ik}))\right),\end{aligned} \tag{4}$$

where u_{ik} and y_{ik}^u are the k-th elements of the predicted logits $\mathbf{u}_i$ and the ground-truth label $\mathbf{y}_{ik}^u$, respectively, corresponding to the k-th label. Besides, $\omega_k = y_{ik}^u e^{1-\rho} + (1 - y_{ik}^u)e^{\rho}$ is the loss weight for the k-th label, depending on its ratio of positive samples $\rho = N_k/N$, and ς is the sigmoid function converting logits in $\mathbb{R}$ to probabilities in the range of [0, 1] by

$$\varsigma(u_{ik}) = 1/(1 + e^{-u_{ik}}). \tag{5}$$

The classification loss $\mathcal{L}_{cls}(\mathbf{r}_j, \mathbf{y}_j^r)$ for the other branch can be defined in the same way.

3.3. Logit Compensation

As discussed in [3, 47], when using the weighted sigmoid cross entropy loss for classification, the imbalance between the numbers of positive and negative samples in each class could push their unbounded logit values away from zero with different distances, leading to class-specific over-fitting. In this section, we address this issue by further compensating the logits of positive and negative samples, respectively.

For simplicity, we assume that logit output of the network for each label recognition conforms to a normal distribution. Suppose the logit for positive samples of the k-th label conforms to a normal distribution with mean μ_k^p and standard deviation σ_k^p, and the logit for negative samples of the same label conforms to a normal distribution with mean μ_k^n and standard deviation σ_k^n. The mean logit values $\{\mu_1^p, \mu_2^p, \cdots, \mu_K^p\}$ and $\{\mu_1^n, \mu_2^n, \cdots, \mu_K^n\}$, and standard deviations $\{\sigma_1^p, \sigma_2^p, \cdots, \sigma_K^p\}$ and $\{\sigma_1^n, \sigma_2^n, \cdots, \sigma_K^n\}$ are then used to compensate the logits before feeding to the classification loss in Eq. (4). Thus, the classification loss (4) is upgraded to

$$\begin{aligned}\mathcal{L}_{cls}(\mathbf{u}_i, \mathbf{y}_i^u) = -\frac{1}{K}\sum_{k=1}^{K} \omega_k \left(y_{ik}^u \log(\varsigma(u_{ik} \cdot \sigma_k^p + \mu_k^p)) \right. \\ \left. + (1 - y_{ik}^u) \log(1 - \varsigma(u_{ik} \cdot \sigma_k^n + \mu_k^n))\right).\end{aligned} \tag{6}$$

The classification loss $\mathcal{L}_{cls}(\mathbf{r}_j, \mathbf{y}_j^r)$ is upgraded with logit compensation in the same way. All the above means and standard deviations are learnable parameters. Compared with previous logit-adjustment methods [3, 47], this simple compensation does not introduce additional empirical hyper-parameters that require manually tuning.

3.4. Logit Consistency between Branches

In the ideal case, when we feed the same input image to the two branches, the output predictions shall approximate the ground-truth labels with the network optimizations. However, since the two branches attempt to fit the different distributions of input data, they may produce different prediction results with the same input, e.g., the two branches may show different recognition performance. As mentioned above, we define a cross-branch consistency loss based on the mean square error of logits computed from the same input image but through different branches. Taking the input from the uniform sampling as an example, this loss is

$$\mathcal{L}_{con}(\mathbf{u}_i, \hat{\mathbf{u}}_i) = \frac{1}{K}\sum_{k=1}^{K}(u_{ik} - \hat{u}_{ik})^2, \tag{7}$$

where u_{ik} and $\hat{u}_{ik}$ are the k-th elements of $\mathbf{u}_i$ and $\hat{\mathbf{u}}_i$, respectively. For the input from the re-balanced sampling, the consistency loss $\mathcal{L}_{con}(\mathbf{r}_j, \hat{\mathbf{r}}_j)$ can be defined in the same way.

Different from existing works on collaborative training [51, 27], which define consistency on probabilities, e.g., softmax/sigmoid outputs, for visual recognition, here we measure the consistency between logits of different branches from the same input. In training multi-label classifiers, due to the sigmoid normalization in Eq. (5), gradients could vanish on highly confident probabilities. For example, when the consistency loss is applied to probabilities, we have loss $\mathcal{L}_{con}(\varsigma(\mathbf{u}_i), \varsigma(\hat{\mathbf{u}}_i))$ and the gradients propagated to the logits $\mathbf{u}_i$ would be:

$$\begin{aligned}\frac{\partial \mathcal{L}_{con}(\varsigma(\mathbf{u}_i), \varsigma(\hat{\mathbf{u}}_i))}{\partial \mathbf{u}_i} &= \frac{\partial \mathcal{L}_{con}(\varsigma(\mathbf{u}_i), \varsigma(\hat{\mathbf{u}}_i))}{\partial \varsigma(\mathbf{u}_i)} \frac{\partial \varsigma(\mathbf{u}_i)}{\partial \mathbf{u}_i} \\ &= \frac{\partial \mathcal{L}_{con}(\varsigma(\mathbf{u}_i), \varsigma(\hat{\mathbf{u}}_i))}{\partial \varsigma(\mathbf{u}_i)} \varsigma(\mathbf{u}_i)(1 - \varsigma(\mathbf{u}_i)).\end{aligned} \tag{8}$$

If the predicted probabilities are highly confident, e.g. $\varsigma(\mathbf{u}_i) \simeq 1$ or $\varsigma(\mathbf{u}_i) \simeq 0$, the gradients from consistency loss

are close to zero. Differently, we define the consistency loss based on logits, with which the gradients propagated to the logits $\mathbf{u}_i$ would be:

$$\frac{\partial \mathcal{L}_{con}(\mathbf{u}_i, \hat{\mathbf{u}}_i)}{\partial \mathbf{u}_i} = \frac{2}{K}(\mathbf{u}_i - \hat{\mathbf{u}}_i). \quad (9)$$

We can see that these gradients do not have the above gradient vanishing issue under high-confident predictions.

3.5. Model Inference

To conduct model inference on test images, we simply feed all the test images to both branches of the trained network one by one. The paths following the dashed arrows in Fig. 2 are not used. For each input test image, the predictions of two branches are averaged as the final prediction result.

4. Experiments

4.1. Datasets and Configurations

As in [47], we conduct experiments on two datasets for long-tailed multi-label visual recognition: VOC-LT and COCO-LT. They are artificially constructed from two multi-label visual recognition benchmarks, VOC [7] and MS-COCO [23], respectively.

VOC-LT is sampled from the 2012 train-val set of VOC [7] based on a Pareto distribution as described in [25]. The training set contains 1,142 images and 20 class labels, and the number of images per class ranges from 4 to 775. The 20 classes are split into three groups according to the number of training samples per class: a head class has more than 100 samples, a medium class has 20 to 100 samples, and a tail class has less than 20 samples. The ratio of head, medium and tail classes after such splitting is 6:6:8. The testing set is constructed on the 2007 test set of VOC, with 4,952 images.

COCO-LT is created from the 2017 version of MS-COCO [23] by following a similar way. The training set of this long-tailed dataset contains 1,909 images and 80 class labels, and the number of images per class ranges from 6 to 1,128. The ratio of head, medium and tail classes is 22:33:25, following a similar split as in VOC-LT. The test set consists of all 5,000 images in the test set of MS-COCO-2017.

Configurations: Following [47] and the conventional multi-label visual recognition [54, 44, 8], we use the mean Average Precision (mAP) to evaluate the performance of long-tailed multi-label visual recognition. We use the similar configurations as in [47] in our experiments for a fair comparison with this prior state-of-the-art method. Specifically, we use the ResNet50 [10] pre-trained on ImageNet [17, 30] as the backbone and input images are resized to the spatial dimension of 224×224. The standard data augmentations are applied as in [47]. The SGD with momentum of 0.9 and weight decay of 0.0001 is adopted as the optimizer. The hyper-parameter λ in Eq. (3) is set to 0.1 constantly. In the classification loss with logit compensation in Eq. (6), the mean values are initialized to 0, while the standard deviations are initialized to 1. The initial learning rate is set to 0.01. All experiments are conducted on PyTorch 1.4.0.

4.2. Comparison with Prior Arts

First of all, to verify the effectiveness of the proposed method, we compare the mAP performance between our method and previous methods on both long-tailed datasets. The comparison methods include Empirical Risk Minimization (ERM), conventional Re-Weighting (RW) using the inverse proportion to the square root of class frequency, Re-Sampling (RS) [32], Focal Loss [22], ML-GCN [5], OLTR [25], LDAM [3], CB Focal [6], BBN [53] and DB Focal [47]. The mAP performance of different methods are shown in Table 1. The prior best performance is achieved by DB Focal [47] – mAP of 78.94% over all classes on VOC-LT and 53.55% over all classes on COCO-LT. We further reproduce DB Focal, denoted as DB Focal* in Table 1, on our platform based on its implementation [1] and achieve similar mAP performances as the ones reported in [47].

We train two baselines for the proposed method with the conventional classification loss and different samplings. Specifically, we train the proposed network only with one branch using the uniform sampling and re-balanced sampling, respectively, with the weighted classification loss in Eq. (4). This way, we obtain two baselines: baseline-uniform and baseline-re-balanced, respectively. From Table 1, we can see that both baselines achieve lower mAP performance than DB Focal (or DB Focal*) – mAP performances of two baselines on VOC-LT are 77.15% and 78.36%, respectively, and those on COCO-LT are 53.15% and 52.76%, respectively. The proposed method can significantly increase the mAP performance on both datasets: mAP performance is improved to 81.44% on VOC-LT (increased by 3.02% from DB Focal*) and to 56.90% on COCO-LT (increased by 2.63% from DB Focal*). Besides, the proposed method also achieves the new state-of-the-art mAP performance for both head, medium and tail classes on both datasets.

4.3. Quantitative Analysis

4.3.1 Ablation Analysis

To further analyze how the proposed method improves mAP performance for long-tailed multi-label recognition, we conduct a set of ablation studies and report the results in Table 2. We first conduct an experiment by using a simple

[1] https://github.com/wutong16/DistributionBalancedLoss

Datasets	VOC-LT				COCO-LT			
Methods	total	head	medium	tail	total	head	medium	tail
ERM	70.86	68.91	80.20	65.31	41.27	48.48	49.06	24.25
RW	74.70	67.58	82.81	73.96	42.27	48.62	45.80	32.02
Focal Loss [22] ICCV'17	73.88	69.41	81.43	71.56	49.46	49.80	54.77	42.14
RS [32] ECCV'16	75.38	70.95	82.94	73.05	46.97	47.58	50.55	41.70
ML-GCN [5] CVPR'19	68.92	70.14	76.41	62.39	44.24	44.04	48.36	38.96
OLTR [25] CVPR'19	71.02	70.31	79.80	64.95	45.83	47.45	50.63	38.05
LDAM [3] NeurIPS'19	70.73	68.73	80.38	69.09	40.53	48.77	48.38	22.92
CB Focal [6] CVPR'19	75.24	70.30	83.53	72.74	49.06	47.91	53.01	44.85
BBN* [53] CVPR'20	73.37	71.31	81.76	68.62	50.00	49.79	53.99	44.91
DB Focal [47] ECCV'20	78.94	73.22	84.18	79.30	53.55	51.13	57.05	51.06
DB Focal* [47] ECCV'20	78.42	74.13	83.19	78.06	54.33	50.06	57.22	54.27
baseline-uniform	77.15	73.14	83.49	75.41	53.15	51.61	57.17	49.21
baseline-re-balanced	78.36	71.72	83.58	79.41	52.76	48.67	56.87	50.94
Ours	**81.44**	**75.68**	**85.53**	**82.69**	**56.90**	**54.13**	**60.59**	**54.47**

Table 1. mAP performance of the proposed method and comparison methods. The notation * indicates the reproduced results based on our experiment environment. Other comparison results are taken from [47].

uniform branch	re-sampled branch	logit consistency	logit compensation	aug-test	VOC-LT				COCO-LT			
					total	head	medium	tail	total	head	medium	tail
√					77.15	73.14	83.49	75.41	53.15	51.61	57.17	49.21
	√				78.36	71.72	83.58	79.41	52.76	48.67	56.87	50.94
√	√				79.42	73.98	84.67	79.56	54.71	51.85	58.62	52.06
√	√	√			81.22	75.42	85.50	82.37	56.62	54.30	60.27	53.86
√	√	√	√		81.44	75.68	85.53	82.69	56.90	54.13	60.59	54.47
√	√	√	√	√	81.79	76.04	85.92	83.01	57.28	54.54	61.10	54.64

Table 2. Ablation analysis on different components of the proposed network.

branch-ensemble method which averages the predictions from the two branches as the final prediction, without considering the consistency and compensation. The achieved mAP performances are 79.42% on VOC-LT and 54.71% on COCO-LT, which are better than the two baselines. One possible reason is that the two branches learned from different label distributions exploit complementary information for recognizing the same label. By considering the proposed cross-branch consistency but not logit compensation, the mAP performance is improved to 81.22% on VOC-LT and 56.62% on COCO-LT, with 1.80% and 1.91% increments, respectively. Finally, we add the logit compensation to the classification loss, the mAP performance is further improved to 81.44% and 56.90%, respectively. This verifies that each component in the proposed method contributes to the mAP performance improvement.

Besides, we also show that incorporating an augmented testing (aug-test) strategy can further improve the mAP performance. In this strategy, the average of the predictions estimated from the original image and its horizontally flipped image is computed as the final prediction. Since this strategy is not widely used in the previous works, we do not consider it when comparing the performance of the proposed method against the previous methods.

4.3.2 Consistency Analysis

We also compare the proposed logit consistency across different training-data distributions with perturbation-based consistency and model-based consistency, as discussed in Sec. 2.3. The mAP performance from different logit consistency is reported in Table 3. Given a single data sampling, we add the perturbations of horizontal flipping as in VAC [8] on the input images and feed both original and perturbed images to the ResNet50 for model learning. The consistency of the estimated logits for the original and perturbed images is considered for multi-label recognition. The perturbation-based consistency based on uniform sampling and re-balanced sampling leads to mAP performance of 78.18% and 79.39% respectively on VOC-LT, and 55.32% and 55.49% respectively on COCO-LT. While the different data distributions are merged directly, i.e. "uniform ∪ re-balanced", to train the network without enforcing the logit consistency, the achieved mAP performance is much lower. This is equivalent to learn the model based on another distribution that combines the uniform and re-balanced samplings.

For model-based consistency, we train the two branches with the same sampling, either the uniform sampling or

number of branches	consistency based on	sampling	VOC-LT				COCO-LT			
			total	head	medium	tail	total	head	medium	tail
single	data perturbations	uniform	78.18	74.09	83.99	76.90	55.32	52.39	59.60	52.26
		re-balanced	79.39	73.35	84.71	79.94	55.49	52.01	59.32	53.50
	N/A	uniform ∪ re-balanced	77.85	72.48	82.68	78.26	53.12	50.14	57.18	50.38
dual	models	uniform×2	80.13	74.71	85.12	80.46	55.70	52.40	59.28	53.89
		re-balanced×2	80.18	74.54	84.99	80.81	55.44	52.01	59.26	53.43
	distributions	uniform; re-balanced	81.22	75.42	85.50	82.37	56.62	54.30	60.27	53.86

Table 3. mAP performance by using different consistencies.

VOC-LT	total	head	medium	tail
logit	81.44	75.68	85.53	82.69
probability	80.32	74.00	85.84	80.92
COCO-LT	total	head	medium	tail
logit	56.90	54.13	60.59	54.47
probability	56.03	53.11	59.85	53.55

Table 4. mAP performance of the proposed network by using the logit consistency and the probability consistency, respectively.

the re-balanced sampling, as well as considering the consistency of logits across two branches, e.g. [51, 27]. The model-based consistency from the uniform and re-balanced samplings yields the mAP performance of 80.13% and 80.18% respectively on VOC-LT, and 55.70% and 55.44% respectively on COCO-LT. We can see that the use of the proposed consistency in our method achieves much better mAP performance than both the uses of perturbation-based and model-based consistencies on both long-tailed datasets.

Finally, we conduct an experiment to justify the proposed logit consistency against the use of the probability consistency after the sigmod normalization in the proposed network. As shown in Table 4, the logit consistency yields better performance than the probability consistency, by avoiding gradient vanishing as discussed in Eq. (8).

4.3.3 Class-wise Analysis

In Fig.3, we show the class-wise average precision (AP) increment made by the re-balanced branch, the branch ensemble and the proposed network, respectively, when compared to solely using the uniform branch. As shown in the top row of Fig. 3, compared with uniform sampling for model training, re-balanced sampling leads to AP increment on tail classes (the right portion of each curve), since it increase the sampling rate of tail-class instances. Meanwhile, it also reduces the sampling rate of some head-class images, resulting in underfitting on head-class recognition and decreased AP performance on head classes, as shown in the left portion of each increment curve in the top row of Fig. 3. We can see that branch ensemble can alleviate the head-class

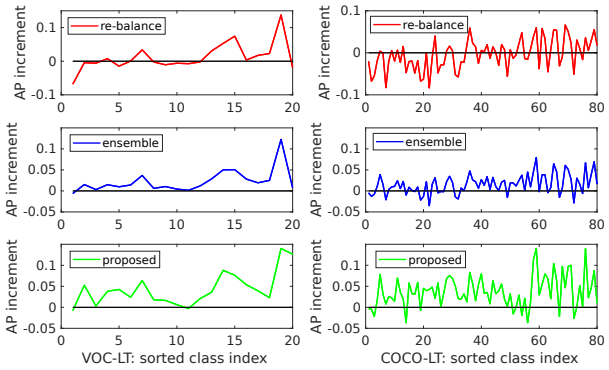

Figure 3. Class-wise AP increment of re-balanced branch, the branch ensemble and the proposed network over the uniform branch. Class labels are sorted from head to tail classes left-right.

performance decrease, while keeping the AP increment in tail classes, as shown in the middle row of Fig. 3. The proposed method further improve the AP performance of most head, medium and tail classes by considering logit consistency between two branches and the logit compensation, as shown in the bottom row of Fig. 3.

To further understand the proposed logit compensation, we visualize the learned distribution parameters of Eq. (6) in Fig. 4. From the top row of Fig. 4, we can see that the mean values for positive and negative logit compensation are almost opposite to each other. The absolute mean value for each class largely follows a positive correlation with the sample number in this class. Since the mean values for compensating logits of positive samples and negative samples are positive and negative, respectively, the absolute values of logits increases for correct predictions. This helps decrease the loss values and prevents the logit values from being away from 0 quickly. The standard deviations also approximately follow a positive correlation with the sample number in each class, as shown in the bottom row of Fig. 4. Besides, we can also notice that the standard deviations learned for positive logits are usually smaller than 1 and those learned for negative logits are usually larger than 1. For most classes, positive samples are usually the minor-

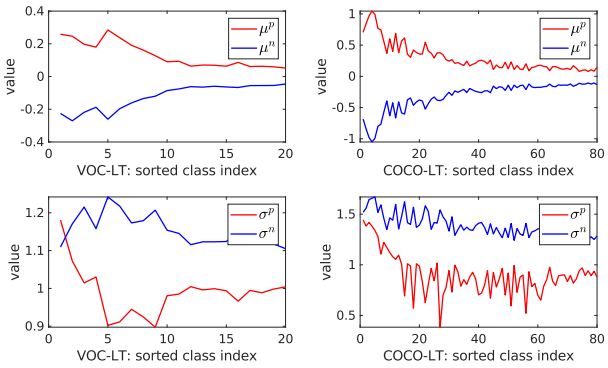

Figure 4. The visualization of learned logit compensation parameters for positive and negative logits, on VOC-LT and COCO-LT. Class labels are sorted from head to tail classes left-right.

ity, while the negative samples are the majority. A standard deviation lower than 1 inclines to increase the classification loss from the logits, while a standard deviation greater than 1 tends to decrease the classification loss from the logits. Therefore, the loss from positive samples, along with the tail classes, are relatively emphasized to address the imbalance issue.

4.3.4 Group-wise Analysis

For all the compared methods in Table 1, we can notice an interesting phenomenon that mAP performance on medium classes is usually higher than those on head classes and on tail classes. The prior work [47] gives an conjecture that sample numbers of medium classes (10 to 100 samples per class) may be more suitable for the specific multi-label learning. We agree with this conjecture. With a simplified assumption that there is only one label associated to each image, a class is balanced if its number of samples is $\frac{N}{K}$. On VOC-LT, $\frac{N}{K} = \frac{1142}{20} = 57$ and on COCO-LT, $\frac{N}{K} = \frac{1909}{80} = 23.9$, both of which are in the range of $[10, 100]$ used for defining medium classes. Therefore, the sample numbers of medium classes are already more balanced than those of the head and tail classes.

In addition, the use of re-balanced sampling, such as DB Focal, baseline-re-balanced, or the proposed method, usually leads to better performance on tail classes than on head classes, as shown in Table 1. One possible reason is that images with head class labels are usually associated with more classes and show more diverse and complex appearance features. As shown in Fig. 5, it is clear that head classes have more co-occurred classes than tail classes. In this case, without sufficient samples, the image diversity and complexity for head classes are more difficult to learn than simpler tail-class images.

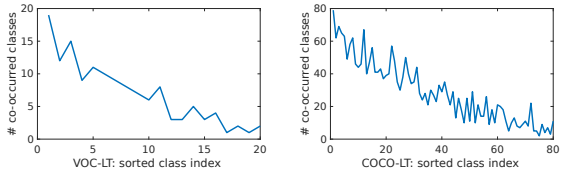

Figure 5. Number of co-occurred classes on the same image in term of class labels sorted from head classes to tail classes on the two datasets.

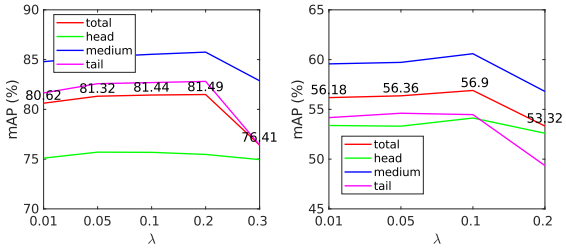

Figure 6. The effect of hyper-parameter λ to the mAP performance.

4.4. Effect of Hyper-parameter λ

Besides the conventional hyper-parameters for deep network learning, the proposed method introduces one more hyper-parameter to tune, i.e. λ in Eq. (3), which is end-to-end training friendly. We further conduct a set of experiments to study the effect of different configurations of λ to the recognition performance. As shown in Fig. 6, when $\lambda = 0.2$, the proposed method achieves the best mAP performance of 81.49% on VOC-LT. When $\lambda = 0.1$, the proposed method achieves the best mAP performance of 56.90% on COCO-LT. An overly small λ may not give sufficient consideration for the consistency, while an overly large λ may make the consistency dominate the training, leading to decreased performance on the original task of multi-label recognition.

5. Conclusion

In this paper, we tackled the task of long-tailed multi-label visual recognition by learning a model using both uniform and re-balanced samplings from the same training set. We proposed a network consisting of two branches for two samplings, respectively. Meanwhile, we incorporated the logit consistency across two branches for the same input to achieve collaborative learning. With extensive experiments on two long-tailed datasets for multi-label recognition, we demonstrated the effectiveness of the proposed method by achieving the new state-of-the-art performance, with significant margins over prior works.

References

[1] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. 1, 2

[2] Ricardo Cabral, Fernando De la Torre, Joao Paulo Costeira, and Alexandre Bernardino. Matrix completion for weakly-supervised multi-label image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):121–135, 2014. 1, 2

[3] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, pages 1567–1578, 2019. 1, 2, 4, 5, 6

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

[5] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2019. 1, 2, 5, 6

[6] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019. 1, 2, 5, 6

[7] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 5

[8] Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. Visual attention consistency under image transforms for multi-label image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 729–739, 2019. 2, 3, 4, 5, 6

[9] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, pages 8527–8537, 2018. 3

[10] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. 1, 2, 5

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2, 3

[12] Hexiang Hu, Guang-Tong Zhou, Zhiwei Deng, Zicheng Liao, and Greg Mori. Learning structured inference neural networks with label relations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2960–2968, 2016. 1, 2

[13] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *IEEE Conference on Computer Vision and pattern recognition*, pages 5375–5384, 2016. 2

[14] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 3, 2017. 2

[15] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7610–7619, 2020. 1

[16] 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. 1, 3

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 2, 5

[18] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 3

[19] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *Asian Conference on Pattern Recognition*, pages 111–115. IEEE, 2015. 2, 4

[20] Qiang Li, Maoying Qiao, Wei Bian, and Dacheng Tao. Conditional graphical lasso for multi-label image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2977–2986, 2016. 1, 2

[21] Xin Li, Feipeng Zhao, and Yuhong Guo. Multi-label image classification with a probabilistic label enhancement model. In *Uncertainty in Artificial Intelligence*, volume 1, pages 1–10, 2014. 1, 2

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. 5, 6

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 5

[24] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2970–2979, 2020. 1

[25] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. 1, 5, 6

[26] Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. In *IEEE International Conference on Computer Vision*, pages 5048–5057, 2017. 3

[27] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Multi-label co-regularization for semi-supervised facial action unit recognition. In *Advances in Neural Information Processing Systems*, pages 909–919, 2019. 1, 2, 3, 4, 7

[28] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *European Conference on Computer Vision*, pages 135–152, 2018. 3

[29] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. *arXiv preprint arXiv:1803.09050*, 2018. 2

[30] 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

[31] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 1163–1171, 2016. 3

[32] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *European Conference on Computer Vision*, pages 467–482. Springer, 2016. 1, 2, 3, 5, 6

[33] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019. 3

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[35] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015. 2

[36] Zichang Tan, Yang Yang, Jun Wan, Guodong Guo, and Stan Z Li. Relation-aware pedestrian attribute recognition with graph convolutional networks. In *AAAI Conference on Artificial Intelligence*, pages 12055–12062, 2020. 2, 4

[37] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, pages 1195–1204, 2017. 3

[38] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007. 2

[39] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018. 1

[40] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2016. 1, 2

[41] Yiru Wang, Weihao Gan, Jie Yang, Wei Wu, and Junjie Yan. Dynamic curriculum learning for imbalanced data classification. In *IEEE International Conference on Computer Vision*, pages 5017–5026, 2019. 1

[42] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020. 3

[43] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems*, pages 7029–7039, 2017. 2

[44] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *IEEE International Conference on Computer Vision*, pages 464–472, 2017. 1, 2, 5

[45] Yunchao Wei, Wei Xia, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. Cnn: Single-label to multi-label. *arXiv preprint arXiv:1406.5726*, 2014. 1, 2

[46] Daniel Worrall and Gabriel Brostow. Cubenet: Equivariance to 3d rotation and translation. In *European Conference on Computer Vision*, pages 567–584, 2018. 3

[47] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. *arXiv preprint arXiv:2007.09654*, 2020. 1, 2, 3, 4, 5, 6, 8

[48] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 3

[49] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007. 2

[50] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2013. 2

[51] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018. 1, 2, 3, 4, 7

[52] Rui-Wei Zhao, Jianguo Li, Yurong Chen, Jia-Ming Liu, Yu-Gang Jiang, and Xiangyang Xue. Regional gating neural networks for multi-label image classification. In *British Machine Vision Conference*, pages 1–12, 2016. 1, 2

[53] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9719–9728, 2020. 1, 2, 5, 6

[54] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5513–5522, 2017. 1, 2, 5