# Reference-Guided Gradient Balancing Loss for Long-Tailed Multi-Label Medical Image Classification

Ying-Chih Lin[1*], Hsiao-Chu Peng[1*], Po-Chih Kuo[2] and Yong-Sheng Chen[1(✉)]

[1] Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
[2] Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
{xup6yj.cs11, penghc.cs12, yschen}@nycu.edu.tw, kuopc@cs.nthu.edu.tw

**Abstract.** Multi-label classification (MLC) is a central problem in medical image analysis (MIA), as a single medical image often contains multiple coexisting abnormalities. While ranking-based loss methods have been widely adopted for MLC, their effectiveness in MIA is limited by two major challenges: the dominance of negative gradients during optimization, and the uniform suppression of all negative logits without accounting for class-specific logit distribution. To address these limitations, we propose Reference-Guided Gradient Balancing (RGB) loss, a novel loss function designed for long-tailed multi-label medical image classification, which comprises an adaptive negative suppression loss and a positive enhancement loss. The adaptive negative suppression loss operates based on whether the correction and incorrect prediction coexist within a batch. When both are present, correct predictions serve as references that reveal the class-specific logit distributions and guide the rectification of hard sample logits, thereby refining the decision boundary of each class. Otherwise, negative gradients are adaptively reduced based on prediction confidence. Additionally, the positive enhancement loss incorporates sample difficulty and class frequency to enhance the recognition of low-prevalence diseases. Extensive experiments on NIH ChestXRay14, CXR-LT, and CheXpert demonstrate that RGB consistently outperforms state-of-the-art methods, with further analysis confirming the effectiveness in suppressing negative gradients and increasing the separation between positive and negative logits. The code is available at `https://github.com/xup6YJ/RGB-Loss`.

**Keywords:** Multi-label classification · Ranking-based loss · Long-tailed distribution.

## 1 Introduction

Multi-label classification (MLC) is a fundamental task in medical image analysis (MIA) [5,8,20,25], as a single medical image often contains multiple coexisting

* Equal contribution.

diseases, particularly in modalities such as chest X-rays (CXR), CT scans, and fundus imaging [17]. However, in MLC for MIA, a major challenge arises from the long-tail nature of clinical datasets, where a few high-prevalence categories (head classes) dominate the data, and many rare but clinically important conditions (tail classes) appear only infrequently [13,24,26]. This imbalance biases model learning toward head classes, leading to poor detection of tail categories [6]. Moreover, gradients from rare disease samples are easily overwhelmed by those from abundant categories, making it difficult for the model to learn robust representations for low-prevalence conditions [22,26]. These issues underscore the complexity posed by long-tailed MLC within MIA.

Ranking-based loss methods have emerged as effective approaches for MLC, aiming to encourage the logits of positive labels to exceed those of negative labels either at the instance or class level [3,12,14,21]. LSEP [12] leverages label co-occurrence information in the training set to guide prediction and improve MLC performance. ZLPR [21] extends this idea to natural language processing tasks by introducing a zero margin for both the logits of positive and negative samples to accommodate an arbitrary number of output categories. Despite their effectiveness, these methods are vulnerable to distribution shifts between the training and test sets, particularly for tail classes with limited samples in long-tailed MLC [14]. To address this challenge, DR [14] reformulates LSEP at the class-wise level and formulates C-LSEP, encouraging each class to be distinguished from the others and reducing spurious correlations caused by dependence on label co-occurrence. In addition, DR [14] incorporates a gradient constraint for negative labels to mitigate overconfident negative predictions, demonstrating improved robustness for long-tailed MLC.

However, these methods still exhibit several limitations for MLC in MIA. First, overlooking the dominance of negative gradients. Large clinical datasets often include an extremely frequent "No Finding" category, while individual images contain only a few positive labels. This results in an extreme imbalance in which negative samples vastly exceed positive ones, causing gradients from rare diseases to be overwhelmed and hindering the learning of positive representations. This issue is further exacerbated by false-positive predictions in medical datasets, which yield excessively strong negative gradients and thereby compromise overall performance and model robustness in real-world diagnostic settings. Second, uniformly constraining all negative class logits without accounting for class-specific logit distributions further limits the performance. Lacking awareness of class-dependent logit characteristics, the model is unable to identify hard samples within each class. Consequently, misaligned logits cannot be effectively corrected with respect to the underlying class distributions, hindering the refinement of class-specific decision boundaries.

To address the aforementioned challenges, we propose Reference-Guided Gradient Balancing (RGB) loss, a novel loss function for long-tailed MLC in MIA. Our key contributions are summarized as follows: (1) We leverage correct predictions as references to identify and rectify hard-sample logits, refining class-specific decision boundaries. (2) We analyze and address the strong gradient

issue of false positive predictions that commonly occur in MIA, suppressing negative sample gradients while amplifying positive ones to achieve more reliable disease recognition. (3) Extensive experiments on three benchmark datasets validate the effectiveness of RGB, achieving state-of-the-art (SoTA) performance among existing loss function designs with well-balanced improvements.

## 2 Methodology

### 2.1 Preliminaries

For MLC, given a batch of $M$ instances, each sample $\mathbf{x}$ is associated with a multi-label set $\mathbf{y} = [y_1, y_2, \ldots, y_N]$, where $y_k \in \{0, 1\}$ indicates the absence or presence of the $k$-th class label for a sample, and $N$ represents the total number of label categories.

To mitigate the overconfidence in negative samples and promote the logits of the positive samples being larger than the negative ones within a class, the distributionally robust (DR) loss [14] is designed as:

$$\mathcal{L}_{\mathrm{DR}} = \sum_{k=1}^{N} \log(1 + \sum_{i \in \Omega_{\mathrm{pos}}} \sum_{j \in \Omega_{\mathrm{neg}}} e^{\gamma_1 (s_{j,k} - s_{i,k})} + \sum_{j \in \Omega_{\mathrm{neg}}} e^{\gamma_2 s_{j,k}}), \tag{1}$$

where $\Omega_{\mathrm{pos}} = \{i \mid y_{i,k} = 1\}$, $\Omega_{\mathrm{neg}} = \{j \mid y_{j,k} = 0\}$ denote the positive and the negative set for class $k$ within a batch, respectively. The notation $s$ represents the output logit of the corresponding class, while $\gamma_1$ controls the overall gradient scaling and $\gamma_2$ regulates the gradient contribution from negative samples. Given the dominance of negative samples in medical datasets, the DR [14] loss applies a uniform constraint on all negative logits (last term). While a scaling factor $\gamma_2$ is implemented to adjust the loss magnitude, this uniform treatment ignores the class-specific logit distributions. As a result, negative gradients can remain dominant, especially for hard false positive cases, hindering effective learning from positive samples. We therefore argue that negative gradient contributions should be further and adaptively suppressed, enabling the model to focus more effectively on informative positive samples, particularly for rare disease classes.

### 2.2 Positive Enhancement

To enhance the positive gradients during optimization, we introduce the positive enhancement loss $\mathcal{L}_{\mathrm{pe}}$, which is defined as:

$$\mathcal{L}_{\mathrm{pe}} = \sum_{i \in \Omega_{\mathrm{pos}}} (1 - p_{i,k}) \log(1 + \sum_{j \in \Omega_{\mathrm{neg}}} e^{-\lambda (s_{i,k} - s_{j,k})}), \ \lambda = 1 - \frac{1}{\log_2(D_k)}, \tag{2}$$

where $p_{i,k} = \frac{1}{1+e^{-s_{i,k}}}$ denotes the prediction probability of disease $k$, and $\lambda$ is the designed class-specific weighting coefficient determined by $D_k$, the number of positive samples in the training dataset for disease $k$. As shown in Fig. 1a,

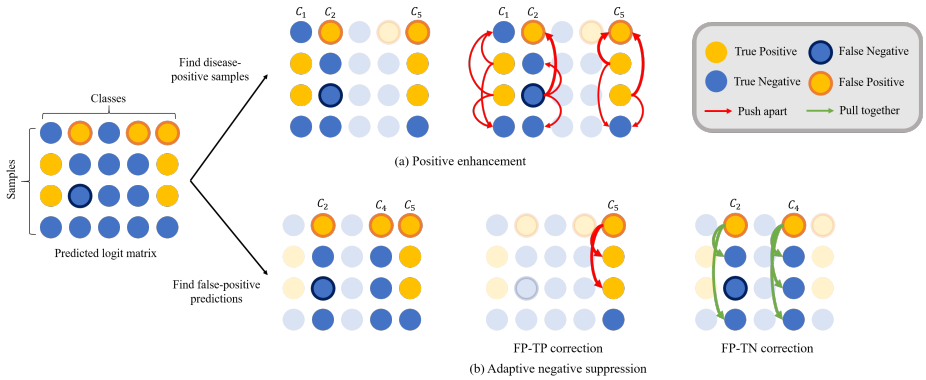



**Fig. 1.** Illustration of $\mathcal{L}_{\text{pe}}$ and $\mathcal{L}_{\text{ref}}$ . (a) The $\mathcal{L}_{\text{pe}}$ explicitly encourages logit separation between positive and negative samples in the logit space. (b) The $\mathcal{L}_{\text{ref}}$ suppresses the negative sample gradient by leveraging the correct predictions. When true positives are present, each false positive logit is driven away from the true positive logits within each class; in their absence, it is guided toward the true negative logits.

this formulation promotes logit separation between positive and negative samples within each class and produces larger gradients for incorrect predictions. The incorporation of prediction confidence in $\mathcal{L}_{\text{pe}}$ further emphasizes hard samples, particularly those that exhibit overly confident false negative predictions, assigning them larger corrective penalties. Moreover, the coefficient $\lambda$ accounts for the long-tailed distribution by assigning smaller values to classes with limited samples, thereby increasing the gradient contribution of rare diseases and preventing frequent categories from dominating the optimization process. This design jointly leverages sample difficulty and class frequency, enabling the model to better learning from the long-tailed distribution and improve recognition of low-prevalence diseases.

### 2.3 Adaptive Negative Suppression

The overall designed adaptive negative suppression loss is composed of two complementary components:

$$\mathcal{L}_{\text{ans}} = \mathcal{L}_{\text{ref}} + \mathcal{L}_{\text{sup}} \ . \tag{3}$$

To mitigate the dominant negative gradients, we focus on hard samples and adaptively suppress negative gradients. Specifically, we target the strong gradients induced by false positive predictions in MIA and introduce the reference loss $\mathcal{L}_{\text{ref}}$, which can be defined as:

$$\mathcal{L}_{\text{ref}} = \sum_{j \in \Omega_k^{\text{FP}}} p_j^{\gamma_1} \log(1 + \mathbb{1}_{\exists TP} \sum_{i \in \Omega_k^{\text{TP}}} e^{-(s_{i,k} - s_{j,k})} + \mathbb{1}_{\nexists TP} \sum_{l \in \Omega_k^{\text{TN}}} e^{s_{j,k} - s_{l,k}}) \,, \tag{4}$$

where $\mathbb{1}$ denotes the indicator function of whether true positives exist within a batch and $\gamma_1$ regulates the contribution of false prediction confidence. The nota-

tion $\Omega_{\mathrm{FP}} = \left\{ i \mid y_{i,k} = 0 \ \cap \hat{y}_{i,k} = 1 \right\}$ denotes the set of false positive predictions for class $k$, the corresponding true positive and true negative sets are defined analogously. Here, $\hat{y}_{i,k}$ denotes the prediction of the sample $i$ for class $k$. The proposed $\mathcal{L}_{\mathrm{ref}}$ prioritizes guiding false positive logits away from the true positive logits; when no true positives are present, it instead guides them toward the true negative logits, enabling a shift back to the correct side of the decision boundary (Fig. 1b). In contrast to DR [14], which uniformly suppresses negative sample gradients yet can still yield excessively large negative gradient magnitudes, our method ingeniously leverages correct predictions as reference signals to balance the gradients of positive and negative samples.

By incorporating these reference logits, the model is enabled to characterize class-specific logit distributions and further identify hard samples within each class, thereby facilitating targeted rectification of false positive logits based on their relative positions with respect to correct prediction logits in logit space. This reference-guided strategy allows for selective suppression of harmful negative gradients while establishing natural, class-dependent thresholds for correction. In addition, the involvement of positive samples in this process simultaneously amplifies positive gradient contributions, facilitating more effective learning of positive features. Consequently, $\mathcal{L}_{\mathrm{ref}}$ enables the model to correct the mispredicted logits more accurately, mitigates the influence of dominant negative gradients, and strengthens the positive gradients.

On the other hand, when a class contains only false positive predictions or only true negative predictions, without a valid reference prediction, we control the negative sample gradients through the suppression loss $\mathcal{L}_{\mathrm{sup}}$:

$$\mathcal{L}_{\mathrm{sup}} = \sum_{j} \begin{cases} p_j^{\gamma_1} \log(1 + e^{s_{j,k}}) & \text{if } j \in \Omega_k^{\mathrm{FP}},\ |\Omega_k^{\mathrm{TP}}| = 0,\ |\Omega_k^{\mathrm{TN}}| = 0 \\ p_j^{\gamma_2} \log(1 + e^{s_{j,k}}) & \text{if } j \in \Omega_k^{\mathrm{TN}}, \end{cases} \tag{5}$$

where the modulation terms $p_j^{\gamma_1}$ and $p_j^{\gamma_2}$ regulate the suppression strength based on prediction results, and we set $\gamma_1 \leq \gamma_2$ to ensure that false predictions receive stronger loss contributions than correct predictions. This design dynamically suppresses negative-sample gradients within each batch according to their confidence, prioritizing the reduction of false positive gradients while imposing weaker constraints on true negatives. Finally, the proposed overall loss function $\mathcal{L}_{\mathrm{rgb}}$ is defined as:

$$\mathcal{L}_{\mathrm{rgb}} = \sum_{k=1}^{N} \mathcal{L}_{\mathrm{pe}} + \mathcal{L}_{\mathrm{ans}}\,. \tag{6}$$

The overall design of $\mathcal{L}_{\mathrm{rgb}}$ suppresses negative sample gradients adaptively while enhancing positive ones, facilitating positive representation learning and accommodating the inherent long-tailed distribution present in medical datasets.

## 3 Experiments

**Datasets and preprocessing.** In this study, we employ three large-scale CXR datasets to evaluate the proposed RGB loss. The NIH ChestXRay14 (CXR-14)

**Table 1.** Results of performance comparison in all/head/medium/tail classes (%). The best scores are highlighted in bold, and the second-best in blue.

| Method | CXR-14 [23] | | | | | | | CXR-LT [15] | | | | | | | CheXpert [9] | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | maF1 | | | | mAUC | ma-R | ma-P | maF1 | | | | mAUC | ma-R | ma-P | maF1 | | | | mAUC | ma-R | ma-P |
| | All | Head | Medium | Tail | | | | All | Head | Medium | Tail | | | | All | Head | Medium | Tail | | | |
| BCE | 16.54 | 29.17 | 15.58 | 6.51 | 82.99 | 11.78 | 39.86 | 14.78 | 42.15 | 12.96 | 7.01 | 78.88 | 11.73 | 32.39 | 35.59 | 75.50 | 29.57 | 9.72 | 78.40 | 31.18 | 56.84 |
| Focal [16] | 17.68 | 28.95 | 16.51 | 9.54 | 82.76 | 12.80 | **40.15** | 12.41 | 39.47 | 8.48 | 5.38 | **79.38** | 9.62 | 33.10 | 34.43 | 75.30 | 29.66 | 4.70 | 78.27 | 31.04 | 57.76 |
| LSEP [12] | 23.12 | 31.30 | 22.37 | 16.96 | 78.60 | 61.53 | 14.81 | 18.16 | 41.90 | 15.32 | 11.79 | 76.76 | 8.79 | 33.93 | 44.71 | 74.24 | 43.01 | 19.13 | 75.64 | 60.13 | 36.49 |
| CBFocal [2] | 16.11 | 25.99 | 14.28 | 11.09 | 82.91 | 11.72 | 37.82 | 11.10 | 36.86 | 4.77 | 5.19 | 78.34 | 3.39 | **34.09** | 35.27 | 75.07 | 31.73 | 3.72 | 78.23 | 31.30 | **58.52** |
| LDAM [1] | 20.26 | 27.09 | 20.94 | 11.63 | 77.97 | 60.43 | 12.34 | 18.89 | 40.91 | 16.31 | 12.97 | 77.07 | **48.09** | 12.61 | 44.35 | 73.07 | 43.15 | 18.45 | 75.41 | 62.43 | 35.42 |
| DBFocal [24] | 6.00 | 3.68 | 5.28 | 10.26 | 82.79 | 3.60 | 33.84 | 5.47 | 17.44 | 2.21 | 2.82 | 79.14 | 43.45 | 12.39 | 18.42 | 46.30 | 13.66 | 1.67 | 78.14 | 13.14 | 56.76 |
| ASL [19] | 30.25 | 43.88 | 30.60 | 15.69 | 82.49 | 35.29 | 29.17 | 17.20 | 48.86 | 15.40 | 8.12 | 79.19 | 17.75 | 28.53 | 47.15 | 75.1 | 45.42 | 23.27 | 78.61 | 67.94 | 36.79 |
| TwoWay [11] | 26.01 | 33.58 | 27.53 | 14.39 | 82.16 | 53.17 | 18.50 | 19.57 | 42.77 | 19.06 | 12.66 | 78.21 | 37.72 | 15.06 | 36.05 | 76.81 | 30.91 | 7.28 | 77.90 | 36.17 | 55.26 |
| RAL [18] | 22.11 | 33.96 | 20.91 | 13.47 | 82.91 | **70.06** | 13.44 | 20.67 | 47.88 | 20.81 | 11.46 | 78.77 | 26.34 | 20.67 | 46.05 | 74.59 | 44.32 | 21.56 | 78.32 | **70.23** | 35.26 |
| DR [14] | 21.63 | 30.04 | 24.12 | 6.57 | 81.80 | 38.84 | 25.01 | 10.92 | 36.27 | 7.21 | 4.33 | 78.12 | 12.49 | 20.48 | 27.75 | 52.56 | 22.50 | 15.20 | 77.82 | 33.29 | 47.93 |
| RGB (Ours) | **33.27** | **45.35** | **33.71** | **20.01** | **84.04** | 39.11 | 30.06 | **22.45** | **51.75** | **23.65** | **13.16** | 79.33 | 24.01 | 24.12 | **48.02** | **78.08** | **45.59** | **23.63** | **78.94** | 52.23 | 45.58 |

dataset [23] includes 112,120 frontal images across 14 disease classes. The CXR-LT dataset [15], an extension of MIMIC [10], contains 377,110 images with 37 disease labels (excluding the "Pleural Other" and "Support Device" classes) for task 1. The CheXpert dataset [9] provides 224,316 images covering 13 disease categories. All experiments are conducted using the standard frontal view images. Each dataset is randomly partitioned at the patient level into training, validation, and test sets with ratios of 70%, 10%, and 20%, respectively, ensuring no patient information leakage. All images are resized into $224 \times 224$. Data augmentation involves random horizontal flipping and random rotation to enhance robustness.

**Implementation details.** We use DenseNet121 [7] pretrained on ImageNet [4] as the backbone. The model is trained for 30 epochs with a batch size of 256 using the Adam optimizer with an initial learning rate of $5 \times 10^{-4}$, adjusted by a scheduler that reduces the rate by 0.1 when the validation loss plateaus for three epochs. The prediction probability threshold is fixed at 0.5 across all experiments. The hyperparameters $\gamma_1$ and $\gamma_2$ are selected via grid search, with stable performance within $\pm 0.1$. We set $\gamma_1$ and $\gamma_2$ to 2.4 and 3 for CXR-LT, 3 and 3 for CXR-14, and 1.9 and 2 for CheXpert. All implementations are conducted in PyTorch and trained on an NVIDIA 4090 GPU.

**Performance comparison.** We conduct a comprehensive comparison of RGB against nine SoTA multi-label classification loss methods, including Focal [16], LSEP [12], CBFocal [2], LDAM [1], DBFocal [24], ASL [19], TWML [11], RAL [18], and DR [14]. To further evaluate the effectiveness of RGB on low-prevalence diseases, we partition each dataset into head, medium, and tail classes based on the "elbow" point of the descending prevalence distribution. Due to dataset-specific prevalence shifts, different thresholds are adopted for each dataset. For CXR-14, diseases with prevalence above 10% are categorized as head classes and those below 1.5% as tail classes (3/8/3 for head, medium, and tail). For CXR-LT, we use thresholds of 10% and 2% (7/7/23), whereas for CheXpert, we use 40% and 5% (3/7/3). Performance is measured using four key metrics: macro-F1, macro-Recall, macro-Precision, and mAUC. To ensure a fair comparison, all methods are trained under identical settings. Source codes are obtained from official repositories where available.

**Table 2.** Ablation study (%).

| Dataset | Method | ma-F1 | mi-F1 | mAUC | ma-R | mi-R | ma-P | mi-P |
|---|---|---|---|---|---|---|---|---|
| CXR14 | $\mathcal{L}_{\text{c-lsep}}$ | 26.37 | 38.58 | 82.71 | 31.54 | 47.75 | 26.18 | 32.36 |
| | $\mathcal{L}_{\text{c-lsep}} + \mathcal{L}_{\text{ans}}$ | 28.02 | 26.69 | 82.94 | 36.77 | 34.80 | 29.47 | 25.89 |
| | $\mathcal{L}_{\text{pe}} + \mathcal{L}_{\text{ans}}$ (Ours) | **33.27** | **39.04** | **84.04** | **39.11** | **48.11** | **30.06** | **32.85** |
| CXR-LT | $\mathcal{L}_{\text{c-lsep}}$ | 16.34 | 51.59 | 77.45 | 19.69 | **62.80** | 18.63 | 43.77 |
| | $\mathcal{L}_{\text{c-lsep}} + \mathcal{L}_{\text{ans}}$ | 17.10 | 28.00 | 78.54 | 23.32 | 30.16 | 22.76 | 30.16 |
| | $\mathcal{L}_{\text{pe}} + \mathcal{L}_{\text{ans}}$ (Ours) | **22.45** | **51.89** | **79.33** | **24.01** | 59.16 | **24.12** | **46.21** |
| CheXpert | $\mathcal{L}_{\text{c-lsep}}$ | 32.27 | 40.30 | 77.81 | 45.85 | 49.87 | 33.53 | 35.95 |
| | $\mathcal{L}_{\text{c-lsep}} + \mathcal{L}_{\text{ans}}$ | 35.61 | 43.75 | 77.84 | 46.79 | 50.10 | 35.06 | 47.23 |
| | $\mathcal{L}_{\text{pe}} + \mathcal{L}_{\text{ans}}$ (Ours) | **48.02** | **64.52** | **78.94** | **52.23** | **71.83** | **45.58** | **58.55** |

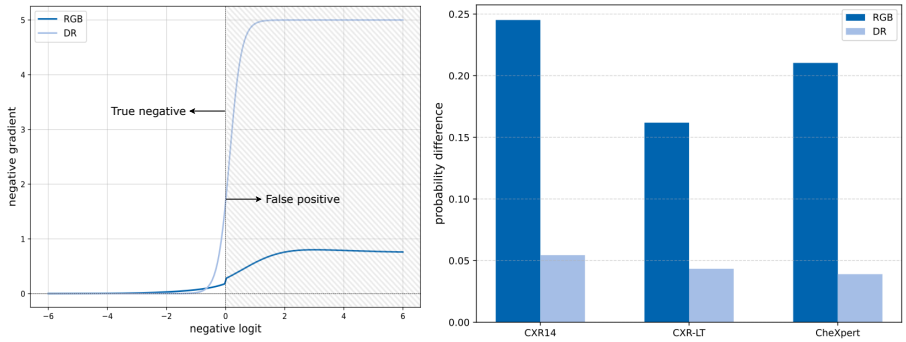
As detailed in Table 1, the proposed RGB loss consistently achieves SoTA performance across all three datasets, achieving the highest macro-F1 and mAUC scores of 33.27% and 84.04% on CXR-14, 22.45% and 79.33% on CXR-LT, and 48.02% and 78.94% on CheXpert, respectively. The consistent improvements in macro-level metrics indicate that RGB is robust to varying disease prevalence and diagnostic difficulty. Compared to the second-best results, RGB improves the overall macro-F1 score by 3.02% on CXR-14, 1.78% on CXR-LT, and 0.87% on CheXpert, respectively, underscoring its effectiveness for long-tailed MLC in MIA. Moreover, a detailed analysis across head, medium, and tail categories shows that RGB consistently achieves the highest macro-F1 scores in all disease groups. Notably, RGB attains the best tail-class macro-F1 on all datasets, highlighting its superior capability in learning from long-tailed distributions and outperforming existing SoTA loss designs.

**Ablation study.** To assess the contribution of each component in RGB, we conduct a comprehensive ablation study across all datasets, as summarized in Table 2. Utilizing $\mathcal{L}_{\text{c-lsep}}$, which was introduced in DR [14], as the baseline, incorporating $\mathcal{L}_{\text{ans}}$ effectively reduces false positives while improving recall, resulting in consistent gains in macro-F1 and mAUC across all datasets. These results confirm that suppressing dominant negative gradients enables the model to better capture positive representations. Furthermore, replacing $\mathcal{L}_{\text{c-lsep}}$ with the proposed $\mathcal{L}_{\text{pe}}$ yields consistent improvements across nearly all evaluation metrics on the three datasets. This demonstrates the effectiveness of our design in jointly suppressing negative gradients and amplifying gradient signals for rare diseases, thereby improving overall performance.

We further evaluate different strategies for computing $\mathcal{L}_{\text{ref}}$ by varying the reference set, as shown in Table 3. The "TP" setting compares false positives only with true positives to maximize separation, whereas the "TN" setting compares them only with true negatives to minimize separation. Our conditional "TP or TN" approach leverages true positives when available and defaults to true negatives otherwise, enabling adaptive reference correction. The results show that "TP or TN" consistently achieves the best performance across all three datasets, providing a robust and flexible solution to balance gradient signals based on correct predictions.

**Table 3.** Comparison of different reference sets in $\mathcal{L}_{\text{ref}}$ (%).

| Dataset | Method | ma-F1 | mi-F1 | mAUC | ma-R | mi-R | ma-P | mi-P |
|---|---|---|---|---|---|---|---|---|
| CXR14 | TP | 32.88 | 38.74 | 83.94 | 39.07 | 48.44 | 28.99 | 32.27 |
| | TN | 7.11 | 9.27 | 80.94 | 4.16 | 5.17 | 29.73 | **45.02** |
| | TP or TN | **33.27** | **39.04** | **84.04** | **39.11** | **48.11** | **30.06** | 32.85 |
| CXR-LT | TP | **22.56** | **51.89** | 79.24 | **24.57** | 57.77 | 21.90 | 47.10 |
| | TN | 5.54 | 25.57 | 75.59 | 4.10 | 16.48 | **25.12** | **57.01** |
| | TP or TN | 22.45 | **51.89** | **79.33** | 24.01 | **59.16** | 24.12 | 46.21 |
| CheXpert | TP | 47.85 | 64.39 | 78.81 | 51.92 | 71.31 | 45.36 | **58.70** |
| | TN | 3.36 | 3.38 | 76.35 | 1.83 | 1.73 | 51.93 | 72.91 |
| | TP or TN | **48.02** | **64.52** | **78.94** | **52.23** | **71.83** | **45.58** | 58.55 |




(a) Comparison of negative sample gradients (b) Comparison of the average predicted probaiblity difference

**Fig. 2.** Gradient and probability difference comparison between DR and ours.

**Visualization.** To provide an intuitive comparison between the proposed method and the latest ranking-based loss (DR [14]), we visualize the gradients of negative logits (Fig. 2a). Rather than uniformly constraining all negative samples, RGB loss selectively focuses on hard negative samples, effectively reducing the negative gradients arising from false positive predictions. It is evident that the proposed method mitigates negative gradient dominance during optimization and facilitates more effective learning from positive samples. Additionally, as ranking-based losses aim to increase the separation between positive and negative logits, we further compare the average predicted probability difference between positive and negative samples (Fig. 2b). The analysis demonstrates that RGB loss consistently produces a larger separation across all datasets compared to DR [14], indicating improved separation between positive and negative logits and a more discriminative decision boundary.

## 4 Conclusion

In this paper, we present RGB, a novel loss function for long-tailed MLC in MIA. RGB alleviates dominant negative gradients through adaptive suppression

of hard negative samples while amplifying positive gradient contributions. To further enhance the learning of rare diseases, we introduce a positive enhancement loss to strengthen positive gradients from rare diseases. Extensive experiments across three CXR datasets show that RGB achieves superior performance compared to SoTA loss function methods, with additional analysis confirming its effectiveness in reducing negative gradient dominance and improving the separation between positive and negative logits.

**Acknowledgments.** This work was supported in part by the National Science and Technology Council, Taiwan, under Grants NSTC 114-2314-B-A49-018 and 114-2628-E-007-008-MY4. We appreciate the National Center for High-performance Computing (NCHC) for providing computational and storage resources.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. Advances in neural information processing systems **32** (2019)
2. 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)
3. Dari, E., Yesilkaynak, V.B., Mertan, A., Unal, G.: Rlsep: Learning label ranks for multi-label classification. arXiv preprint arXiv:2212.04022 (2022)
4. 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)
5. Holste, G., Jiang, Z., Jaiswal, A., Hanna, M., Minkowitz, S., Legasto, A.C., Escalon, J.G., Steinberger, S., Bittman, M., Shen, T.C., et al.: How does pruning impact long-tailed multi-label medical image classifiers? In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 663–673. Springer (2023)
6. Holste, G., Zhou, Y., Wang, S., Jaiswal, A., Lin, M., Zhuge, S., Yang, Y., Kim, D., Nguyen-Mau, T.H., Tran, M.T., et al.: Towards long-tailed, multi-label disease classification from chest x-ray: Overview of the cxr-lt challenge. Medical Image Analysis **97**, 103224 (2024)
7. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
8. Huang, J., Chen, Z.M., Chen, Z.M., Zhang, X., Zhang, X., Ge, Y., Ge, Y., Ye, L., Ye, L., Zhang, G., et al.: Label decoupling and reconstruction: A two-stage training framework for long-tailed multi-label medical image recognition. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 2861–2869 (2024)
9. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)

10. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
11. Kobayashi, T.: Two-way multi-label loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7476–7485 (2023)
12. 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)
13. 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)
14. Lin, D., Peng, T., Chen, R., Xie, X., Qin, X., Cui, Z.: Distributionally robust loss for long-tailed multi-label image classification. In: European Conference on Computer Vision. pp. 417–433. Springer (2024)
15. Lin, M., Holste, G., Wang, S., Zhou, Y., Wei, Y., Banerjee, I., Chen, P., Dai, T., Du, Y., Dvornek, N.C., et al.: Cxr-lt 2024: A miccai challenge on long-tailed, multi-label, and zero-shot disease classification from chest x-ray. Medical Image Analysis p. 103739 (2025)
16. 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)
17. Lin, Y.C., Chen, Y.S.: Weighted stratification in multi-label contrastive learning for long-tailed medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 677–687. Springer (2025)
18. Park, W., Park, I., Kim, S., Ryu, J.: Robust asymmetric loss for multi-label long-tailed learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2711–2720 (2023)
19. 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)
20. Seo, H., Lee, M., Cheong, W., Yoon, H., Kim, S., Kang, M.: Enhancing multi-label long-tailed classification on chest x-rays through ml-gcn augmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2747–2756 (2023)
21. 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)
22. Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C., Yan, J.: Equalization loss for long-tailed object recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11662–11671 (2020)
23. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)
24. 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)
25. Xiao, J., Li, S., Lin, T., Zhu, J., Yuan, X., Feng, D.D., Sheng, B.: Multi-label chest x-ray image classification with single positive labels. IEEE transactions on medical imaging **43**(12), 4404–4418 (2024)

26. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. IEEE transactions on pattern analysis and machine intelligence **45**(9), 10795–10816 (2023)