Pattern Recognition Letters 172 (2023) 51–57

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Local pseudo-attributes for long-tailed recognition

Dong-Jin Kim[a], Tsung-Wei Ke[b], Stella X. Yu[b,c,*]

[a] Hanyang University, South Korea
[b] University of California, Berkeley, USA
[c] University of Michigan, USA

ARTICLE INFO

Article history:
Received 5 January 2023
Revised 14 April 2023
Accepted 31 May 2023
Available online 3 June 2023

Edited by: Jiwen Lu

Keywords:
Long-tailed recognition
Pseudo-attributes
Self-supervised learning

ABSTRACT

Existing long-tailed recognition methods focus on learning global image representation by re-weighing, re-sampling, or global representation learning. However, we observe that solving real-world long-tailed recognition problems requires a fine-grained understanding of local parts within the image in order to avoid confusion among images with similar global configurations. We propose a novel self-supervised learning framework based on local pseudo-attributes (LPA) that are learned via clustering of local features without any human annotations. Such pseudo-attributes are often more balanced compared to image-level class labels. Our method outperforms the state-of-the-art on various long-tailed image classification datasets, such as CIFAR100-LT, iNaturalist, and ImageNet-LT.

1. Introduction

Deep learning has remarkable successes in visual recognition [2–4], by virtue of large-scale annotated datasets such as ImageNet [5] for visual recognition, MS COCO [6] for object detection, or Places [7] for place recognition.

However, these datasets are often curated to be class balanced, unlike long-tailed data distributed real-world datasets (Fig. 1) where a few non-rare classes own most instances, and most classes have a few instances [1,8–16].

Training a classifier with imbalanced data leads to biased predictions towards the few non-rare classes [15], with very low accuracies for most of the rare classes [8,17–19]. Long-tailed recognition has thus remained an active research topic in the past few years [14,20–25].

Several approaches have been proposed for long-tailed recognition, including re-sampling [8,17,26–28], re-weighting [20,29–33], and multi-expert approaches [14,34,35].

Existing methods treat long-tailed recognition as *a machine learning* problem, where images are indivisible given data points. In contrast, we take a *vision perspective* and break the image into divisible components: Not only can they be shared across classes, but their local relationships can also be studied and characterized for better recognition and generalization.

Different from tiny images in CIFAR-LT [20,30], images in large-scale datasets such as ImageNet-LT [1] contain larger scenes with fine-grained classes within *dogs* or *birds* super-classes. A model can be easily confused among rare class images (*Robin* or *Hummingbird* in Fig. 2) with similar *global configurations*. The key to distinguishing such rare-class images from their majority counterparts lies in the *local parts* of the object, such as the color of the body or the length of the beak. We observe in Fig. 1 that a pair of rare- and common- class images also share attributes such as *long tail* for the *bird* or *pointed ears* for the *dog*. These common class instances help us identify local attributes that are also present in rare classes, effectively transferring knowledge from common to rare classes and providing more reliable mid-level representations for further visual discrimination (Fig. 2).

We propose a novel concept, *local pseudo-attributes* (LPA), to capture our intuition. LPA can be obtained without using any human annotations: We run K-means clustering on the pixel-level feature vectors from a model to obtain the pseudo-attribute *labels*. We conjecture that each cluster indicates a specific visual pattern such as *forest-like* or *sky-like*. We then aggregate pseudo-attribute scores of all the instances in the same class to compute class-level pseudo-attribute *labels*.

For example, in Fig. 2, we have local pseudo-attributes *Blue sky, Blunt beak, Long tail* for the rare *Robin* class, *Blue sky, Pointed beak, Short tail* for the rare *Hummingbird* class, and *Green forest, Blunt beak, Long tail* for the common *Bulbul* class. Please note that we only name attributes for the sake of description; our method does not need naming or require knowledge of the language.

Treating each image as a global pattern, the sheer size of *Blue sky* would easily confuse the two rare classes: *Robin* with *Hummingbird*. However, when we consider the image as a collection of these local pseudo-attributes, they are pushed further apart due

[*] Corresponding author.
E-mail addresses: djdkim@hanyang.ac.kr (D.-J. Kim), twke@berkeley.edu (T.-W. Ke), stellayu@umich.edu (S.X. Yu).

https://doi.org/10.1016/j.patrec.2023.05.035
0167-8655/© 2023 Elsevier B.V. All rights reserved.

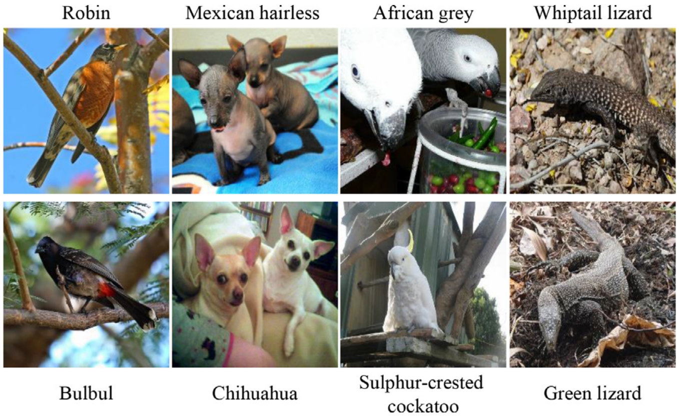

Fig. 1. Paired rare (top row) and common (bottom row) class instances from ImageNet-LT [1]. Although two images in the same column appear different, they share local parts with similar attributes, e.g., from Columns 1 to 4: *long tail of birds, pointed ears of dogs, white body of birds*, and *dotted patterns of lizards*. We propose to leverage such local psuedo-attributes, which can be obtained without human annotations, to learn a model that is more discriminative for rare classes.

to the two other distinctive attributes. The model is guided away from *Blue sky* distractions and focuses instead on parts truly discriminative for these rare classes.

Specifically, given a pseudo-attribute label for each class, we use supervised contrastive learning [36] to force the predictions between instances of similar (different) pseudo-attribute labels to be similar (different). Consequently, *Robin* becomes closer to *Bulbul* instead of *Hummingbird*, which shares big *Blue sky* with *Robin* in their very few instances, thereby increasing visual discrimination among many tail classes.

Please note that, since pseudo-attributes are based on visual features of local parts, they are more likely to be less imbalanced compared to image-level class labels, another aspect that leads to better conditioning during model training.

Our method outperforms the state-of-the-art on extensive long-tailed image classification datasets such as CIFAR100-LT [20,30], iNaturalist [16], and ImageNet-LT [1].

The key contributions of our work can be summarized as follows: **1)** We propose the novel concept of local pseudo-attributes (LPA) which can be learned via clustering of local features without requiring any human annotation. **2)** We leverage pseudo-attribute *labels* to train a classifier via contrastive learning so that the model receives discriminative supervision on detailed local object parts. **3)** We benchmark our method, local pseudo-attributes, on various long-tailed recognition datasets such as CIFAR100-LT, iNaturalist, and ImageNet-LT dataset. Our LPA surpasses the existing state-of-the-art long-tailed recognition methods.

An earlier version of this work was first presented at a non-archival workshop [37]. Here we provide more analysis, additional results, and comprehensive implementation details.

2. Related work

To address long-tail distributed visual recognition, various methods have been proposed [14,38–42] for long years. The previous works for long-tailed recognition can be divided into three different categories: (1) class re-balancing [12,30,43,44]. (2) multi-stage training [20,21], and (3) multi-expert methods [14,34,35,41,45].

Class re-balancing. The most straightforward and classic way to address the long-tailed distribution problem is class re-balancing which modulate the effect of the each samples in the training set on the target model. In particular, class re-balancing can be further divided into of (1) re-sampling the training data [8,21,26,29,46], (2) re-weighting loss function [40,47–51], and (3) increasing the diversity of the rare class via data augmentation [46,52]. It has been shown that many class re-balancing methods improve the image classification performance overall. However, these methods tend to show degraded accuracy on non-rare classes especially on small datasets.

Multi-stage training. The methods with multi-stage training have several different stages during training. For instance, the training procedure by Kang et al. [21] is divided into representation learning stage and classifier learning stage. The multiple stages in the training method of Li et al. [53] includes knowledge distillation. Moreover, other works like [54,55] suggest an additional post-process for shifting the model logits. Such additional processes might increase the overhead of the training whereas our method only adds a simple K-means clustering process which almost doesn't harm the training efficiency.

Multi-expert methods. Finally, multi-expert based methods have been studied including LFME [41], BBN [42], RIDE [14], TADE [35], and ACE [34]. Multi-expert methods have shown favorable performance in long-tailed recognition by virtue of extra models that learn more diverse knowledge. However, such extra models lead to the increase of the complexity not only during training but also during inference. In contrast, we propose a totally new and efficient approach by introducing pseudo-attributes that are

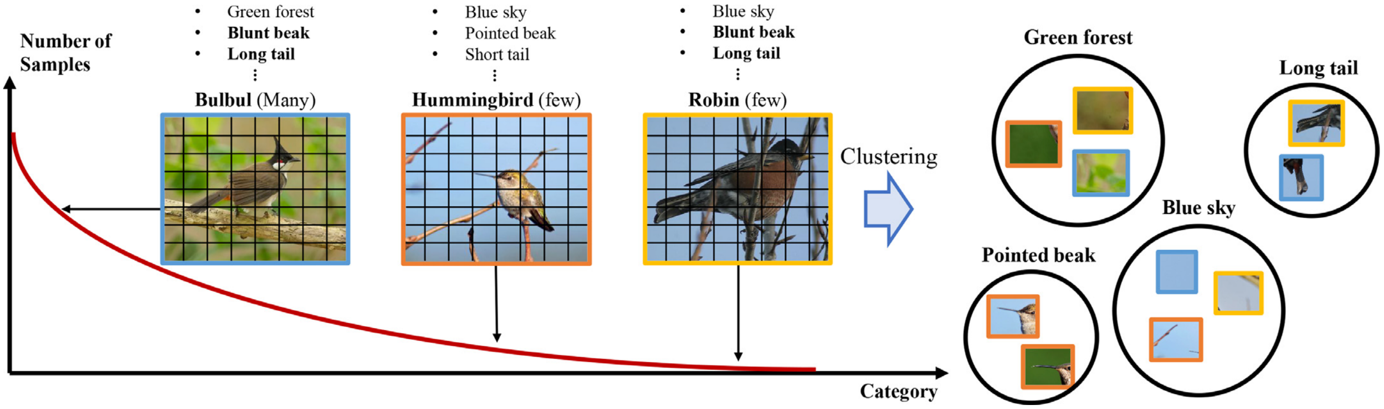

Fig. 2. Illustration of our method with image examples from ImageNet-LT [1]. We run K-means clustering to obtain a set of clusters of local features; we call them local pseudo-attributes. Then, we assign each training image with pseudo-attribute *labels* and utilize these labels to refine the model in a self-supervised learning fashion. Our model learns to distinguish the rare class, *Robin*, from the *Hummingbird* class by giving different supervision signals (*pointed* vs. *blunt beak* and *short* vs. *long beak*) by leveraging knowledge from common *Bulbul* class.

more balanced and help improve the discriminative ability of the model.

3. Local pseudo-attributes

We introduce *local pseudo-attributes* and our contrastive learning with pseudo-attributes for long-tailed recognition.

3.1. Problem definition

Given the input image in the training data $\mathbf{x}$ and its class label y, a neural network consists of encoder $f(\cdot)$ and decoder $g(\cdot)$ and computes the class probability $\hat{\mathbf{y}}$. Let W, H, D, and C denote the width and height of the latent spatial feature, the dimension of the feature vector, and the number of classes. We have:

$$Z = f(\mathbf{x}) \in \mathbb{R}^{W\times H\times D} \tag{1}$$

$$\mathbf{z} = GAP(Z), \qquad z^k = \frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H} Z^{ijk} \tag{2}$$

$$\hat{\mathbf{y}} = g(\mathbf{z}) \in \mathbb{R}^{C}, \tag{3}$$

where $\mathbf{z}$ is the latent feature vector from the global average pooling (GAP). Traditional image classification methods, especially for long-tailed recognition, mostly utilize the global image representation $\mathbf{z}$ by averaging the whole latent representation along the spatial dimension according to the ResNet architecture [3].

In PaCo [36] for long-tailed recognition, the pooled feature vector $\mathbf{z}$ and the class prediction $\hat{\mathbf{y}}$ are used to compute the supervised contrastive loss function $\mathcal{L}_{PaCo}(\hat{\mathbf{y}}, \mathbf{z}, y)$ to train the classifier model.

3.2. Learning with pseudo-attributes

Averaging local representations loses important details that are essential for distinguishing fine-grained details, especially between many long-tailed rare classes. We retain local information by proposing *local pseudo-attribute* (LPA), obtained simply by clustering the $W \times H$ number of D-dimensional local features from Z of the training data which is popularly used in self-supervised [56,57] or unsupervised [58,59] learning works. In particular, we use K-means clustering to obtain the K number of cluster centers.

As illustrated in Fig. 2, each LPA cluster represents a *pseudo-attribute* automatically derived from the training data without any human annotations, as opposed to actual attributes that are labeled by human experts. Please note that pseudo-attributes are less biased than images' semantic class labels in that common and rare classes could share many similar local parts.

We then aggregate pseudo-attribute scores of instances within the same class to compute the class-level pseudo-attribute *label* $\mathbf{a}(y)$. Similar to the recent pseudo-label-based semi-supervised learning works [60], we use the supervised contrastive learning objective [61] to force the predictions between the instances of similar (different) pseudo-attribute labels to be similar (different).

Our final training loss is a combination of the pseudo-attributes based contrastive learning loss and the standard supervised classification loss, weighted by hyperparameter α:

$$\mathcal{L}_{PaCo}(\hat{\mathbf{y}}, \mathbf{z}, y) + \alpha \cdot \mathcal{L}_{SupCon}(\mathbf{z}, \mathbf{a}(y)). \tag{4}$$

4. Experiments

We first describe our experimental setup. We then compare our method with other baseline methods on various datasets. Finally, we conduct additional analysis to validate the effect of our pseudo-attributes.

Table 1
Our model consistently outperforms the state-of-the-art methods (rows) across different imbalance factors (columns) on CIFAR100-LT.

Top-1 accuracy (%) \ imbalance ratio	100	50	10
Softmax	39.1	44.0	55.8
Focal [47]	38.4	44.3	55.8
LDAM-DRW [20]	42.0	46.6	58.7
LWS [21]	42.3	46.0	58.1
BBN [42]	42.6	47.0	59.1
LDAM+DAP [51]	44.1	49.1	58.0
LFME [41]	43.8	-	-
Causal Norm [62]	44.1	50.3	59.6
Balanced Softmax [40]	45.1	49.9	61.6
M2m [52]	43.5	-	57.6
LADE [63]	45.4	50.5	61.7
Hybrid-SC [64]	46.7	51.9	63.1
MisLAS [65]	47.0	52.3	63.2
Logit Adj [54]	43.9	-	-
RIDE [14]	49.1	-	-
DiVE [66]	45.4	51.1	62.0
SSD [53]	46.0	50.5	62.3
ACE [34]	49.6	51.9	-
PaCo [36]	52.0	56.0	64.2
TSC [67]	43.8	47.4	59.0
GCL [68]	48.7	53.6	-
BS+CMO [69]	46.6	51.4	62.3
CMO [69]	50.0	53.0	60.2
Our Baseline	51.3	55.8	63.2
LPA **(Ours)**	**53.3**	**57.1**	**65.1**

Table 2
Our method outperforms the state-of-the-art methods on iNaturalist 2018 with the ResNet-50 backbone.

Method	Top-1 accuracy (%)
Focal	58.0
CB-Focal [30]	61.1
LDAM-DRW [20]	68.0
τ-norm [21]	65.6
LWS [21]	65.9
DAP [51]	67.6
BBN [42]	66.3
FSA [46]	65.9
Balanced Softmax [40]	70.6
LADE [63]	69.3
MiSLAS [65]	70.7
Logit Adj [54]	66.4
RIDE [14]	72.6
IB [72]	65.4
SSD [53]	69.3
TIDE [35]	72.9
ACE [34]	72.9
PaCo [36]	73.2
TSC [67]	69.7
WD&Max [73]	70.2
CMO [69]	72.8
GCL [68]	72.0
LPA **(Ours)**	**73.6**

4.1. Experimental setup

Datasets. We follow the popular evaluation setup described in [1,36] for the long-tailed recognition tasks. In particular, the training data has a long-tailed distribution, whereas the test dataset follows a uniform distribution. Following previous works, we conduct experiments on popular datasets, including the long-tailed version of CIFAR-100 [20,30], iNaturalist 2018 [16], and ImageNet [1] datasets.

CIFAR100-LT. CIFAR-100 consists of 60K images categorized into 100 classes (50K for training and 10K for testing). To implement the CIFAR dataset with long-tailed distribution, we borrow the setting described in [20,30,42]. The imbalance ratio of the dataset γ is defined as $\gamma = \frac{N_{\max}}{N_{\min}}$ where $N_{\max}$ and $N_{\min}$ are the numbers of

Table 3
Our method outperforms the state-of-the-art across all the metrics on ImageNet-LT with ResNeXt-50 backbone.

Top-1 accuracy (%)	Many	Med.	Few	All
Cross Entropy	65.9	37.5	7.7	44.4
OLTR [1]	-	-	-	46.3
cRT [21]	61.8	46.2	27.4	49.6
LWS [21]	60.2	47.2	30.3	49.9
SSD [53]	64.2	50.8	34.5	53.8
TSC [74]	63.5	49.7	30.4	52.4
CMO [69]	66.4	53.9	35.6	56.2
LPA **(Ours)**	**66.7**	**55.4**	**39.0**	**57.5**

Table 4
The prediction statistics of the rare class samples in ImageNet-LT before and after applying our method. Applying our method not only improves the classification accuracy for the rare samples but also alleviates the bias of the wrong predictions toward non-rare classes.

Ratio (%)	Correct	Many	Med.	Few-{GT}
Our Baseline	32.8	30.4	31.5	5.3
LPA **(Ours)**	**39.0**	21.5	30.5	9.0

samples in the training data for the classes with the largest and smallest number of samples. We use imbalance ratios of 100, 50, and 10 in our experiments following [42].

iNaturalist 2018. The iNaturalist 2018 [16] is the large-scale and fine-grained species classification dataset, which is also known to have heavy long-tailed distribution. This dataset consists of 437.5K images categorized into 8142 classes.

ImageNet-LT. ImageNet-LT is a long-tailed version of ImageNet dataset [70] by sampling a subset following the Pareto distribution with power value 6. This dataset has 115.8K images with 1K categories, with samples ranging between 5 and 1,246.

Implementation Details. We use ResNet-50 [3] as our basic backbone architecture. All models are trained via SGD with momentum $\mu = 0.9$. Similar to [36,71], we use a cosine scheduler from 0.02 to 0 for the learning rate and set the batch size to 128 per GPU. The model is trained for 400 epochs using 4 GPUs. The temperature for contrastive learning is set to 0.2. For CIFAR-100-LT, we define the temperature hyper-parameter to be 0.05 and set the remaining to follow [40].

4.2. Results

CIFAR100-LT. Our method is implemented in the same setting as that of PaCo [36]. The baseline “PaCo” in Table 1 refers to the scores reported in the PaCo [36] paper, and “Our Baseline” refers to the model trained with the method that we reproduce described in [36]. Table 1 shows that our LPA model consistently outperforms PaCo [36] and CMO [72], across all different imbalance factors by a large margin. In particular, our method surpasses PaCo by 1.3%, 1.1%, and 0.9% under imbalance factors 100, 50, and 10, respectively, which shows the efficacy of our method. Note that our baseline model shows lower performance than PaCo across different imbalance factors.

iNaturalist and ImageNet. Table 2 lists experimental results on iNaturalist 2018. In this setting, our LPA method consistently surpasses recent state-of-the-art methods, *e.g.*, TSC [67], CMO [72], and GCL [68]. Table 3 shows results on ImageNet-LT. Our method still outperforms the current state-of-the-art long-tailed recognition method, CMO [69], across all the metrics, corroborating the effectiveness of our method. Finally, Table 5 shows results on the Full ImageNet dataset. We compare the classification accuracy between our method and our baseline method (PaCo) with the ResNet-50 backbone. As no other long-tailed recognition works evaluate their methods on Full ImageNet, we only compare them with the PaCo. Our method is also able to improve the classification performance on a balanced dataset as well.

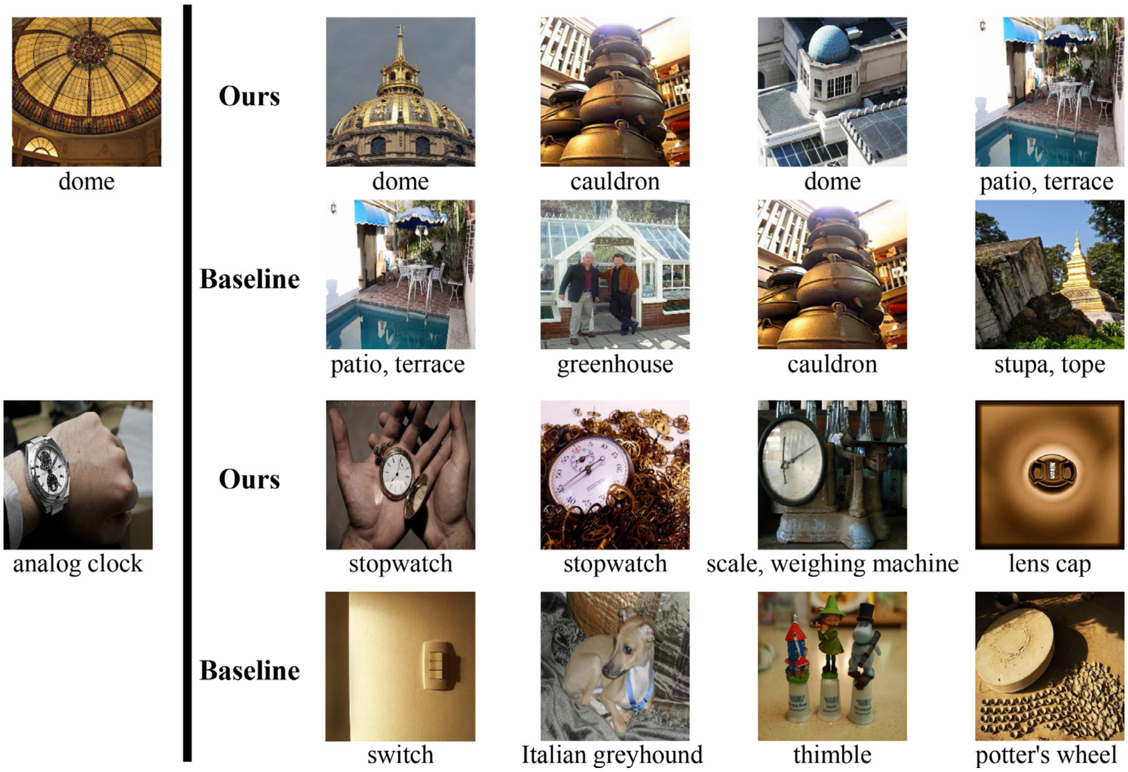

Fig. 3. Image retrievals with the same pseudo-attributes for a rare class sample. Row 1 shows that the model correctly learns the pseudo-attribute for the *dome-like* structure. Row 2 shows the results per the pseudo-attribute for the *watch-like* structure whereas the tendency is less prominent for our baseline.

Table 5
Classification accuracy of our method and our baseline method (PaCo) on the Full ImageNet dataset with the ResNet-50 backbone. Our method is also able to improve the classification performance on a balanced dataset as well.

Method	Top-1 accuracy (%)
PaCo [36]	78.7
PaCo [36] + LPA **(Ours)**	**79.6**

Learned Local Pseudo-attributes. In order to understand what these local pseudo-attributes look like, we retrieve images with the same pseudo-attribute labels given the rare class sample as a query and compare it with that of our baseline model. Figure 3 shows that the learned pseudo-attributes help the model correctly predict the class of the test images, whereas the tendency is less prominent for our baseline. For example, the first and second rows show the retrieval result of the pseudo-attributes for the *dome-like* and *watch-like* structures, respectively. Note that these pseudo-attributes are *self-supervisedly* learned and do not require human labor to label them.

Prediction Statistics. We analyze how our LPA improves recognition performance on many tail classes. There are 385, 479, and 136 in many, medium, and few-shot classes, respectively, in ImageNet-LT. For each instance in the few-shot classes, we look at the top prediction: It could be the correct few-shot class or mistaken as one in the many, medium, or any other few-shot classes. We collect these statistics and calculate the distribution of these prediction types. Note that the sum of these ratios is always 100%. We compare the prediction statistics between our baseline and final models (LPA). Table 4 shows that the model is heavily biased towards majority classes and confuses rare class instances mostly as many- or medium-shot classes. Our method not only improves the few-shot class accuracy from 32.8 to 39.0, but the distribution of mistakes is also flattened more among many-, medium-, and other few-shot classes, from 5.8 : 5.9 : 1 to 2.4 : 3.4 : 1 respectively, reducing the bias of confusion by roughly a factor of 2. That is, our LPA improves long-tailed recognition by reducing both tail-head confusion and confusion bias. We also show the distribution of the incorrect predictions for given all the classes in the confusion matrix form in Fig. 4. For the baseline, it is notable that the predictions are biased to many- and medium-shot classes regardless of the ground truth class. In addition, for any ground truth classes, our method flattens the distribution of the mistakes of the model.

Hyper-parameters. We show the hyper-parameter analysis on the number of clusters during the K-means clustering in Fig. 5. In particular, we show the classification accuracy on the CIFAR100-LT dataset with different imbalance ratios ($\gamma = 10, 50, 100$) with a different number of clusters K. We also show the performance of our baseline model as a reference (dotted line). Our method outperforms the baseline models overall regardless of the number of clusters. Among various design choices, the model with K=200 generally shows the best performance.

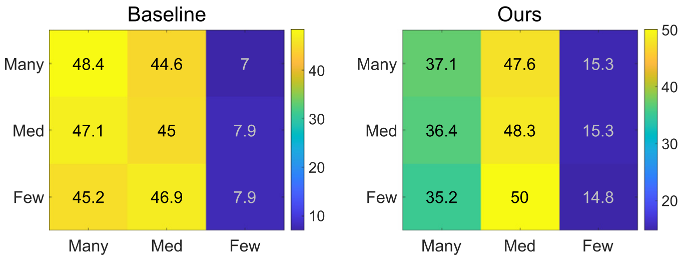

Fig. 4. Distribution of the incorrect predictions in ImageNet-LT before and after applying our method. Y-axis indicates whether the input samples are from many-, med-, or few-shot classes. The X-axis indicates what class the model predicted to be. For the baseline, the predictions are biased to many- and medium-shot classes regardless of the ground truth class. Our method flattens the distribution of the mistakes.

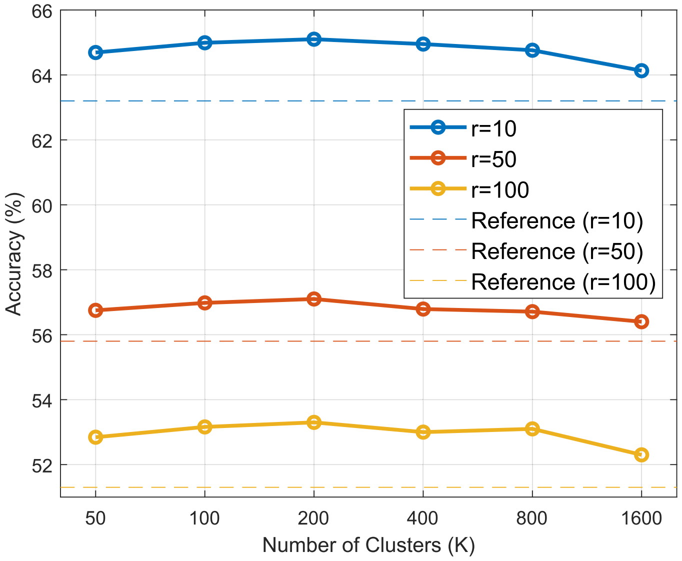

Fig. 5. Classification accuracy on CIFAR100-LT dataset with different imbalance ratios ($\gamma = 10, 50, 100$) with a different number of clusters K. We also show the performance of our baseline model as a reference (dotted line). Our method outperforms the baseline models overall regardless of the number of clusters. Among various design choices, the model with K=200 generally shows the best performance.

5. Conclusion

We introduce the novel concept of *local pseudo-attributes* (LPA), derived from data automatically without requiring human annotations, and incorporate it into long-tailed recognition with an additional self-supervised learning loss.

Our experimental results on CIFAR100-LT, iNaturalist, and ImageNet-LT demonstrate that our method (LPA) consistently outperforms various state-of-the-art methods.

Our ideas can also be extended to other tasks that suffer from similar data bias problems, such as visual question answering [75], semi-supervised learning [24,76] or active learning [77–79].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

The work of Dong-Jin Kim was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government(MSIT) (No.2020-0-01373, Artificial Intelligence Graduate School Program(Hanyang University)) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00245661).

References

[1] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, S.X. Yu, Large-scale long-tailed recognition in an open world, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[2] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: IEEE International Conference on Computer Vision (ICCV), 2017.

[3] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[4] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems (NIPS), Vol. 28, 2015, pp. 91–99.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.

[6] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: European Conference on Computer Vision (ECCV), 2014.

[7] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: a 10 million image database for scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 40 (6) (2017) 1452–1464.

[8] M. Buda, A. Maki, M.A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks, Neural Netw. 106 (2018) 249–259.

[9] A. Gupta, P. Dollar, R. Girshick, LVIS: a dataset for large vocabulary instance segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5356–5364.

[10] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D.A. Shamma, et al., Visual genome: connecting language and vision using crowdsourced dense image annotations, Int. J. Comput. Vis. 123 (1) (2017) 32–73.

[11] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: IEEE International Conference on Computer Vision (ICCV), 2015.

[12] Y.-X. Wang, D. Ramanan, M. Hebert, Learning to model the tail, Advances in Neural Information Processing Systems (NIPS), 2017.

[13] G. Van Horn, P. Perona, The devil is in the tails: fine-grained classification in the wild, arXiv preprint arXiv:1709.01450(2017).

[14] X. Wang, L. Lian, Z. Miao, Z. Liu, S.X. Yu, Long-tailed recognition by routing diverse distribution-aware experts, in: International Conference on Learning Representations (ICLR), 2021.

[15] Y. Zhang, B. Kang, B. Hooi, S. Yan, J. Feng, Deep long-tailed learning: a survey, arXiv preprint arXiv:2110.04596(2021).

[16] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, S. Belongie, The inaturalist species classification and detection dataset, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[17] S. Ando, C.Y. Huang, Deep over-sampling framework for classifying imbalanced data, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2017, pp. 770–785.

[18] R. Collobert, J. Weston, A unified architecture for natural language processing: deep neural networks with multitask learning, in: International Conference on Machine Learning (ICML), 2008.

[19] T.-Y. Wu, P. Morgado, P. Wang, C.-H. Ho, N. Vasconcelos, Solving long-tailed recognition with deep realistic taxonomic classifier, in: European Conference on Computer Vision (ECCV), 2020.

[20] K. Cao, C. Wei, A. Gaidon, N. Arechiga, T. Ma, Learning imbalanced datasets with label-distribution-aware margin loss, Advances in Neural Information Processing Systems (NIPS), 2019.

[21] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, Y. Kalantidis, Decoupling representation and classifier for long-tailed recognition, in: International Conference on Learning Representations (ICLR), 2020.

[22] D.-J. Kim, X. Sun, J. Choi, S. Lin, I.S. Kweon, Detecting human-object interactions with action co-occurrence priors, in: European Conference on Computer Vision (ECCV), 2020.

[23] D.-J. Kim, X. Sun, J. Choi, S. Lin, I.S. Kweon, ACP++: action co-occurrence priors for human-object interaction detection, IEEE Trans. Image Process. (TIP) 30 (2021) 9150–9163.

[24] Y. Oh, D.-J. Kim, I.S. Kweon, DASO: distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[25] Y. Yang, Z. Xu, Rethinking the value of labels for improving class-imbalanced learning, Advances in Neural Information Processing Systems (NIPS), 2020.

[26] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.

[27] L. Shen, Z. Lin, Q. Huang, Relay backpropagation for effective learning of deep convolutional neural networks, in: European Conference on Computer Vision (ECCV), 2016.

[28] J. Van Hulse, T.M. Khoshgoftaar, A. Napolitano, Experimental perspectives on learning from imbalanced data, in: International Conference on Machine Learning (ICML), 2007.

[29] J. Byrd, Z. Lipton, What is the effect of importance weighting in deep learning? in: International Conference on Machine Learning (ICML), 2019.

[30] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[31] Q. Dong, S. Gong, X. Zhu, Imbalanced deep learning by minority class incremental rectification, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 41 (6) (2018) 1367–1381.

[32] S. Khan, M. Hayat, S.W. Zamir, J. Shen, L. Shao, Striking the right balance with uncertainty, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[33] S.H. Khan, M. Hayat, M. Bennamoun, F.A. Sohel, R. Togneri, Cost-sensitive learning of deep feature representations from imbalanced data, IEEE Trans. Neural Netw. Learn. Syst. 29 (8) (2017) 3573–3587.

[34] J. Cai, Y. Wang, J.-N. Hwang, ACE: ally complementary experts for solving long–tailed recognition in one-shot, in: IEEE International Conference on Computer Vision (ICCV), 2021.

[35] Y. Zhang, B. Hooi, L. Hong, J. Feng, Test-agnostic long-tailed recognition by test-time aggregating diverse experts with self-supervision, in: IEEE International Conference on Computer Vision (ICCV), 2021.

[36] J. Cui, Z. Zhong, S. Liu, B. Yu, J. Jia, Parametric contrastive learning, in: IEEE International Conference on Computer Vision (ICCV), 2021.

[37] D.-J. Kim, T.-W. Ke, X.Y. Stella, Local pseudo-attributes for long-tailed recognition, NeurIPS Workshop: Self-Supervised Learning - Theory and Practice, 2022.

[38] D. Cao, X. Zhu, X. Huang, J. Guo, Z. Lei, Domain balancing: Face recognition on long-tailed domains, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[39] J. Liu, J. Zhang, W. Li, C. Zhang, Y. Sun, et al., Memory-based jitter: Improving visual recognition on long-tailed data with diversity in memory, in: AAAI Conference on Artificial Intelligence (AAAI), 2020.

[40] J. Ren, C. Yu, X. Ma, H. Zhao, S. Yi, et al., Balanced meta-softmax for long-tailed visual recognition, Advances in Neural Information Processing Systems (NIPS), 2020.

[41] L. Xiang, G. Ding, J. Han, Learning from multiple experts: self-paced knowledge distillation for long-tailed classification, in: European Conference on Computer Vision (ECCV), 2020.

[42] B. Zhou, Q. Cui, X.-S. Wei, Z.-M. Chen, BBN: bilateral-branch network with cumulative learning for long-tailed visual recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[43] C. Huang, Y. Li, C.C. Loy, X. Tang, Learning deep representation for imbalanced classification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5375–5384.

[44] M. Ren, W. Zeng, B. Yang, R. Urtasun, Learning to reweight examples for robust deep learning, in: International Conference on Machine Learning (ICML), PMLR, 2018, pp. 4334–4343.

[45] Y. Li, T. Wang, B. Kang, S. Tang, C. Wang, J. Li, J. Feng, Overcoming classifier imbalance for long-tail object detection with balanced group softmax, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[46] P. Chu, X. Bian, S. Liu, H. Ling, Feature space augmentation for long-tailed data, in: European Conference on Computer Vision (ECCV), 2020.

[47] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: IEEE International Conference on Computer Vision (ICCV), 2017.

[48] J. Tan, C. Wang, B. Li, Q. Li, W. Ouyang, C. Yin, J. Yan, Equalization loss for long–tailed object recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[49] J. Wang, W. Zhang, Y. Zang, Y. Cao, J. Pang, T. Gong, K. Chen, Z. Liu, C.C. Loy, D. Lin, Seesaw loss for long-tailed instance segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[50] J. Shu, Q. Xie, L. Yi, Q. Zhao, S. Zhou, Z. Xu, D. Meng, Meta-weight-net: learning an explicit mapping for sample weighting, Advances in Neural Information Processing Systems (NIPS), 2019.

[51] M.A. Jamal, M. Brown, M.-H. Yang, L. Wang, B. Gong, Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[52] J. Kim, J. Jeong, J. Shin, M2m: imbalanced classification via major-to-minor translation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[53] T. Li, L. Wang, G. Wu, Self supervision to distillation for long-tailed visual recognition, in: IEEE International Conference on Computer Vision (ICCV), 2021.

[54] A.K. Menon, S. Jayasumana, A.S. Rawat, H. Jain, A. Veit, S. Kumar, Long-tail learning via logit adjustment, in: International Conference on Learning Representations (ICLR), 2021.

[55] S. Zhang, Z. Li, S. Yan, X. He, J. Sun, Distribution alignment: a unified framework for long-tail visual recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[56] B.X. Nguyen, B.D. Nguyen, G. Carneiro, E. Tjiputra, Q.D. Tran, T.-T. Do, Deep metric learning meets deep clustering: an novel unsupervised approach for feature embedding, in: British Machine Vision Conference (BMVC), 2020.

[57] Y. Pan, T. Yao, Y. Li, C.-W. Ngo, T. Mei, Exploring category-agnostic clusters for open-set domain adaptation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[58] Q. Cai, Y. Wang, Y. Pan, T. Yao, T. Mei, Joint contrastive learning with infinite possibilities, Advances in Neural Information Processing Systems (NIPS), 2020.

[59] C. Wei, H. Wang, W. Shen, A. Yuille, Co2: consistent contrast for unsupervised visual representation learning, in: International Conference on Learning Representations (ICLR), 2021.

[60] K. Sohn, D. Berthelot, C. Li, Z. Zhang, N. Carlini, E.D. Cubuk, A. Kurakin, H. Zhang, C. Raffel, FixMatch: simplifying semi-supervised learning with consistency and confidence, Advances in Neural Information Processing Systems (NIPS), 2020.
[61] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, Advances in Neural Information Processing Systems (NIPS), 2020.
[62] K. Tang, J. Huang, H. Zhang, Long-tailed classification by keeping the good and removing the bad momentum causal effect, Advances in Neural Information Processing Systems (NIPS), 2020.
[63] Y. Hong, S. Han, K. Choi, S. Seo, B. Kim, B. Chang, Disentangling label distribution for long-tailed visual recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6626–6636.
[64] P. Wang, K. Han, X.-S. Wei, L. Zhang, L. Wang, Contrastive learning based hybrid networks for long-tailed image classification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.
[65] Z. Zhong, J. Cui, S. Liu, J. Jia, Improving calibration for long-tailed recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.
[66] Y.-Y. He, J. Wu, X.-S. Wei, Distilling virtual examples for long-tailed recognition, in: IEEE International Conference on Computer Vision (ICCV), 2021.
[67] T. Li, P. Cao, Y. Yuan, L. Fan, Y. Yang, R.S. Feris, P. Indyk, D. Katabi, Targeted supervised contrastive learning for long-tailed recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
[68] M. Li, Y.-m. Cheung, Y. Lu, Long-tailed visual recognition via Gaussian clouded logit adjustment, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
[69] S. Park, Y. Hong, B. Heo, S. Yun, J.Y. Choi, The majority can help the minority: Context-rich minority oversampling for long-tailed classification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
[70] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.
[71] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9729–9738.
[72] S. Park, J. Lim, Y. Jeon, J.Y. Choi, Influence-balanced loss for imbalanced visual classification, in: IEEE International Conference on Computer Vision (ICCV), 2021, pp. 735–744.
[73] S. Alshammari, Y.-X. Wang, D. Ramanan, S. Kong, Long-tailed recognition via weight balancing, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
[74] B. Li, Z. Han, H. Li, H. Fu, C. Zhang, Trustworthy long-tailed classification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
[75] J.W. Cho, D.-j. Kim, H. Ryu, I.S. Kweon, Generative bias for robust visual question answering, 2023.
[76] D.-J. Kim, J. Choi, T.-H. Oh, I.S. Kweon, Image captioning with very scarce supervised data: adversarial semi-supervised learning approach, in: Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019.
[77] J.W. Cho, D.-J. Kim, Y. Jung, I.S. Kweon, MCDAL: maximum classifier discrepancy for active learning, IEEE Trans. Neural Netw. Learn. Syst. (2022).
[78] D.-J. Kim, J.W. Cho, J. Choi, Y. Jung, I.S. Kweon, Single-modal entropy based active learning for visual question answering, in: British Machine Vision Conference (BMVC), 2021.
[79] I. Shin, D.-J. Kim, J.W. Cho, S. Woo, K. Park, I.S. Kweon, Labor: labeling only if required for domain adaptive semantic segmentation, in: IEEE International Conference on Computer Vision (ICCV), 2021.