Distributionally Robust Loss for Long-Tailed Multi-Label Image Classification

Dekun Lin[1,2], Tailai Peng[1,2], Rui Chen[1,2], Xinran Xie[1,2], Xiaolin Qin[1,2]✉, and Zhe Cui[1,2]✉

[1] Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu 610213, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
kunonkey@163.com

Abstract. The binary cross-entropy (BCE) loss function is widely utilized in multi-label classification (MLC) tasks, treating each label independently. The log-sum-exp pairwise (LSEP) loss, which emphasizes higher logits for positive classes over negative ones within a sample and accounts for label dependencies, has demonstrated effectiveness for MLC. However, our experiments suggest that its performance in long-tailed multi-label classification (LTMLC) appears to be inferior to that of BCE. In this study, we investigate the impact of the log-sum-exp operation on recognition and explore optimization avenues. Our observations reveal two primary shortcomings of LSEP that lead to its poor performance in LTMLC: 1) the indiscriminate use of label dependencies without consideration of the distribution shift between training and test sets, and 2) the overconfidence in negative labels with features similar to those of positive labels. To mitigate these problems, we propose a distributionally robust loss (DR), which includes class-wise LSEP and a negative gradient constraint. Additionally, our findings indicate that the BCE-based loss is somewhat complementary to the LSEP-based loss, offering enhanced performance upon integration. Extensive experiments conducted on two LTMLC datasets, VOC-LT and COCO-LT, demonstrate the consistent effectiveness of our proposed method. Code: https://github.com/Kunmonkey/DR-Loss.

Keywords: Long-tailed learning · Multi-label classification · Loss

1 Introduction

Among various visual recognition tasks in the real world, multi-label classification (MLC) is a prevalent scenario. This task requires identifying multiple objects within a single image, exemplified by tasks such as human attribute recognition [11, 23, 25], image retrieval [16, 21, 24], *etc*. Real-world datasets usually exhibit a long-tailed distribution [31,36,38,40], characterized by a few dominant categories (also known as head classes) that encompass the majority of samples,

✉ Corresponding author.

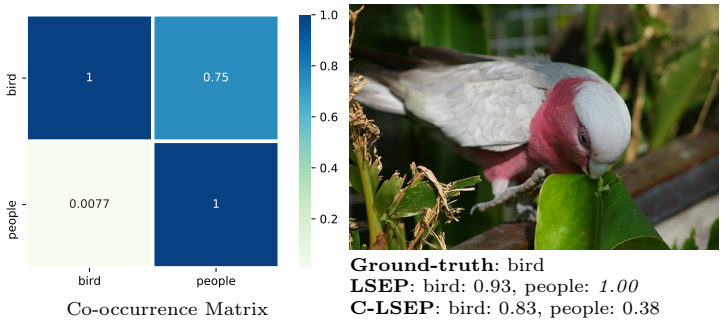


Fig. 1: The co-occurrence matrix, representing the relationship between birds and people within the VOC-LT training set, is depicted on the left. On the right, the results predicted by LSEP and C-LSEP for an image labeled as 'bird' from the VOC-LT testing set are presented. All numbers denote the predicted probabilities, ranging from 0 to 1.

while the majority of classes are instance-scarce (also known as tail classes). Such a distribution pattern in MLC (LTMLC) [18,35] may lead to overfitting of deep models on tail classes, significantly diminishing performance.

The log-sum-exp pairwise (LSEP) [17] is a ranking-based loss designed for MLC. It leverages the dependencies among labels with the goal of ensuring the predicted logits for positive labels surpass those for negative labels within a single image. In essence, it utilizes co-occurrence information from the training set to aid in prediction. For instance, as Fig. 1 shows, the co-occurrence matrix for birds and people in the VOC-LT training set is depicted. The first row represents the likelihood of birds co-occurring with other instances, with diagonal elements valued at 1.0. The probability of birds co-occurring with people is notably high at 0.75. For right figure labeled solely with 'bird', LSEP assigns a high prediction probability of 0.93 for the presence of birds. Having learned the co-occurrence information (*i.e.* high co-occurrence probability of birds with people (0.75)), LSEP erroneously predicts the presence of people with a probability of 1.0, despite the absence of people-related features. This issue highlights a potential drawback: significant distribution shifts between training and testing data, especially for tail classes with fewer samples, can be detrimental to LTMLC. Fundamentally, models should capture robust features for each class to accurately distinguish one class from another. However, relying on co-occurrence information as a shortcut can hinder LSEP's capability to extract the true features of specific classes.

In this paper, we propose calculating the LSEP at the class-wise level (C-LSEP) to learn more robust features. Specifically, positive labels belong to the same categories across various images. In calculating the loss for a certain category, C-LSEP aims to distinguish the features of this class from others, enabling the model to learn more robust and distinguishable features, akin to those in face recognition [30]. This approach effectively avoids reliance on incorrect co-occurrence information and focuses on extracting robust features. Typically, in

Ground-truth: cow, people
C-LSEP: cow: 0.68, people: 0.57, dog: *0.61*, horse: *0.69*
C-LSEP + NGC: cow: 0.54, people: 0.58, dog: 0.41, horse: 0.45

Fig. 2: A practical example labeled with both 'cow' and 'people' simultaneously from the testing set of VOC-LT. The NGC, our proposed negative gradient constraint, is applied exclusively to negative logits. All numbers represent predicted probabilities, which range between 0 and 1.

LTMLC, the number of negative labels significantly exceeds that of positive labels, leading to the excessive accumulation of gradients from negative labels with BCE-based loss. Inspired by DB loss [35] and ASL loss [27], we incorporate gradient scaling factors into C-LSEP. As shown in Fig. 1, with C-LSEP, the predicted probability of 'people' decreases to a normal level of 0.38.

C-LSEP still encounters the problem of excessively trusting certain negative labels, leading to their prediction with overly high probabilities. For example, consider Fig. 2, depicting an image labeled as both 'people' and 'cows.' When applying our C-LSEP, the model correctly classifies 'people' and 'cows' with probabilities higher than 0.5. However, the model predicts 'dogs' and 'horses' with even higher probabilities than those for 'people', leading to inaccurate results. In multi-label settings, our objective is to assign higher probabilities to positive labels than to negative labels. Typically, errors arise from false positives, where the probabilities assigned to negative labels are excessively high, sometimes even surpassing those assigned to positive labels. If certain classes are absent in a mini-batch, the gradient of negative labels by C-LSEP will be zero, resulting in excessively high negative logits. To address this issue, we introduce the gradient constraint for negative labels (NGC) , ensuring they receive a gradient and preventing the negative logits from becoming too high. As illustrated in Fig. 2, the NGC significantly alleviates the problem of models excessively trusting in incorrect negative labels, successfully decreasing the probabilities for both 'dogs' and 'horses'. Our contributions can be summarized as follows:

– We find that the LSEP loss tends to rely excessively on label dependencies information, which hinders the model's ability to learn robust features. Therefore, we propose the C-LSEP to tackle this problem, aiming to enhance the model's generalization capabilities. Additionally, we introduce gradient scaling factors to moderate the learning rate for easier samples.

- To alleviate the problem of models being overly confident in incorrect negative labels, we propose the negative gradient constraint to guarantee gradients for negative labels.
- Our proposed DR loss shows a marked improvement over the LSEP loss in the LTMLC task. We identify complementarity between LSEP-based and BCE-based losses, which enables their integration for further improvement.

2 Related Work

Multi-Label Classification. High-level semantic information conveyed by labels tends to attract researchers' attention in MLC. With the widespread use of CNNs for various computer vision tasks, methods employed Recurrent Neural Networks (RNNs) [32, 34] and Graph Convolutional Networks (GCNs) [5, 15, 37] to capture label dependencies. In addition to model design, cost-sensitive methods utilize semantic information among labels. The LSEP loss [17] strives to prioritize the predicted probability of positive labels over negative labels for each sample. Building upon LSEP loss, ZLRP loss [29] proposes to add zero margins for an adaptive determination on the number of target categories. While these methods have shown undeniable effectiveness in MLC, we contend that the distribution shift between the training and testing sets in LTMLC may lead to excessive and erroneous use of semantic information, resulting in poor prediction.
Long-Tailed Single-Label Classification (LTSLC). As widely used in long-tailed learning, re-sampling [3,28] is a technique that involves either over-sampling the minority categories or under-sampling the frequent categories during training [2, 12]. Class-aware sampling [28] selects samples of each category with equal probabilities. Re-weighting the loss is another technique by adjusting the weights of each class based on their frequency in the dataset to control the training loss at a class-wise level [7, 14, 33]. Alternatively, loss can also be controlled at the sample level [19, 26]. Recent advanced work ensemble multiple experts [1, 25, 39] to extract more robust features. Despite their effectiveness in LTSLC, these approaches do not necessarily translate well to a multi-label setting.
Long-Tailed Multi-Label Classification. Compared to MLC and LTSLC, which have attracted increasing attention, research on LTMLC remains limited. Based on the BCE loss, which is commonly used for MLC, Wu *et al.* [35] propose the DB loss with the goal of suppressing the gradients of negative labels. This design effectively mitigates the over-suppression of negative labels. Another cost-sensitive method for MLC, the ranking-based loss function such as LSEP, has yet to be explored in the context of LTMLC. In this paper, we adapt this approach to LTMLC with necessary modifications and achieve notable advancements.

3 Approach

3.1 Preliminaries

Cross-entropy (CE) is the most commonly used loss function in single-label classification (SLC) tasks. Let x^k be the k-th sample in the training set, with its

corresponding label denoted as $y^k \in \{1, 2, \ldots, N\}$, where N represents the total number of categories. Let s denote the predicted logit output by the classifier. The CE loss for sample x^k is given by:

$$L_{ce}(x^k, y^k) = -\log \frac{e^{s_i}}{\sum_{j=1}^{C} e^{s_j}} = \log(1 + \sum_{j=1, j\neq i}^{C} e^{s_j - s_i}), \tag{1}$$

where s_i is the output logit of the i-th class, with i denoting the ground-truth, and s_j represents the logit of the j-th category. The log-sum-exp (LSE) operator is actually a smooth approximation of the maximum operator, *i.e.*,

$$\log(1 + \sum_{j=1, j\neq i}^{C} e^{s_j - s_i}) \approx max(0, s_j - s_i, \ldots), \tag{2}$$

as pointed out by [29, 30]. The CE loss is actually to make the logit of target category bigger than the negative ones within a sample.

When a sample contains multiple labels simultaneously, the task transitions to multi-label classification. In this context, the LSE-based loss becomes the log-sum-exp pairwise (LSEP) function, as defined by:

$$L_{lsep}(x^k, y^k) = \log(1 + \sum_{i \in \Omega_{\text{pos}}} \sum_{j \in \Omega_{\text{neg}}} e^{s_j - s_i}), \tag{3}$$

where Ω_{pos} and Ω_{neg} denote the sets of positive and negative labels of x^k, respectively. Note that $|\Omega_{\text{pos}}| \geq 1$, indicating that an image may have one or multiple labels.

Based on the LSEP loss, the ZLPR loss [29] extends its application to MLC tasks in natural language processing (NLP) by incorporating a zero margin for both positive and negative labels. This addition explicitly allows for an adaptive determination of the number of target categories. The loss for sample x^k is:

$$L_{zlpr}(x^k, y^k) = \log(1 + \sum_{i \in \Omega_{\text{pos}}} e^{-s_i}) + \log(1 + \sum_{j \in \Omega_{\text{neg}}} e^{s_j}). \tag{4}$$

3.2 Class-Wise LSEP

Both the LSEP and ZLPR losses are calculated on a per-sample basis, with the aim that the predicted positive logits for a single image sample exceed those of the negative ones. For a mini-batch of size M, the total loss of LSEP is:

$$L_{tot\_lsep} = \sum_{i=1}^{M} L_{lsep}(x^i, y^i), \tag{5}$$

where x^i is the i-th sample in a mini-batch, and y^i is the corresponding label.

Although calculating on a sample-wise basis may seem more intuitive, we propose calculating the LSEP loss at a class-wise level (C-LSEP) to extract more

robust features and to avoid incorrect adoption of co-occurrence information. The total loss of C-LSEP in a mini-batch is given by:

$$L_{tot\_c-lsep} = \sum_{k=1}^{N} L_{c-lsep}(C^k), \tag{6}$$

where C^k represents the k-th class among N categories across M samples. The loss for the k-th class is defined as follows:

$$L_{c-lsep}(C^k) = \log(1 + \sum_{i \in \Omega_{\text{pos}}} \sum_{j \in \Omega_{\text{neg}}} e^{\gamma(s_j - s_i)}), \tag{7}$$

where Ω_{pos} and Ω_{neg} denote the sets of positive labels and negative labels for C^k, respectively. The predicted matrix dimensions are $M \times N$. The LSEP calculates the loss row by row (sample by sample), where the categories of positive logits differ and co-occur within the same image. In contrast, our C-LSEP calculates the loss column by column, with the positive and negative labels for C^k located in the k-th column, belonging to different samples. All positive logits are associated with the same category. Motivated by the DB and ASL losses, we introduce a gradient scaling hyperparameter, γ.

3.3 Negative Gradient Constraint

If certain classes are absent in a mini-batch, the gradients for negative labels computed by the C-LSEP will be zero. This situation can lead to models exhibiting overconfidence in negative labels due to the lack of gradients. To address this problem, we introduce a gradient constraint for negative labels (NGC), denoted by $\sum_{j \in \Omega_{\text{neg}}} e^{\gamma_2 s_j}$. The NGC ensures that negative labels receive gradients, assisting in lowering the probability threshold and thereby preventing the negative probabilities from becoming excessively high. For further details, please refer to the gradient analysis in Sec. 3.4. In combination with the C-LSEP loss, our final loss, termed DR loss, is defined as follows:

$$L_{DR}(C^k) = \log(1 + \sum_{i \in \Omega_{\text{pos}}} \sum_{j \in \Omega_{\text{neg}}} e^{\gamma_1(s_j - s_i)} + \sum_{j \in \Omega_{\text{neg}}} e^{\gamma_2 s_j}). \tag{8}$$

Notably, there are two hyperparameters: γ_1 for C-LSEP and γ_2 for NGC, respectively. The value of γ_2 can be adjusted independently of γ_1, offering increased flexibility in controlling the probability threshold.

3.4 Gradient Analysis

The gradient of negative logit s_k ($s_k \in \Omega_{\text{neg}}$), *i.e.* negative gradient, is given by:

$$\frac{\partial L_{c-lsep}}{\partial s_k} = \frac{\gamma * e^{\gamma s_k} * \sum\limits_{i \in \Omega_{\text{pos}}} e^{-\gamma si}}{1 + \sum\limits_{i \in \Omega_{\text{pos}}} \sum\limits_{j \in \Omega_{\text{neg}}} e^{\gamma(s_j - s_i)}}. \tag{9}$$

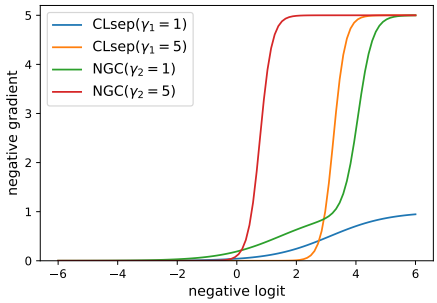


Fig. 3: The gradient curves on negative labels. The x-axis denotes the logit values of negative labels, and the y-axis represents the corresponding gradients. The NGC is based on C-LSEP with $\gamma_1 = 5$.

After applying the negative gradient constraint (NGC), the negative gradient of s_k is then given by:

$$\frac{\partial L_{DR}}{\partial s_k} = \frac{\gamma_1 * e^{\gamma_1 s_k} * \sum\limits_{i \in \Omega_{\text{pos}}} e^{-\gamma_1 s_i} + \gamma_2 e^{\gamma_2 s_k}}{1 + \sum\limits_{i \in \Omega_{\text{pos}}} \sum\limits_{j \in \Omega_{\text{neg}}} e^{\gamma_1 (s_j - s_i)} + \sum\limits_{j \in \Omega_{\text{neg}}} e^{\gamma_2 s_j}}. \tag{10}$$

Direct comparison of the negative gradients of these losses can be challenging. To simplify observations, we visualize the negative gradient curves for different losses under varying hyperparameters. As depicted in Fig. 3, with an increased γ_1, C-LSEP exhibits a smaller gradient for low predicted logits and a larger gradient for high predicted logits. This phenomenon emphasizes the optimization of hard samples while decelerating the optimization of easy samples.

Please note that a probability threshold exists within the negative gradient curve. When the predicted probability exceeds this threshold, gradients are generated for negative labels. For instance, in the case of C-LSEP with $\gamma_1 = 1$ (blue curve), the probability threshold is approximately 0.5, with the corresponding logit being about 0. This probability threshold of 0.5 is derived using the sigmoid function on the logit. If the predicted probability is below 0.5, the negative gradient is nearly zero. However, it gradually increases as the probability surpasses 0.5. For C-LSEP with $\gamma_1 = 5$, the probability threshold is around 0.85, indicating that the hyperparameter γ_1 can increase the probability threshold. A threshold being too high means negative labels only generate gradients when their probability is above this threshold, potentially leading to large logits and a higher likelihood of being misclassified as positive. This phenomenon explains why the C-LSEP may exhibit overconfidence for negative labels that share similar features with positive labels. With our NGC, the probability threshold is significantly reduced. For example, with $\gamma_2 = 5$, the probability threshold decreases to about 0.5. Despite this reduction, the gradient for hard samples remains elevated, which is promising.

4 Experiments

4.1 Experimental Settings

Just as in [35], we conduct experiments on two long-tailed multi-label image classification datasets: VOC-LT and COCO-LT. These datasets are artificially constructed from the Pascal Visual Object Classes Challenge (VOC) [9] and Microsoft COCO (MS-COCO) [20], respectively.

VOC-LT. The VOC-LT is derived from the 2012 train-val set of VOC based on the Pareto distribution as in [22]. It encompasses 1142 images across 20 classes in the training set, with the number of each class ranging from 4 to 775. These 20 classes are categorized into three groups according to the number of samples per class: the head class for classes with more than 100 samples, the medium for those with 20 to 100 samples, and the tail for classes with fewer than 20 samples. After the splitting, the ratio of head, medium, and tail classes is 6:6:8. The test set contains 4952 images, which is identical to the 2007 VOC test set.

COCO-LT. The COCO-LT is derived from the 2017 version of MS-COCO through a similar sampling method. It comprises 1909 images across 80 classes in the training set, with the number of images per class ranging from 6 to 1128. The method for dividing classes is akin to that used in VOC-LT, resulting in a distribution ratio of head, medium, and tail classes of 22:33:25, respectively, after division. The test set includes 5000 images, which is identical to the test set of the 2017 MS-COCO.

Implementation Details. For fair comparision, both the configurations and evaluation metrics used for LTMLC are similar to those described in [35]. More specifically, we adopt the mean average precision (mAP) as the evaluation metric. We utilize ResNet50 [13], which is pre-trained on ImageNet [8], as the backbone of the model. The input images are randomly cropped and resized to dimensions of 224×224. Standard data augmentations, as described in [35], are employed, alongside a class-aware resampling strategy identical to that used in [7, 28]. The samples are organized into batches of size 32. For optimization, we employ SGD with a momentum of 0.9 and a weight decay of 0.0001. The initial learning rates are set to 0.0003 for VOC-LT and 0.0001 for COCO-LT, with a warm-up learning rate schedule [10] for the first 500 iterations at a ratio of $\frac{1}{3}$. All core codes from the DB loss [35] are extracted and retrained by us. All experiments are conducted using PyTorch 1.8.0.

4.2 Benchmark Results

To begin, we compare the mean Average Precision (mAP) scores of our method with those of previous methods across two long-tailed benchmark datasets to verify the effectiveness of our proposed method. Our work focuses on cost-sensitive methods, whereby these comparision methods contain Empirical Risk Minimization (ERM), Re-Weighting (RW)—which re-weights by the inverse proportion of the square root of class frequency, Re-Sampling (RS) [28], Focal Loss [19], LSEP Loss [17], ML-GCN [6], LDAM [4], CB Focal [7], Circle Loss [30], DB Loss [35],

Table 1: The mAP performance of our proposed method and other comparision methods is presented. The results for DB Focal is obtained by re-training. This is the first time Circle, ASL, ZLPR loss and BalPoE being applied in this task. The results for the other methods are directly cited from the DB loss [35]. Best results are bolded.

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
RM	74.70	67.58	82.81	73.96	42.27	48.62	45.80	32.02
RS [28]	75.38	70.95	82.94	73.05	46.97	47.58	50.55	41.70
Focal Loss [19]	73.88	69.41	81.43	71.56	49.46	49.80	54.77	42.14
LSEP [17]	72.99	69.00	79.83	70.88	47.05	46.18	50.91	42.88
ML-GCN [6]	68.92	70.14	76.41	62.39	44.24	44.04	48.36	38.96
LDAM [4]	70.73	68.73	80.38	69.09	40.53	48.77	48.38	22.92
CB Focal [7]	75.24	70.30	83.53	72.74	49.06	47.91	53.01	44.85
Circle Loss [30]	75.20	70.00	82.00	73.88	52.00	48.64	55.52	50.28
DB Focal [35]	78.29	72.67	83.17	78.75	53.45	50.91	56.58	51.52
ASL Loss [27]	76.40	70.70	82.26	76.29	50.21	49.05	53.65	46.68
ZLPR Loss [29]	75.10	71.00	82.67	72.38	49.90	47.59	53.73	47.00
BalPoE [1]	73.76	69.00	82.17	71.00	50.02	48.45	54.18	45.96
DR Loss (**Ours**)	78.01	72.19	83.83	77.69	54.10	50.27	57.00	**53.76**
DR (**Ours**) + DB Focal	**78.75**	**73.17**	**83.87**	**78.82**	**54.55**	**51.18**	**57.64**	53.40

ASL Loss [27], ZLPR Loss [29] and BalPoE [1]. BalPoE is the most advanced method for LTSLC, we use BCE as its base loss function. It is noteworthy that the results of LSEP, Circle, DB Focal, ZLPR, BalPoE and our method are all obtained using RS. Accordingly, it is observed that LSEP does not enhance performance compared to RS.

The mAP performance of different methods is presented in Tab. 1. These benchmark results underscore the superiority of our method, which surpasses previous methods by a certain margin. Specifically, our proposed DR loss achieves total mAP scores of 78.01% and 54.1% on VOC-LT and COCO-LT, respectively. In comparison to the most related method, the LSEP loss, our method demonstrates an mAP improvement of approximately 5% on VOC-LT and 7% on COCO-LT, which is remarkable. When compared to previous state-of-the-art (SOTA) method, the DB Focal loss, our method achieves an approximately 0.7% higher mAP on COCO-LT, albeit a 0.3% lower mAP on VOC-LT. These results suggest that: 1) our method is capable of achieving SOTA performance from a novel perspective; and 2) this innovative approach is especially effective for datasets with multiple classes (such as COCO-LT), thereby demonstrating its enhanced efficacy in managing challenging datasets.

Furthermore, by observing performances across three subset classes, *i.e.* head, medium and tail classes, we find that our LSEP-based loss complements the BCE-based loss, specifically, DB Focal. Concretely, our method performs bet-

Table 2: Ablation analysis on different components of our proposed method. CW, which stands for class-wise, refers to calculate the loss for each class individually. NGC denotes negative gradient constraint. The Hyperparameter γ_1 of C-LSEP is set to 5 for both VOC-LT and COCO-LT. We set $\gamma_2 = 7$ for VOC-LT and $\gamma_2 = 10$ for COCO-LT.

Methods					VOC-LT				COCO-LT			
LSEP	CW	NGC	DB	Focal	total	head	medium	tail	total	head	medium	tail
√					72.99	69.00	79.83	70.88	47.05	46.18	50.91	42.88
√	√				74.93 (+1.94)	71.00	81.00	73.25	50.35 (+3.30)	49.14	54.18	46.36
√	√	√			78.01 (+3.08)	72.19	83.83	77.69	54.10 (+3.75)	50.27	57.00	53.76
√	√	√	√		78.75 (+0.74)	73.17	83.87	78.82	54.55 (+0.45)	51.18	57.64	53.40

ter on medium classes for VOC-LT, whereas the BCE-based loss shows stronger performance in both head and tail classes. Yet for COCO-LT, our methods outperforms in both medium and tail classes. With a combination of our method with DB Focal, the performance gets improved further. We achieve a total mAP gain of about 0.5% over DB Focal on VOC-LT and a 0.5% mAP gain over our own approach on COCO-LT, with the highest results across all subsets.

4.3 Ablation Study

The overall ablation results are reported in Tab. 2. We observe that applying class-wise calculation to the LSEP loss resulted in a performance gain of about 1.9% on VOC-LT and 3.3% on COCO-LT. Additionally, there has been a significant improvement in performance across all three subset classes. Then with our proposed NGC, performance further increases by about 3% mAP on VOC-LT and 3.8% on COCO-LT, which is substantial. The advance takes place primarily in the tail classes. Compared to the baseline method of LSEP loss, our approach achieves an overall performance improvement of approximately 5% on VOC-LT and 6.7% on COCO-LT. Combined with DB Focal loss, there is a further mAP performance gain of about 0.8% on VOC-LT and 0.5% on COCO-LT.

Class-Wise LSEP To verify that our class-wise calculation indeed enhances performance, we conduct experiments with both LSEP and ZLPR losses, comparing scenarios with and without class-wise calculation. Regarding LSEP loss, it inherently lacks a hyperparameter. However, in C-LSEP, we introduce a scaling hyperparameter, γ_1. It might be mistakenly assumed that the improvement in performance is solely due to this parameter. To prevent such a misinterpretation, we also apply varying values of γ_1 to LSEP. This enables a fair comparison of the effectiveness of class-wise calculation between LSEP and C-LSEP.

As presented in Tab. 3, the performance of the LSEP loss indeed improves slightly with the increase of γ_1 on both VOC-LT and COCO-LT. Yet, our class-wise operation exhibits a stable 1% gain in the total mAP score for VOC-LT,

Table 3: Ablation study on class-wise calculations. C-ZLPR denotes the calculation of ZLPR loss at the class level.

	VOC-LT				COCO-LT			
Methods	total	head	medium	tail	total	head	medium	tail
LSEP($\gamma = 1$)	72.99	69.00	79.83	70.88	47.36	46.32	50.52	44.16
C-LSEP($\gamma = 1$)	73.98	69.33	81.83	71.62	49.60	49.09	53.21	45.24
LSEP($\gamma = 3$)	73.58	67.83	80.17	73.00	47.72	45.68	51.09	45.00
C-LSEP($\gamma = 3$)	74.59	71.00	81.33	72.13	50.16	49.23	53.85	46.16
LSEP($\gamma = 5$)	73.86	68.67	79.33	73.50	47.44	45.09	51.18	44.52
C-LSEP($\gamma = 5$)	74.93	71.00	81.00	73.25	50.35	49.14	54.18	46.36
ZLPR	75.10	71.00	82.67	72.38	49.90	47.59	53.73	47.00
C-ZLPR	76.12	71.33	83.00	74.63	51.76	49.41	54.91	49.60

Table 4: Ablation study on the negativity of gradient constraint. PGC denotes the positive gradient constraint, enforcing a gradient constraint solely on positive logits. NGC represents our proposed negative gradient constraint, applied solely to negative logits. GC refers to the simultaneous use of PGC and NGC. Best results are bolded.

	VOC-LT				COCO-LT			
Methods	total	head	medium	tail	total	head	medium	tail
C-LSEP	74.93	71.00	81.00	73.25	50.35	49.14	54.18	46.36
C-LSEP + PGC	74.74	71.00	81.33	72.75	50.29	49.55	53.97	46.08
C-LSEP + NGC	**78.01**	**72.19**	**83.83**	**77.69**	**54.10**	**50.27**	**57.00**	**53.76**
C-LSEP + GC	76.90	71.50	82.83	76.50	53.58	49.82	56.70	52.84

while for COCO-LT, it consistently enhances performance by about 2%-3%. ZLPR, an LSEP-based loss proposed for MLC tasks in NLP, calculates the original loss on a sample-wise basis, similar to the LSEP loss. Notably, the application of class-wise calculation also benefits ZLPR loss, specifically improving the mAP performance by about 1% on VOC-LT and 1.9% on COCO-LT. These experiments effectively demonstrate that class-wise calculation is advantageous and useful for LSEP-based loss, enabling the learning of more robust features.

Negative Gradient Constraint In previous sections, we have provided a detailed analysis of the reasons for introducing the NGC for two primary purposes: 1) to ensure that negative labels can receive gradients; and 2) to prevent the probability threshold for negative labels from being excessively high. The experimental results, as presented in Tab. 2, demonstrate the effectiveness of our proposed NGC. One might question the potential outcomes of incorporating a similar mechanism for positive labels, namely, a positive gradient constraint (PGC). **Why** then, is such a term added only for **negative** labels? As shown in

Table 5: Ablation study on the combination of different methods. For the baseline method, C-LSEP, the hyperparameter γ_1 is set to 1.

	VOC-LT				COCO-LT			
Methods	total	head	medium	tail	total	head	medium	tail
BCE	75.96	71.17	83.17	74.00	47.27	47.72	50.79	42.12
C-LSEP	73.98	69.00	79.83	70.88	49.60	49.09	53.12	45.24
BCE + C-LSEP	75.89	70.83	82.33	75.00	50.11	47.45	53.52	47.84
DB Focal	78.17	73.50	83.67	77.63	53.45	50.91	56.58	51.52
C-LSEP + DB Focal	76.21	68.33	82.83	77.00	51.09	45.73	54.30	51.36
DR	78.01	72.19	83.83	77.69	54.10	50.27	57.00	53.76
BCE + **DR**	77.12	72.17	83.67	76.00	52.84	49.95	56.30	50.80
DR + DB Focal	78.75	73.17	83.87	78.82	54.55	51.18	57.64	53.40

Tab. 4, incorporating a PGC can actually impair performance. It fails to enhance performance in the absence of NGC. When both PGC and NGC are applied concurrently (GC), the performance is even worse than using NGC alone. NGC lowers the probability threshold, allowing negative logits to receive gradients without needing to reach a high value. While PGC also lowers the probability threshold for positive labels, this leads to gradients being generated only when the predicted probability is below this threshold. However, this contradicts our objective of encouraging high probabilities for positive labels, which is achieved by maintaining a high threshold for these labels.

Why can We Integrate DB Focal with Our Loss We conduct a range of experiments to demonstrate the rationality behind combining two losses and to explore how we can integrate them. Initially, we merge two baseline methods, BCE and C-LSEP. As illustrated in Tab. 5, the mAP scores for C-LSEP across all splits are consistently lower than those for the BCE loss on VOC-LT. After their combination, the overall performance mirrors that of BCE when used alone. However, there is an improvement in the mAP for tail classes, while the mAP for head classes decreases. On COCO-LT, the performance trend between these two baselines reverses; that is, C-LSEP outperforms BCE in mAP scores across all splits. Upon integrating the two baseline losses, we observe an improvement in the total mAP and for tail classes, but a decline for head classes. This indicates that C-LSEP may be advantageous for tail classes, whereas BCE loss is more effective for head classes. Furthermore, the LSEP-based and BCE-based losses appear to be partially complementary, suggesting their integration could potentially enhance performance.

DB Focal is an advanced loss function tailored for LTMLC, based on BCE . When combining the baseline method, C-LSEP, with DB Focal, a performance drop is observed on both VOC-LT and COCO-LT. Similarly, integrating BCE with our advanced loss function, DR, also leads to a decline in performance.

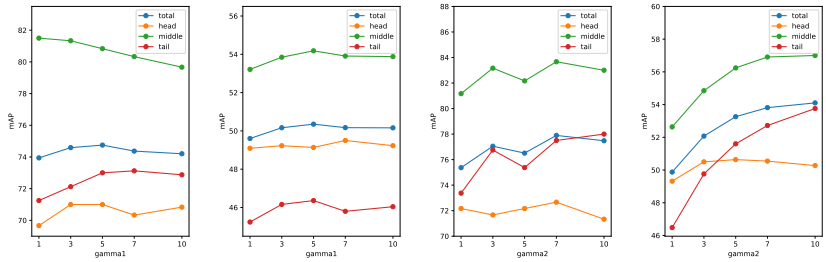
Table 6: Ablation study on the effect of batch size (BS), which ranges from 16 to 64.

Methods	VOC-LT				COCO-LT			
	total	head	medium	tail	total	head	medium	tail
LSEP (BS = 16)	73.06	70.50	79.83	69.88	47.92	46.50	51.39	44.56
C-LSEP (BS = 16)	73.96	70.33	81.33	70.88	48.19	47.68	52.03	43.44
LSEP (BS = 32)	72.99	69.00	79.83	70.88	47.27	47.72	50.79	42.12
C-LSEP (BS = 32)	73.98	69.33	81.83	71.62	49.60	49.09	53.12	45.24
LSEP (BS = 64)	72.75	68.83	80.00	70.63	47.11	46.59	50.24	43.64
C-LSEP (BS = 64)	73.55	68.83	81.33	71.38	50.33	48.68	54.24	46.72

The performance after integration finds a middle ground between the advanced and baseline methods. Given the significant performance gap between the advanced and baseline methods, any compromise is likely to result in diminished overall performance. A critical prerequisite for their combination is ensuring compatibility. The performance gap between our approach and DB Focal is not substantial, and each offers unique advantages, which is the primary reason why their integration is feasible and allows for further progress.

Influence of Batch Size The LSEP calculates the loss for each sample individually, with each sample being associated with a fixed number of categories (*e.g.*, VOC-LT has 20 classes). As a result, the gradient of the labels is unaffected by variations in batch sizes (BSs). In contrast, the C-LSEP calculates at a class-wise level, where the ratio of negative to positive labels may fluctuate with changes in BS. With an increase in BS, the number of negative labels may rise, potentially influencing the performance of the C-LSEP approach differently. Consequently, we conduct ablation studies on BS, both with and without class-wise calculation.

As reported in Tab. 6, the total mAP performance of LSEP reaches its optimal level with a BS of 16 for both VOC-LT and COCO-LT datasets. For VOC-LT, C-LSEP can consistently increase the mAP performance by approximately 0.8%-1.0%. In contrast, the improvement in mAP performance by C-LSEP for COCO-LT is relatively limited, at about 0.3%, when the BS is 16. However, performance significantly improves as the BS increases. Specifically, there is a 2.3% mAP gain when $BS = 32$ and a 3.2% mAP gain when $BS = 64$. This suggests that C-LSEP exhibits more potent capabilities for datasets with a larger number of categories, such as COCO-LT, which includes 80 categories. In the context of the current trend towards larger BSs, this advantage is particularly beneficial. Class-wise calculations remain effective across various BSs, showing relatively stable improvement for datasets with fewer classes. However, for more challenging datasets, C-LSEP benefits significantly from larger BSs, marking a departure from the performance trends observed with the original LSEP.



(a) γ_1 on VOC-LT **(b)** γ_1 on COCO-LT **(c)** γ_2 on VOC-LT **(d)** γ_2 on COCO-LT

Fig. 4: Impact of γ_1 and γ_2 on the mAP performance across two long-tailed datasets. In (a) and (b), NGC is not employed. For both VOC-LT and COCO-LT datasets, peak performance is observed at $\gamma_1 = 5$. Consequently, γ_1 is set to 5 in (c) and (d), where it remains constant.

Influence of Hyperparameter γ We have two hyperparameters, γ_1 for C-LSEP and γ_2 for NGC, respectively. The values of both γ_1 and γ_2 range from 1 to 10. Given that C-LSEP is the foundational term, we initially fine-tune γ_1 without NGC. As shown in Fig. 4a and Fig. 4b, optimal performance is achieved at $\gamma_1 = 5$ for both VOC-LT and COCO-LT. The introduction of γ_1 results in a enhancement of approximately 1% mAP for VOC-LT and 0.7% for COCO-LT.

We set $\gamma_1 = 5$ and examine the effect of γ_2 as depicted in Fig. 4c and Fig. 4d. It is observed that performance is poor when $\gamma_2 = 1$. However, as γ_2 increases, the performance significantly improves. Specifically, for VOC-LT, a value of $\gamma_2 = 7$ yields the best result, bringing an approximately 3% mAP gain. With regard to COCO-LT, setting γ_2 to 10 achieves the highest mAP improvement, approximately 3.7%. In both instances, the optimal values of γ_2 are found to be greater than that of γ_1. The performance gains achieved by fine-tuning γ_2 are substantial.

5 Conclusion

In this paper, we analyze the disadvantages of LSEP loss for LTMLC, noting its tendency to use label dependency information excessively, which leads to poor generalization. To learn more robust features, we propose calculating LSEP on a class-wise basis, effectively addressing this issue. Moreover, it is observed that C-LSEP may exhibit overconfidence for certain negative labels with features similar to those of positive labels, resulting in false positives. This issue is attributed to the vanishing negative gradient for certain classes in mini-batches and a large probability threshold for negative labels. To mitigate this issue, we introduce a gradient constraint for negative labels, significantly alleviating this problem. Extensive experiments demonstrate the effectiveness of our DR loss for LTMLC. Additionally, we identify a complementarity between BCE-based and LSEP-based losses, suggesting the potential of their integration. We believe that more reasonable integration methods can be explored in future work.

Acknowledgements

This research was supported by Sichuan Science and Technology Program (Nos. 2022ZHCG0007, 2024NSFJQ0035), the Talents by Sichuan provincial Party Committee Organization Department, and Chengdu - Chinese Academy of Sciences Science and Technology Cooperation Fund Project (Major Scientific and Technological Innovation Projects).

References

1. Aimar, E.S., Jonnarth, A., Felsberg, M., Kuhlmann, M.: Balanced product of calibrated experts for long-tailed recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19967–19977 (2023)
2. Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks **106**, 249–259 (2018)
3. Byrd, J., Lipton, Z.: What is the effect of importance weighting in deep learning? In: International Conference on Machine Learning. pp. 872–881. PMLR (2019)
4. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. arXiv preprint arXiv:1906.07413 (2019)
5. Chen, T., Xu, M., Hui, X., Wu, H., Lin, L.: Learning semantic-specific graph representation for multi-label image recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 522–531 (2019)
6. Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5177–5186 (2019)
7. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9268–9277 (2019)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
9. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**(1), 98–136 (2015)
10. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677 (2017)
11. Guo, H., Fan, X., Wang, S.: Human attribute recognition by refining attention heat map. Pattern Recognition Letters **94**, 38–45 (2017)
12. Han, H., Wang, W.Y., Mao, B.H.: Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: International conference on intelligent computing. pp. 878–887. Springer (2005)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Huang, C., Li, Y., Loy, C.C., Tang, X.: Learning deep representation for imbalanced classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5375–5384 (2016)

15. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
16. Lee, S., Kim, D., Han, B.: Cosmo: Content-style modulation for image retrieval with text feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 802–812 (2021)
17. Li, Y., Song, Y., Luo, J.: Improving pairwise ranking for multi-label image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3617–3625 (2017)
18. Lin, D.: Probability guided loss for long-tailed multi-label image classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1577–1585 (2023)
19. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
21. Liu, Z., Sun, W., Hong, Y., Teney, D., Gould, S.: Bi-directional training for composed image retrieval via text prompt learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5753–5762 (2024)
22. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2537–2546 (2019)
23. Metwaly, K., Kim, A., Branson, E., Monga, V.: Glidenet: Global, local and intrinsic based dense embedding network for multi-category attributes prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4835–4846 (2022)
24. Neculai, A., Chen, Y., Akata, Z.: Probabilistic compositional embeddings for multimodal image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4547–4557 (2022)
25. Ramanathan, V., Kalia, A., Petrovic, V., Wen, Y., Zheng, B., Guo, B., Wang, R., Marquez, A., Kovvuri, R., Kadian, A., et al.: Paco: Parts and attributes of common objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7141–7151 (2023)
26. Ren, M., Zeng, W., Yang, B., Urtasun, R.: Learning to reweight examples for robust deep learning. In: International Conference on Machine Learning. pp. 4334–4343. PMLR (2018)
27. Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric loss for multi-label classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 82–91 (2021)
28. Shen, L., Lin, Z., Huang, Q.: Relay backpropagation for effective learning of deep convolutional neural networks. In: European conference on computer vision. pp. 467–482. Springer (2016)
29. Su, J., Zhu, M., Murtadha, A., Pan, S., Wen, B., Liu, Y.: Zlpr: A novel loss for multi-label classification. arXiv preprint arXiv:2208.02955 (2022)
30. Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., Wei, Y.: Circle loss: A unified perspective of pair similarity optimization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6398–6407 (2020)

31. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8769–8778 (2018)
32. Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W.: Cnn-rnn: A unified framework for multi-label image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2285–2294 (2016)
33. Wang, Y.X., Ramanan, D., Hebert, M.: Learning to model the tail. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 7032–7042 (2017)
34. Wang, Z., Chen, T., Li, G., Xu, R., Lin, L.: Multi-label image recognition by recurrently discovering attentional regions. In: Proceedings of the IEEE international conference on computer vision. pp. 464–472 (2017)
35. Wu, T., Huang, Q., Liu, Z., Wang, Y., Lin, D.: Distribution-balanced loss for multi-label classification in long-tailed datasets. In: European Conference on Computer Vision. pp. 162–178. Springer (2020)
36. Yang, Y., Zha, K., Chen, Y., Wang, H., Katabi, D.: Delving into deep imbalanced regression. In: International Conference on Machine Learning. pp. 11842–11851. PMLR (2021)
37. Ye, J., He, J., Peng, X., Wu, W., Qiao, Y.: Attention-driven dynamic graph convolutional network for multi-label image recognition. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16. pp. 649–665. Springer (2020)
38. Zhang, X., Wu, Z., Weng, Z., Fu, H., Chen, J., Jiang, Y.G., Davis, L.S.: Videolt: Large-scale long-tailed video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7960–7969 (2021)
39. Zhang, Y., Hooi, B., Hong, L., Feng, J.: Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. Advances in Neural Information Processing Systems **35**, 34077–34090 (2022)
40. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)