# BPaCo: Balanced Parametric Contrastive Learning for Long-tailed Medical Image Classification

Zhiyuan Cai[1,2], Tianyunxi Wei[1], Li Lin[1,3], Hao Chen[2(✉)], and Xiaoying Tang[1,4(✉)]

[1] Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen, China
tangxy@sustech.edu.cn
[2] Department of Computer Science Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China
jhc@ust.hk
[3] Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong SAR, China
[4] Jiaxing Research Institute, Southern University of Science and Technology, Jiaxing, China

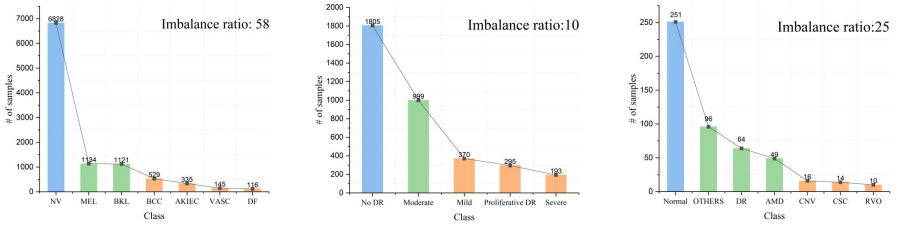
**Abstract.** Medical image classification is an essential medical image analysis tasks. However, due to data scarcity of rare diseases in clinical scenarios, the acquired medical image datasets may exhibit long-tailed distributions. Previous works employ class re-balancing to address this issue yet the representation is usually not discriminative enough. Inspired by contrastive learning's power in representation learning, in this paper, we propose and validate a contrastive learning based framework, named **B**alanced **Pa**rametric **Co**ntrastive learning (BPaCo), to tackle long-tailed medical image classification. There are three key components in BPaCo: across-batch class-averaging to balance the gradient contribution from negative classes; hybrid class-complement to have all classes appear in every mini-batch for discriminative prototypes; cross-entropy logit compensation to formulate an end-to-end classification framework with even stronger feature representations. Our BPaCo shows outstanding classification performance and high computational efficiency on three highly-imbalanced medical image classification datasets. The source code is available at https://github.com/Davidczy/BPaCo.

**Keywords:** Long-tailed · Contrastive learning · Medical image classification.

## 1 Introduction

Deep neural networks have achieved remarkable success in medical image analysis tasks, such as disease diagnosis [1] and lesion detection [14], benefiting from

large-scale, balanced, and high-quality labeled data. However, for some rare diseases, it is challenging to collect comprehensive datasets. Consequently, the acquired medical image datasets may exhibit long-tailed distribution patterns [19], where head classes dominate most data and tailed classes have only a few samples (Fig. 1). Deep models trained on such imbalanced data typically perform poorly on balanced testing data, especially for rare classes. Addressing this issue and improving the recognition performance with imbalanced data present significant challenges for modern deep learning methods. To tackle this problem, early meth-




ISIC2018 APTOS2019 OCTA500

**Fig. 1.** Long-tailed distributions of the ISIC2018, APTOS2019 and OCTA500 datasets. The imbalance ratio is computed as $N_{max}/N_{min}$, where $N$ is the number of samples in each class.

ods mainly focus on class re-balancing [2, 17], information augmentation [16, 30] and module improvement [6, 31]. Class re-balancing methods aim to adjust the distribution of the training samples to achieve balanced data over all categories, e.g., over-sampling on tailed classes [8] or under-sampling on head classes [23]. Even though the performance on tailed classes gets improved, such techniques sacrifice the performance on head classes, inducing a performance seesaw [29]. Information augmentation methods seek to introduce additional information into model training to improve a model's performance in long-tailed learning, e.g., head-to-tail knowledge transfer [18, 27], knowledge distillation [11, 13] or data augmentation [24, 28]. Benefiting from introducing additional knowledge, these methods avoid sacrificing head classes' performance while improving the tailed classes' performance. However, information augmentation methods require additional training time and professional domain knowledge may be required.

Recently, methods aiming to improve network modules become the mainstream solution to long-tailed tasks. In general, these methods modify different modules of a model of interest to boost its performance, e.g., representation learning for improving the feature extractor [6, 31], classifier designing for enhancing the performance [9] or decoupled training for promoting both the feature extractor and the classifier [20]. Since representation learning is the most remarkable capability of deep models, supervised contrastive learning (SCL) methods exhibit great potential, e.g., generalized parametric contrastive learning (GPaCo) [7] and balanced contrastive learning (BCL) [31]. Despite that SCL achieves better performance than cross-entropy based supervision on large-scale

classification problems, some recent works indicate that high-frequency classes dominate SCL when representing imbalanced data, which may still lead to unsatisfactory performance. BCL alleviates such issues by introducing in-batch class-averaging, class-complement and logit compensation to attain a regular simplex configuration, yet its performance is still restricted by its mini-batch's size. GPaCo relieves the performance degradation problem by introducing a set of class-wise learnable centers, but it sacrifices the gradient contribution of the in-batch samples by assigning a small weight to the in-batch loss.

Here, in this paper, we propose **B**alanced **Pa**rametric **Co**ntrastive learning (BPaCo) to tackle the long-tailed medical image classification task, with outstanding performance having been observed. Assisted by a well-designed BPaCo loss and the cross-entropy loss's logit compensation, BPaCo simultaneously improves representation learning and classifier learning in an end-to-end manner. Collectively, our main contributions are three-fold: (1) We analyze the space for improvement in the BCL and GPaCo losses, and propose a novel BPaCo loss with across-batch class-averaging to avoid the dominance of head classes. (2) We combine the class-wise learnable parametric centers with the weights of the linear classifier to form a more comprehensive hybrid class-complement module, which provides different types of class-wise prototypes to enhance representation learning. (3) Through rigorous experimental evaluations on three medical image datasets, we show that BPaCo exhibits better performance and higher computational efficiency than various competitive baselines.

## 2 Methodology

### 2.1 Preliminaries

For a medical image classification task, we aim to learn a complex function mapping $\phi$ from an input space $\mathcal{X}$ to a target space $\mathcal{Y} = \{1, 2, ..., K\}$, where $K$ is the number of classes. The function $\phi$ is usually composed of an encoder $f : \mathcal{X} \to \mathcal{Z} \in \mathbb{R}^h$ and a linear classifier $w : \mathcal{Z} \to \mathcal{Y}$. We here aim to learn a good $f$ for better performance in long-tailed classification through representation learning.

### 2.2 Framework

The pipeline of our proposed framework is shown in Fig. 2. It consists of a classification branch and a contrastive learning branch. Both branches are simultaneously trained and share the same feature extractor (encoder). Similar to BCL, BPaCo is end-to-end instead of two-stage. We adopt two different augmentation strategies for an input of interest. The two different views of the input ($v_1$ and $v_2$) are fed into the same encoder to output the corresponding features $f_1$ and $f_2$. After that, in the contrastive learning branch, $f_1$ and $f_2$ are both fed into an MLP with one hidden layer to obtain representations $z_1$ and $z_2$. $f_1$ is also fed into an identity mapping (IM) module to obtain the class-wise parametric

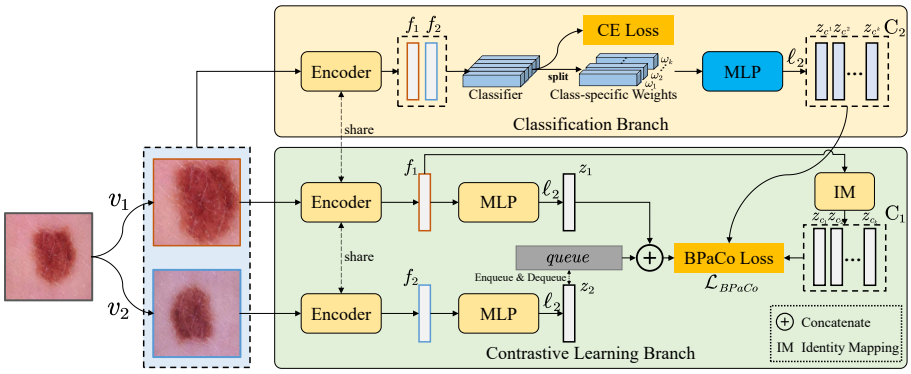



**Fig. 2.** The overall framwork of BPaCo. It consists of a classification branch and a contrastive learning branch. Both branches are simultaneously trained and share the same feature extractor (encoder). $C_1$ and $C_2$ are different types of prototypes.

learnable centers $C_1 = \{z_{c_1}, z_{c_2}, ..., z_{c_k}\}$. $z_2$ enqueues and dequeues the *queue*, keeping in line with MoCo [10]. In the classification branch, $f_1$ and $f_2$ are concatenated and fed into a classifier to conduct cross-entropy logit compensation. Since the weights of the linear classifier are co-linear with the class prototypes to which the classes collapse [31], we can easily obtain the class-specific weights $\omega_1, \omega_2, ..., \omega_k$. Those weights then go through a nonlinear MLP to obtain the centers $C_2 = \{z_{c^1}, z_{c^2}, ..., z_{c^k}\}$. Finally, the two sets of centers ($C_1$ and $C_2$) as well as the representations of all samples in the mini-batch and the *queue* are used to jointly calculate the BPaCo Loss $\mathcal{L}_{BPaCo}$. All representations going through each MLP are $\ell_2$ normalized to ensure the feature space is a unit hypersphere. In our framework, the BPaCo loss is the key to solve the long-tailed classification problem, consisting of two important compositions, namely across-batch class-averaging and hybrid class-complement. We now delve into their details.

### 2.3 Across-batch Class-averaging

SCL builds its basis on self-supervised contrastive loss by making use of the label information. Although SCL allows class-wise gradient contribution, it cannot prevent the dominance of head classes and is not applicable to long-tailed data. To make SCL accommodate long-tailed data, BCL introduces in-batch class-averaging to formulate the BCL loss as

$$\mathcal{L}_{BCL} = -\sum_{i \in B_y} \frac{1}{|B_y| - 1} \sum_{p \in \{B_y \setminus \{i\}\}} \log \frac{\exp\left(z_p \cdot z_i\right)}{\sum_{j \in \mathcal{Y}} \frac{1}{|B_j|} \sum_{k \in B_j} \exp\left(z_k \cdot z_i\right)}, \tag{1}$$

where $B_y$ is a subset of $B$ that contains all samples of class $y$. $|\cdot|$ stands for the number of samples in a specific set. The key idea of in-batch class-averaging is to average the instances of each same class in a mini-batch so that each class has similar contribution during model optimization. However, such averaging is

only based on in-batch representations, so the performance is strictly restricted by the mini-batch's size.

Our across-batch class-averaging overcomes the restriction of the mini-batch's size by introducing a *queue* in MoCo [10] to our framework. The *queue* enqueues the new encoded representations and dequeues the older ones. All the encoded representations in the *queue* and representations in the mini-batch are jointly used to calculate the contrastive loss at each training epoch. The implementation of across-batch class-averaging forms as follows

$$\mathcal{L}_{ave} = -\sum_{i \in A_y} \frac{1}{|A_y| - 1} \sum_{p \in \{A_y \setminus \{i\}\}} \log \frac{\exp\left(z_p \cdot z_i\right)}{\sum_{j \in \mathcal{Y}} \frac{1}{|A_j|} \sum_{k \in A_j} \exp\left(z_k \cdot z_i\right)}, \tag{2}$$

where $A_y$ is a subset of $A$ that contains all samples of class $y$ and $A$ is a combination of in-batch representations as well as representations in the *queue*. Below, following notations in preliminaries, we give the loss's lower bound, which guarantees on the optimality of model training, after performing across-batch class-averaging.

**Theorem 1.** *Assuming the normalization function is applied for feature embedding, let* $Z = (z_1, ..., z_N) \in \mathcal{Z}^N$ *be an* $N$ *point configuration with labels* $Y = (y_1, ..., y_N) \in [K]^N$*, where* $\mathcal{Z} = \{z \in \mathbb{R}^h : ||z|| = 1\}$*,* $\mathcal{Y}_A \subseteq [K]$ *denotes the set of classes that appear within* $A$*. The class-specific batch-wise loss after performing across-batch class averaging is bounded by*

$$\begin{aligned} \mathcal{L}_{ave}(Z; Y, A, y) \geq \sum_{i \in A_y} log(1 + (|\mathcal{Y}_A| - 1) \times exp(\frac{1}{|\mathcal{Y}_A| - 1} \\ \sum_{q \in \mathcal{Y}_A \setminus \{y\}} \frac{1}{|A_q|} \sum_{k \in A_q} z_i \cdot z_k - \frac{1}{|A_y - 1|} \sum_{j \in A_y \setminus \{i\}} z_i \cdot z_j)). \end{aligned} \tag{3}$$

*Proof. See the Supplementary Material.*

Assisted by the *queue*, our $\mathcal{L}_{BPaCo}$ breaks the limitation of the mini-batch's size, achieving similarly excellent performance even under small batch settings (See Table 3). To be more specific, BPaCo can perform contrastive learning with in-the-*queue* representations that are obtained from previous batches, so that near-optimal performance can be attained even with a mini-batch of a small size, indicating its being computationally efficient.

### 2.4 Hybrid Class-complement

Even though across-batch class-averaging can maintain class-wise balance during training, samples in the tailed classes still occupy only a small portion in the *queue*, or even none in extreme data imbalance situations. To have all classes appear in every mini-batch, we introduce hybrid class-complement for balanced contrastive learning. We decouple the hybrid class-complement module into class-wise parametric learnable centers $C_1$ as a contrastive learning term and linear classifier's weights $C_2$ as a logit compensation term. The logit compensation term aims to eliminate the bias caused by the data imbalance and

learn the rectification of the boundary [2]. There are various ways to implement logit compensation, through the cross-entropy loss or the LDAM loss [2]. We use cross-entropy logit compensation in our framework. Getting combined with Eq. 2 and hybrid class-complement, our BPaCo loss is formulated as

$$\mathcal{L}_{BPaCo} = -\sum_{i \in A_y} \frac{1}{|A_y| - 1} \sum_{p \in \{A_y \setminus \{i\}\} \cup \{c_y\} \cup \{c^y\}} \log \frac{\exp\left(z_p \cdot z_i\right)}{\sum_{j \in \mathcal{Y}} \frac{1}{|A_j|} \sum_{k \in A_j \cup \{c_j\} \cup \{c^j\}} \exp\left(z_k \cdot z_i\right)}. \tag{4}$$

After hybrid class-complement, we have the overall lower bound of $\mathcal{L}_{BPaCo}$.

**Theorem 2.** *Let $Z, Y$ be defined as in Theorem 1, if we have $\mathcal{Y}_a = \mathcal{Y}$ for every $A$, the overall loss is given by*

$$\mathcal{L}_{BPaCo}\left(Z; Y\right) \geq |\mathcal{D}| \log \left(1 + (K-1) \exp\left(-\frac{K}{K-1}\right)\right), \tag{5}$$

*where $\mathcal{D}$ represents the whole dataset and $K$ is the number of classes.*
*Proof. See the Supplementary Material.*

From Eq. 5, the lower bound of our $\mathcal{L}_{BPaCo}$ is a constant that is only related to the number of classes and the size of the whole dataset but unrelated to the number of samples in each class. By applying both across-batch class-averaging and hybrid class-complement, we avoid the model's preference to head classes. Finally, when optimizing with cross-entropy logit compensation, we have the following loss for training

$$\mathcal{L} = \mathcal{L}_{ce} + \beta \mathcal{L}_{BPaCo}, \tag{6}$$

where $\beta$ is a hyperparameter that control the impacts of $\mathcal{L}_{BPaCo}$.

## 3 Experiment

**Datasets and Evaluations.** We conduct experiments on three publicly available datasets. The ISIC2018 [3] dataset is published by the International Skin Imaging Collaboration (ISIC) organizers, which is a large-scale dataset of dermoscopy images. The APTOS2019 [21] dataset is a fundus dataset from the APTOS 2019 blindness detection competition. And the OCTA500 [15] dataset contains multiple optical coherence tomography angiography (OCTA) modalities. In our experiment, we only use the OCT en-face (Full) modality in OCTA500 to perform long-tailed classification. We illustrate the distribution details of these datasets in Fig. 1, and more details are provided in Table A1 of the supplementary material. Following the experimental setup for long-tailed problems, we split the original dataset into train and test sets with a ratio of 9:1 and maintain balanced sample sizes for all classes in the test sets. All experimental results are reported with the evaluation metrics of accuracy and F1-score.

As mentioned above, there are three main categories of methods for long-tailed problems: class re-balancing (CR), information augmentation (IA) and

**Table 1.** Comparisons with state-of-the-art methods. The best ones are bolded while the second best ones are underlined. Keys: CR - Class Re-balancing; IA - Information Augmentation; MI - Module Improvement; (∘ - resample) - training ∘ first and then another classifier with balanced sampling.

| Method | Publication | CR | IA | MI | ISIC2018 | | APTOS2019 | | OCTA500 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Accuracy | F1-score | Accuracy | F1-score | Accuracy | F1-score |
| CE(baseline) | - | - | - | - | 0.713 | 0.700 | 0.580 | 0.569 | 0.320 | 0.195 |
| CE - resample [26] | MICCAI'2022 | ✓ | - | - | 0.742 | 0.725 | 0.536 | 0.509 | 0.333 | 0.214 |
| Focal loss [10] | ICCV'2017 | ✓ | - | - | 0.781 | 0.748 | <u>0.660</u> | 0.632 | <u>0.383</u> | 0.204 |
| LDAM [2] | NeurIPS'2019 | ✓ | - | - | 0.752 | 0.715 | 0.593 | 0.580 | 0.278 | 0.124 |
| Decouple-cRT [20] | arXiv | ✓ | - | ✓ | 0.685 | 0.613 | 0.422 | 0.403 | 0.254 | 0.105 |
| Decouple-$\tau$-norm [20] | arXiv | ✓ | - | ✓ | 0.699 | 0.654 | 0.431 | 0.407 | 0.255 | 0.102 |
| MiSLAS [30] | CVPR'2021 | ✓ | ✓ | ✓ | 0.800 | 0.798 | 0.464 | 0.430 | 0.200 | 0.071 |
| ResLT [5] | TPAMI'2022 | - | - | ✓ | 0.621 | 0.615 | 0.444 | 0.403 | 0.240 | 0.100 |
| CL [10] | CVPR'2020 | - | - | ✓ | 0.350 | 0.304 | 0.464 | 0.441 | 0.293 | 0.226 |
| CL - resample [26] | MICCAI'2022 | ✓ | - | ✓ | 0.600 | 0.602 | 0.460 | 0.452 | 0.267 | 0.108 |
| SCL [22] | NeurIPS'2020 | - | - | ✓ | 0.594 | 0.435 | 0.604 | 0.585 | 0.213 | 0.105 |
| BCL [31] | CVPR'2022 | - | - | ✓ | 0.786 | 0.786 | <u>0.660</u> | <u>0.654</u> | **0.400** | <u>0.214</u> |
| PaCo [6] | ICCV'2021 | - | - | ✓ | 0.715 | 0.702 | 0.596 | 0.517 | 0.240 | 0.149 |
| GPaCo [7] | arXiv | - | - | ✓ | <u>0.800</u> | <u>0.803</u> | 0.596 | 0.518 | 0.260 | 0.186 |
| **BPaCo (Ours)** | - | - | - | ✓ | **0.822** | **0.820** | **0.676** | **0.671** | **0.400** | **0.216** |

module improvement (MI). We select 14 representative methods as our comparison methods. The belonging and characteristic of each method are listed in Table 1.

**Implementation Details.** We perform all experiments on a workstation configured with eight NVIDIA RTX 3090 GPUs. The data augmentation policy is RandAugment [4]. ResNet50 [9] is our backbone network. The batch size is 128 with an initial learning rate of 0.01 and the default optimizer is SGD with a momentum of 0.999 and a weight decay of 1e-4. The input images are resized to $224 \times 224$ and the feature dimension is 128. Hyperparameters $\beta$ is set to 0.25. The training epochs are set as 1,000 for all datasets.

**Comparison with State-of-the-art.** We compare our proposed BPaCo with 14 state-of-the-art methods on all three datasets in Table 1. Apparently from that table, our proposed BPaCo achieves the best performance on all three datasets. Its accuracy outperforms the second best one by 2.2% and 1.6% respectively on ISIC2018 and APTOS2019. For the OCTA500 dataset, BPaCo's F1-score is 0.2% higher than that of BCL even though they have the same accuracy. We also show the influence of the mini-batch's size on different contrastive learning methods to establish the computational efficiency of BPaCo in Fig. 3. As shown in that figure, our BPaCo achieves comparably high performance regardless of the mini-batch's size, which greatly reduces computational cost and avoids performance loss. To be more specific, our BPaCo achieves 78.6% accuracy with a mini-batch of size 64 while other methods need at least a mini-batch of size 128 to attain comparable performance.

**Ablation Study.** Our proposed BPaCo has three fundamental components: cross-entropy logit compensation, across-batch class-averaging and hybrid class-

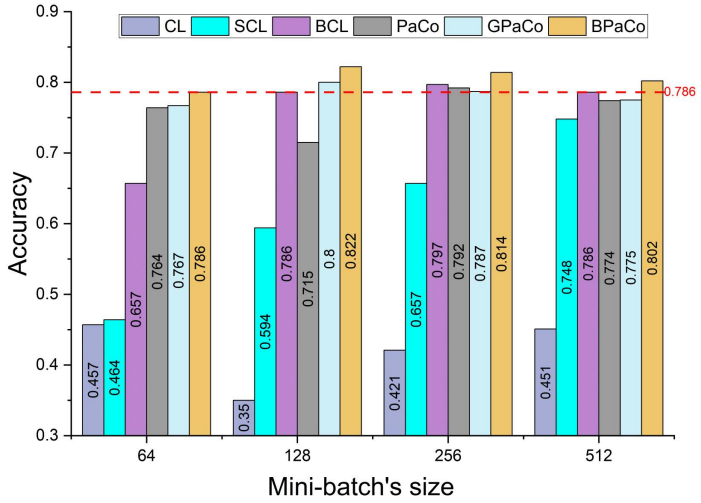



**Fig. 3.** The accuracy of contrastive learning methods with different mini-batch's sizes on ISIC2018. The red dash line represents BPaCo's performance under a size of 64.

complement. To evaluate the effectiveness of each component, we conduct ablation studies. Specifically, seven experiments are conducted by arranging and combining these three components. As shown in Table 2, from the first three rows, we demonstrate that using either across-batch class-averaging or hybrid class-complement alone does not benefit the overall performance. From rows 4, 5 and 6, it is obvious that with the assistance of cross-entropy logit compensation, averaging and complement respectively boost the performance by 13.1% and 13.0%. Finally, our proposed method attains the best performance in terms of both accuracy and F1-score, as shown in the last row with all three elements included.

**Table 2.** Effectiveness of cross-entropy logit compensation (Cross-entropy), across-batch class-averaging (Averaging) and hybrid class-complement (Complement) in our BPaCo framework on ISIC2018. The best ones are bolded while the second best ones are underlined.

| Cross-entropy | Averaging | Complement | Accuracy | F1-score |
|---|---|---|---|---|
| ✓ | | | 0.713 | 0.700 |
| | ✓ | | 0.684 | 0.650 |
| | | ✓ | 0.691 | 0.673 |
| ✓ | ✓ | | 0.773 | 0.740 |
| ✓ | | ✓ | <u>0.781</u> | <u>0.736</u> |
| | ✓ | ✓ | 0.733 | 0.721 |
| ✓ | ✓ | ✓ | **0.822** | **0.820** |

## 4 Conclusion

This paper proposes BPaCo, a novel method addressing the long-tailed classification problem from a contrastive learning perspective. Our BPaCo consists of three essential components: cross-entropy logit compensation, across-batch class-averaging and hybrid class-complement. Extensive experiments on three publicly available datasets exhibit the superiority of our framework in terms of both performance and computational efficiency. BPaCo outperforms existing SOTA long-tailed methods by large margins.

**Disclosure of Interests.** This study was supported by the National Key Research and Development Program of China (2023YFC2415400); the National Natural Science Foundation of China (62071210); the Shenzhen Science and Technology Program (RCYX202106-09103056042); the Shenzhen Science and Technology Innovation Committee (KCX-FZ2020122117340001); the Guangdong Basic and Applied Basic Research (2021A-1515220131).

## References


1. Cai, Z., Lin, L., et al.: Corolla: An Efficient Multi-Modality Fusion Framework with Supervised Contrastive Learning for Glaucoma Grading. In: IEEE International Symposium on Biomedical Imaging, ISBI, pp. 1-4, doi: 10.1109/ISBI52829.2022.9761712 (2022)
2. Cao, K., Wei, C., et al.: Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. Advances in Neural Information Processing Systems, NeurIPS (2019)
3. Codella, N., Rotemberg, V., et al.: Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). arXiv preprint arXiv:1902.03368 (2018)
4. Cubuk, E. D., Zoph, B., et al.: Randaugment: Practical Automated Data Augmentation with a Reduced Search Space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops. pp. 702-703 (2020)
5. Cui, J., Liu, S., et al. Reslt: Residual Learning for Long-tailed Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
6. Cui, J., Zhong, Z., et al.: Parametric Contrastive Learning. In: Proceedings of the IEEE/CVF international Conference on Computer Vision, ICCV, pp. 715-724 (2021)
7. Cui, J., Zhong, Z., et al. Generalized Parametric Contrastive Learning. arXiv preprint arXiv:2209.12400 (2022)
8. Douzas, G., Bacao, F., et al.: Improving Imbalanced Learning Through a Heuristic Oversampling Method Based on K-means and SMOTE. Information Sciences **465**, 1-20 (2018)
9. He, K., Zhang, X., et al.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 770-778 (2016)
10. He, K., Fan, H., et al. Momentum Contrast for Unsupervised Visual Representation Learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, pp. 9729-9738 (2020)
11. Hinton, G., Vinyals, O., Dean, J.: Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531 (2015)
12. Holste, G., Wang, S., et al.: Long-Tailed Classification of Thorax Diseases on Chest X-Ray: A New Benchmark Study. In: Data Augmentation, Labelling, and Imperfections: Second MICCAI Workshop, Proceedings, Cham: Springer Nature Switzerland, pp. 22-32 (2022)
13. Hu, X., Jiang, Y., et al.: Learning to Segment the Tail. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, pp. 14045-14054 (2020)
14. Huang, Y., Huang, W., et al.: Lesion2void: Unsupervised Anomaly Detection in Fundus Images. In: IEEE International Symposium on Biomedical Imaging, ISBI, pp. 1-5, doi: 10.1109/ISBI52829.2022.9761593 (2022)
15. Li, M., Chen, Y., et al., December 23, 2019, "OCTA-500", IEEE Dataport, doi: https://dx.doi.org/10.1109/TMI.2020.2992244
16. Li, T., Wang, L., et al.: Self-supervision to Distillation for Long-tailed Visual Recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, pp. 630-639 (2021)
17. Lin, T., Goyal, P., et al.: Focal Loss for Dense Object Detection. In: Proceedings of the IEEE International Conference on Computer Vision, ICCV, pp. 2980-2988 (2017)

18. Liu, J., Sun, Y., et al.: Deep Representation Learning on Long-tailed Data: A Learnable Embedding Augmentation Perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, pp. 2970-2979 (2020)
19. Liu, Z., Miao, Z., et al.: Large-scale Long-tailed Recognition in An Open World. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, pp. 2537-2546 (2019)
20. Kang, B., Xie, S., et al.: Decoupling Representation and Classifier for Long-tailed Recognition. arXiv preprint arXiv:1910.09217 (2019)
21. Karthik, Maggie, et al.: APTOS 2019 Blindness Detection. https://kaggle.com/competitions/aptos2019-blindness-detection, Kaggle (2019)
22. Khosla, P., Teterwak, P., et al.: Supervised Contrastive Learning. Advances in Neural Information Processing Systems, NeurIPS, pp. 18661-18673 (2020)
23. Koziarski, M.: Radial-based Undersampling for Imbalanced Data Classification. Pattern Recognition **102**: 107262 (2020)
24. Wang, J., Perez, L.: The Effectiveness of Data Augmentation in Image Classification Using Deep Learning. Convolutional Neural Networks Vis. Recognit **11**: 1-8 (2017)
25. Wu, T., Liu, Z., et al.: Adversarial Robustness under Long-tailed Distribution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, pp. 8659-8668 (2021)
26. Yang, Z., Pan, J., et al.: ProCo: Prototype-Aware Contrastive Learning for Long-Tailed Medical Image Classification. In: Medical Image Computing and Computer Assisted Intervention, MICCAI, Proceedings, Part VIII. Cham: Springer Nature Switzerland, pp. 173-182 (2022)
27. Yin, X., Yu, X., et al.: Feature Transfer Learning for Face Recognition with Under-represented Data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVOR, pp. 5704-5713 (2019)
28. Zang, Y., Huang, C., Loy, CC.: Fasa: Feature Augmentation and Sampling Adaptation for Long-tailed Instance Segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, pp. 3457-3466 (2021)
29. Zhang, Y., Kang, B., et al.: Deep Long-tailed Learning: A Survey. arXiv preprint arXiv:2110.04596 (2021)
30. Zhong, Z., Cui, J., et al.: Improving Calibration for Long-tailed Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, pp. 16489-16498 (2021)
31. Zhu, J., Wang, Z., et al.: Balanced Contrastive Learning for Long-tailed Visual Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, pp. 6908-6917 (2022)