Adaptive Unified Contrastive Learning for Imbalanced Classification

Cong Cong[1(✉)], Yixing Yang[1], Sidong Liu[2,3], Maurice Pagnucco[1], Antonio Di Ieva[3], Shlomo Berkovsky[2], and Yang Song[1]

[1] School of Computer Science and Engineering, University of New South Wales, Kensington, Australia
z3414050@ad.unsw.edu.au
[2] Centre for Health Informatics, Macquarie University, Sydney, Australia
[3] Computational NeuroSurgery Lab, Macquarie University, Sydney, Australia

Abstract. Medical image classifiers often suffer from the imbalanced class distribution of datasets. For example, among the 7 classes in the ISIC2018 skin lesion detection dataset, over 67% of the instances belong to melanocytic nevus while only 1% belong to dermatofibroma. Contrastive feature learning has been shown to achieve promising results in enhancing the performance for imbalanced classification tasks. However, the contrastive learning methods are either not end-to-end or require extra memory, which may lead to less compatible and sub-optimal features and classifiers. In this paper, we propose a novel unified feature and classifier learning framework for imbalanced medical image datasets. We equip our model with an adaptive unified contrastive (AduC) loss which progressively adapts model learning between feature learning and classifier learning. Furthermore, we explore the impact of different sampling methods on model training under data sparsity. The experimental results on two long-tailed medical datasets demonstrate that our methods can substantially improve the classification accuracy and F1-score over all classes without using extra memory storage. Our code is available at https://github.com/thomascong121/AdUni.

Keywords: Contrastive learning · Imbalanced classification

1 Introduction

Deep learning has shown great progress in medical image classification in recent years. However, many medical datasets are imbalanced, *i.e.*, majority classes have substantially more samples than other rare classes. As shown in Fig. 1, the majority classes (*e.g.*, "melanocytic nevus" in ISIC2018 [4] and "No Diabetic Retinopathy" in APTOS2019 [1]) dominate the other categories. Classification models that are trained under such imbalanced settings tend to be biased towards the majority classes and perform poorly on the minority classes.

To alleviate this issue in medical image datasets, data re-sampling techniques are widely used [2,12,25], which either up-sample the minority classes or down-sample the majority classes. Conditional generative networks [21] and the mix-up

C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 348–357, 2022.
https://doi.org/10.1007/978-3-031-21014-3_36

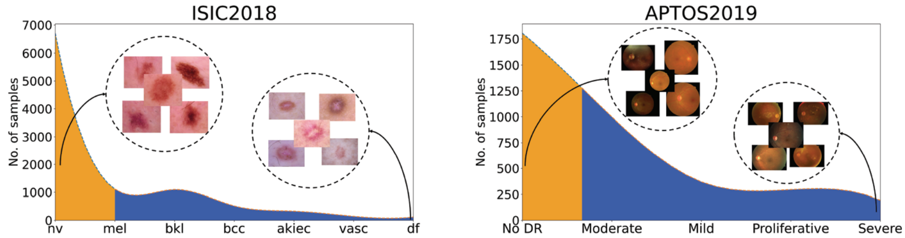


Fig. 1. Class distribution of the dataset used in this paper. Both datasets pose highly imbalanced distributions.

strategy [28] are also commonly used for generating synthetic samples [23,29]. Recently, Marrakchi *et al.* [17] adopted a supervised contrastive learning framework [11] and achieved substantially improved performance for lesion diagnosis and blindness detection under highly imbalanced settings. However, solutions for class imbalance in the medical domain are still under explored.

On the other hand, methods addressing class imbalanced learning for general classification tasks have been well studied. Re-sampling [19,24] and loss-sensitive learning [3,6,20] are the most widely used strategies. However, they normally gain improved performance on minority classes by sacrificing performance on majority classes and they ignore the different properties of feature learning and classifier learning [29]. Decoupled training [9,10,26,27,30] generally conducts feature learning in the first stage and performs classifier learning via cross-entropy (CE) in the second stage. Based on different feature learning methods, they can be divided into two categories: *classification pre-training* and *instance discrimination pre-training*. For classification pre-training [10,30], CE is typically used as the loss function which, however, has limited robustness against class imbalance (see Fig. 2 (left)). To alleviate this issue, these methods either use extra sampling methods (*e.g.*, square-root sampling [16]) or data mixup [28]. Contrastive learning (CL) is commonly used for instance discrimination pre-training, as it helps forming a more balanced feature space. For instance, Yang *et al.* [27] pre-trained the model using self-supervised contrastive learning; Kang *et al.* [9] proposed κ-positive contrastive learning for enhanced balancedness; and, Wang *et al.* [26] forced samples to be closer to their class prototype.

Although the decoupled training methods have demonstrated improved performance, these models are not trained end-to-end and are learned separately with two different targets for the feature learning and classifier learning stages, which makes the features learned from the first stage less compatible with the classifier in the second stage. To learn features and classifier compatibly, Parametric Contrastive Learning (PaCo) [5] provided a unified framework to perform both feature learning and classifier learning concurrently via a modified supervised contrastive loss and showed promising results. However, PaCo uses a memory queue [8] to store negative samples which impairs the class imbalance learning. To further illustrate this, we record the average number of instances of the majority class and minority class in a batch on APTOS2019 for PaCo and our framework. On average, for each batch in our method, the number of majority class samples (70) is approximately 5 times greater than that of the minority class (14). On the other hand, since PaCo uses a memory queue of length 1024 (default value in [5]), the number of majority class samples (795) is over 10 times as many as the minority class samples (68). This indicates that using a memory queue exacerbates the biased learning towards the majority classes. Moreover, PaCo focuses more on classifier learning throughout the training. We argue such a biased learning is sub-optimal for medical datasets in which samples between classes share high semantic similarity (see Fig. 1).

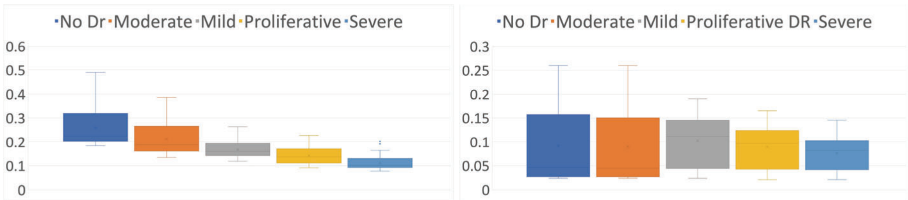


Fig. 2. We compare the loss magnitudes (y-axis) of 5 classes in the APTOS2019 dataset based on the model trained using cross-entropy loss (left) and AdUni (right). Classes are sorted by their cardinality. Clearly, AdUni shows a more balanced distribution of each class.

In this paper, we explore the effectiveness of feature learning in class imbalanced medical image classification. Since the memory queue mechanism in previous feature learning literatures may exacerbate the biased learning for imbalanced classification, we conduct both feature and classifier learning using batch contents only. Furthermore, inspired by the unified feature and classifier learning framework (PaCo), we propose an Adaptive Unified Contrastive (AduC) loss which well balances feature learning and classifier learning throughout training to obtain enhanced performance on datasets with high inter-class similarity. We also design an end-to-end unified learning framework (AdUni) which jointly performs feature learning and classifier learning using batch contents only with AduC. We conduct extensive experiments on the public ISIC2018 and APTOS2019 datasets, demonstrating that our method provides significant performance improvement over existing approaches.

2 Methodology

We design an Adaptive Unified learning framework (AdUni) (see Fig. 3) for imbalanced classification of medical images. AdUni, which is equipped with the proposed Adaptive Unified Contrastive (AduC) loss, provides a compatible learning process for feature and classifier without using any memory queue. As can be seen in Fig. 2, AdUni shows a relatively uniform loss distribution across classes which indicates that AdUni keeps a more balanced learning of each class. Both

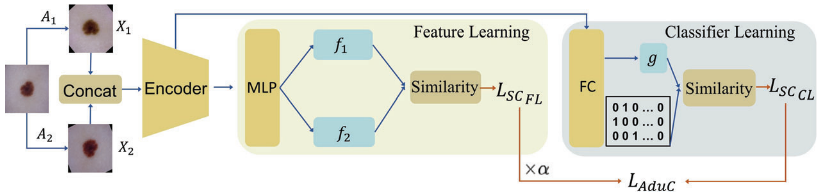


Fig. 3. The proposed AdUni framework which performs both feature learning and classifier learning via contrastive learning with Batch samples only. The model progressively transitions its learning focus from feature learning to classifier learning.

feature learning and classifier learning are conducted via supervised contrastive loss based on batch contents only and are jointly learned via our proposed adaptive unified mechanism to boost performance.

2.1 Feature Learning and Classifier Learning

Feature Learning via Contrastive Learning. We conduct feature learning using contrastive loss. Specifically, given a batch of N input images, for each image $x^i(1 \leq i \leq N)$, we obtain two augmentations x_1^i and x_2^i, respectively, which forms a total of $2N$ augmented training samples. Then, we obtain normalised feature vectors f_1^i and f_2^i using a ResNet-based encoder [7] followed by a two-layer multi-layer perceptron (MLP) (see Fig. 3). Finally, the classifier is trained with the features that are optimised by a supervised contrastive loss $\mathcal{L}_{SC_{FL}}$:

$$\mathcal{L}_{SC_{FL}} = -\sum_{p \in P(i)} \log \frac{\exp(f_i \cdot f_p)/\tau}{\sum\limits_{j \in B(i)} \exp(f_i \cdot f_j)/\tau} \tag{1}$$

For each anchor sample with index i, $P(i)$ collects indices of positive samples (i.e., samples with the same label of sample i), $B(i)$ represents the other $2N-1$ samples and τ is the temperature parameter. Here, $\mathcal{L}_{SC_{FL}}$, which explicitly favours higher similarity between samples with the same class labels, pulls positive pairs closer to each other and forms denser clusters.

Classifier Learning via Contrastive Learning. Previous methods [10,17] use cross-entropy loss (CE) to perform classifier learning separately from feature learning. However, CE is not robust enough under class imbalance. Thus, we replace cross-entropy loss with contrastive loss which is shown to be more robust to class imbalance [9,13]. To achieve this, we use the form of supervised contrastive loss for classifier learning in [5] where we maximise similarities between the class logits $g \in \mathcal{R}^{2N \times C}$ produced by the encoder with the ground-truth label in one-hot format $l \in \mathcal{R}^{2N \times C}$, where C indicates the number of classes. This can be expressed as:

$$\mathcal{L}_{SC_{CL}} = -\sum_{p \in P(i)} \log \frac{\exp(g_i \cdot l_p)/\tau}{\sum\limits_{j \in C} \exp(g_i \cdot l_j)/\tau} \tag{2}$$

where i indicates the index of anchor sample. $\mathcal{L}_{SC_{CL}}$ explicitly encourages the model to maximise the similarities between the logit g_i and the ground-truth label l_p over other classes.

2.2 Adaptive Unified Contrastive (AduC) Learning

To achieve joint feature and classifier learning, we first define a unified loss function L_{AduC}, which merges $\mathcal{L}_{SC_{FL}}$ and $\mathcal{L}_{SC_{CL}}$ by modifying the denominators:

$$\mathcal{L}_{AduC} = -\alpha \sum_{p \in \{P(i)\}} \log \frac{\exp(f_i \cdot f_p)/\tau}{\sum\limits_{j \in C} \exp(g_i \cdot l_j)/\tau + \sum\limits_{k \in B(i)} \exp(f_i \cdot f_k)/\tau}$$
$$- \sum_{p \in \{P(i)\}} \log \frac{\exp(g_i \cdot l_p)/\tau}{\sum\limits_{j \in C} \exp(g_i \cdot l_j)/\tau + \sum\limits_{k \in B(i)} \exp(f_i \cdot f_k)/\tau}$$

The parameter α in Eq. 3 balances the loss function between feature learning and classifier learning. Generally, a smaller α weakens the influence of feature learning and strengthens the influence of classifier learning; and a larger α would make the model focus more on feature learning.

It has been observed that excessively focusing on classifier learning would lead to overfitting, whereas over-emphasising feature learning prevents the model from achieving a better accuracy. To capture the advantages of both feature learning and classifier learning and keep a balance between them, we propose to adaptively unify both training by gradually reducing α to smoothly transition model training from feature learning in the early stage of training to classifier learning in the later stage:

$$\alpha = \omega(t)\alpha_{max} + (1 - \omega(t))\alpha_{min} \tag{3}$$

Here, $\alpha_{min}/\alpha_{max}$ are the minimum/maximum values of α, t is the current epoch number and ω is a function which moderates the decrease of α.

3 Experiments and Results

3.1 Datasets and Implementation

We evaluated our methods on the ISIC2018 skin lesion detection dataset [4] and APTOS2019 blindness detection dataset [1]. ISIC2018 contains 10,015 images of size 450×600 pixels which are categorised into 7 disease states, including 6,705 melanocytic nevus (nv, 67%), 1,113 melanoma (mel, 11%), 1,099 benign keratosis (bkl, 11%), 514 basal cell carcinoma (bcc, 5%), 327 actinic keratosis (akiec, 3%), 142 vascular lesion (vasc, 1%) and 115 dermatofibroma (df, 1%). APTOS2019 contains 3,662 images with varying image sizes, which are categorised into 5 classes based on severity of diabetic retinopathy (DR): 1,805 (49%) no DR, 999 (27%) moderate, 370 (10%) mild, 295 (8%) proliferative, and 193 (5%) severe DR. Both the datasets pose data imbalance issues.

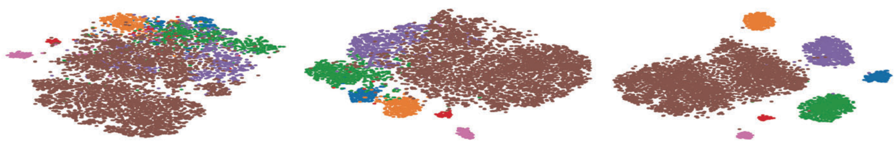
For both datasets, we conduct 5-fold cross validation and apply resize, random affine transformation, horizontal and vertical flip and color jittering for data augmentation. Furthermore, we resize images to 384 × 384 for training a ResNet18 backbone and evaluate its performance by measuring classification accuracy, averaged macro F1-score over all classes. For both datasets, we train the model for 400 epochs with batch size of 128 on 4 NVIDIA RTX3090 GPUs. We develop the framework using PyTorch [22] and set the initial learning rate to 0.1, which is decayed to 0.001 using the cosine schedule [14]. Training takes 5.5 h and 6.5 h on APTOS2019 and ISIC2018, respectively. We set $\tau = 0.1$, $\alpha_{min} = 0.01$, $\alpha_{max} = 1.0$, $\omega(t) = 1 + cos(t/t_{decay}\pi)$ and $t_{decay} = 1000$. For fair comparison, we re-implement all the baseline models with the same setting.

3.2 Results

Table 1. Classification accuracy and macro F1-score on the test set on both datasets using different methods.

Methods	ISIC2018		APTOS2019	
	Acc	Macro F1	Acc	Macro F1
CE [18]	0.841 ± 5.37e–03	0.701±1.21e–03	0.813 ± 6.09e–03	0.603 ± 7.81e–03
BALMS [24]	0.862 ± 2.28e–03	0.756 ± 1.94e–03	0.826 ± 3.46e–03	0.649 ± 3.19e–03
LDAM [3]	0.851 ± 2.28e–03	0.728 ± 1.94e–03	0.815 ± 2.05e–03	0.636 ± 3.46e–03
Decouple [10]	0.861 ± 2.32e–04	0.738 ± 4.01e–03	0.822 ± 2.83e–03	0.661 ± 3.31e–03
CICL [17]	0.866 ± 7.39e–03	0.760 ± 9.71e–03	0.828 ± 2.05e–03	0.676 ± 1.24e–03
PaCo [5]	0.864 ± 2.33e–03	0.768 ± 4.49e–03	0.825 ± 2.00e–03	0.680 ± 3.31e–03
PaCo+AduC	0.869 ± 1.79e–03	0.773 ± 1.25e–03	0.828 ± 2.05e–03	0.687 ± 1.73e–03
AdUni w/o AduC	0.876 ± 2.65e–03	0.778 ± 1.94e–03	0.826 ± 2.44e–03	0.682 ± 1.63e–03
AdUni	**0.878** ± 4.36e–03	**0.805** ± 1.18e–03	**0.839** ± 4.92e–03	**0.695** ± 5.51e–03

Comparison to the State-of-the-Art Methods. In Table 1, we compare the AdUni-based framework to state-of-the-art imbalanced classification studies. The compared methods [3,10,17,24] cover the three categories mentioned in Sect. 1, i.e., re-sampling, loss-sensitive learning and decoupled training, and a recent proposed unified method, i.e., PaCo [5]. Besides, we conduct PaCo+AduC in which we simply change the original fixed α with an adaptively changing α in AduC. It can be seen that our method consistently achieves the highest classification accuracy and macro F1-score.



(a) Decoupled [17] Features. (b) PaCo [5] Features. (c) AdUni Features.

Fig. 4. Feature visualisation of training images on ISIC2018 using t-SNE [15]

Among all the compared methods, the model trained with cross-entropy loss (CE) [18] has the lowest performance which indicates the limitation of CE for imbalanced classification. Decoupled training [10,17,24] produces comparable results; however, their learning schema are not end-to-end and the features learned from the first stage may become less compatible with the second stage learning. The unified training framework, PaCo [5], shows improved performance over the decoupled methods. PaCo+AduC can further enhance the performance of PaCo on both datasets which indicates the usefulness of AduC on datasets with similar semantic information across categories. Furthermore, we observe improved performance over PaCo even by removing the AduC loss from AdUni (AdUni w/o AduC) which illustrates the benefit of removing the memory queue. Moreover, as shown in Fig. 4, compared to the decoupled training method [17] which demonstrates superior performance on both datasets, the features learned by unified frameworks (PaCo [5] and AdUni are more visually separable. The proposed AdUni framework obtains a more distinguishable feature space than PaCo without using a memory queue. Furthermore, compared to the other methods, our proposed framework not only has the best overall classification performance but also achieves improved F1-score on every individual class for both the datasets, as shown in Fig. 5, demonstrating the consistent effectiveness of our method for both majority and minority classes.

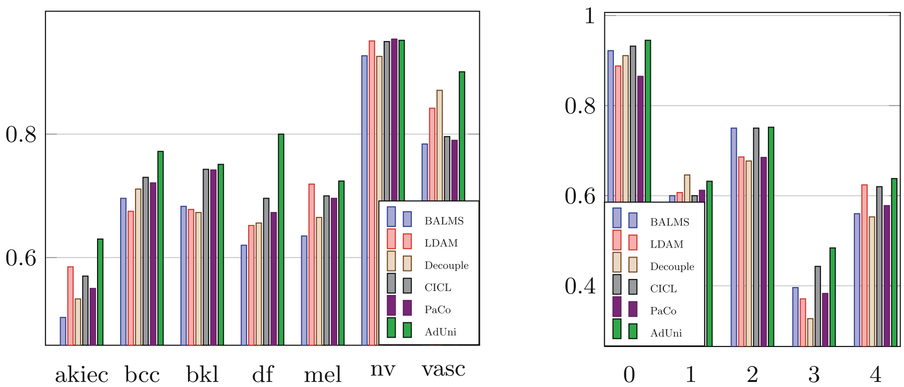


Fig. 5. F1-score for each class on the test set on ISIC2018 (left) and APTOS2019 (right) using different class-imbalanced studies.

Ablation Studies and Discussion. For ablation studies, we study the importance of different α values and show that our designed progressive adaptive α shows improved performance. Finally, we analyse the usefulness of data sampling for further performance improvement.

Table 2 presents the classification accuracy and macro F1-score using different α values and different sampling approaches. Unlike natural images, medical images typically have similar semantic information across categories; thus, we

Table 2. Classification accuracy and macro F1-score on the test set on both datasets with different sampling approaches.

		ISIC2018		APSTOS2019	
		Acc	Macro F1	Acc	Macro F1
$\alpha_{0.01}$	w/o sampling	0.867 ± 4.49e–03	0.760 ± 2.16e–03	0.822 ± 2.44e–03	0.675 ± 1.63e–03
	up-sampling	**0.876** ± 3.55e–03	**0.782** ± 5.79e–03	**0.830** ± 1.34e–03	**0.684** ± 3.74e–03
	SqRt sampling [16]	0.869 ± 3.29e–03	0.769 ± 2.94e–03	0.789 ± 3.87e–03	0.643 ± 1.89e–03
	ProBal sampling [3]	0.873 ± 1.70e–03	0.767 ± 4.77e–03	0.787 ± 6.48e–03	0.641 ± 6.02e–03
$\alpha_{1\rightarrow 0.01}$	w/o sampling	0.878 ± 4.36e–03	0.805 ± 1.18e–03	0.839 ± 4.92e–03	0.695 ± 5.51e–03
	up-sampling	**0.886** ± 2.87e–03	**0.808** ± 2.78e–03	**0.841** ± 2.78e–03	**0.711** ± 3.54e–03
	SqRt sampling [16]	0.877 ± 3.19e–03	0.776 ± 1.72e–03	0.823 ± 2.49e–03	0.676 ± 4.90e–03
	ProBal sampling [3]	0.880 ± 1.72e–03	0.793 ± 3.27e–03	0.810 ± 3.99e–03	0.682 ± 3.26e–03

believe that a more separable feature space can make classifier learning easier. One way to focus the model on learning separable features is to use a large α throughout the training. However, we experimented with $\alpha = \{1, 100\}$ and found that setting large α normally leads to slow convergence. Thus, it is important to balance both feature learning and classifier learning. Here, we choose to adaptively change α using a cosine updating function $\omega(t) = 1 + cos(t/t_{decay}\pi)$ based on epochs t. By choosing suitable t_{decay}, we can keep α large in early training, then gradually drop it to smaller values in the later stage. We observe that using $\alpha_{max} = 1.0$ and progressively reducing it to 0.01 offers a reasonable choice. Table 2 shows that such mechanism improves the accuracy and F1-score by 2∼3%.

Moreover, we compare the results with different sampling approaches where up-sampling means we re-sample the minority classes to ensure every class has the same number of instances in each batch. Square-root (SqRt) sampling [16] calculates the sampling probability of each class using the square root of its sample numbers and progressive-balanced (ProBal) sampling [3] progressively increases the probability of minority samples until each class has equal chances to be selected. The results in Table 2 also show an interesting finding that the simple up-sampling approach works the best with AdUni and provides further improvement. We posit this may be due to the fact that contrastive learning-based approaches are more data-demanding. Although advanced sampling techniques like SqRt and ProBal increase the chance of minority classes to be selected, they cannot provide additional training data.

4 Conclusions

In this paper, we attempt to address the imbalanced data distribution issue for medical image classification. A novel end-to-end training framework is proposed, which jointly performs feature learning and classifier learning based on supervised contrastive loss and an adaptively unified mechanism. Our framework, while not requiring extra memory queue, improves classification performance for all classes in two public datasets ISIC2018 and APTOS2019. Our findings highlight the benefits of enhanced feature learning in improving classification performance for medical image datasets which have an imbalanced data distribution and high similarity between classes.

References

1. APTOS 2019 blindness detection (2019). www.kaggle.com/c/aptos2019-blindness-detection/data
2. Bokhorst, J.M., Pinckaers, H., van Zwam, P., Nagtegaal, I., van der Laak, J., Ciompi, F.: Learning from sparsely annotated data for semantic segmentation in histopathology images. In: MIDL (2018)
3. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. Adv. Neural Inform. Process. Syst. **32** (2019)
4. Codella, N., et al.: Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). arXiv preprint arXiv:1902.03368 (2019)
5. Cui, J., Zhong, Z., Liu, S., Yu, B., Jia, J.: Parametric contrastive learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 715–724 (2021)
6. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: CVPR, pp. 9268–9277 (2019)
7. He, K., Zhang, X., et al.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
8. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR, pp. 9729–9738 (2020)
9. Kang, B., Li, Y., Xie, S., Yuan, Z., Feng, J.: Exploring balanced feature spaces for representation learning. In: International Conference on Learning Representations (2020)
10. Kang, B., et al.: Decoupling representation and classifier for long-tailed recognition. arXiv preprint arXiv:1910.09217 (2019)
11. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Adv. Neural Inform. Process. Syst. **33**, 18661–18673 (2020)
12. Li, Y., Shen, L.: Skin lesion analysis towards melanoma detection using deep learning network. Sensors **18**(2), 556 (2018)
13. Liu, H., HaoChen, J.Z., Gaidon, A., Ma, T.: Self-supervised learning is more robust to dataset imbalance. arXiv preprint arXiv:2110.05025 (2021)
14. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
15. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11) (2008)
16. Mahajan, D., et al.: Exploring the limits of weakly supervised pretraining. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 181–196 (2018)
17. Marrakchi, Y., Makansi, O., Brox, T.: Fighting class imbalance with contrastive learning. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 466–476. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_44
18. Murphy, K.P.: Machine Learning: A Probabilistic Perspective. MIT press (2012)

19. Park, S., Hong, Y., Heo, B., Yun, S., Choi, J.Y.: The majority can help the minority: context-rich minority oversampling for long-tailed classification. arXiv preprint arXiv:2112.00412 (2021)
20. Park, S., Lim, J., Jeon, Y., Choi, J.Y.: Influence-balanced loss for imbalanced visual classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 735–744 (2021)
21. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2337–2346 (2019)
22. Paszke, A., Gross, S., Massa, F., et al.: Pytorch: an imperative style, high-performance deep learning library. In: NeurIPS, pp. 8024–8035. Curran Associates, Inc. (2019)
23. Qasim, A.B., et al.: Red-GAN: attacking class imbalance via conditioned generation. yet another medical imaging perspective. In: Medical Imaging with Deep Learning, pp. 655–668. PMLR (2020)
24. Ren, J., Yu, C., Ma, X., Zhao, H., Yi, S., et al.: Balanced meta-softmax for long-tailed visual recognition. Adv. Neural Inform. Process. Syst. **33**, 4175–4186 (2020)
25. Reza, M.S., Ma, J.: Imbalanced histopathological breast cancer image classification with convolutional neural network. In: ICSP, pp. 619–624 (2018)
26. Wang, P., Han, K., Wei, X.S., Zhang, L., Wang, L.: Contrastive learning based hybrid networks for long-tailed image classification. In: CVPR, pp. 943–952 (2021)
27. Yang, Y., Xu, Z.: Rethinking the value of labels for improving class-imbalanced learning. Adv. Neural Inform. Process. Syst. **33**, 19290–19301 (2020)
28. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: Mixup: beyond Empirical Risk Minimization. In: International Conference on Learning Representations (2018)
29. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: a survey. arXiv preprint arXiv:2110.04596 (2021)
30. Zhong, Z., Cui, J., Liu, S., Jia, J.: Improving calibration for long-tailed recognition. In: CVPR, pp. 16489–16498 (2021)