Local Pseudo-Attributes for Long-Tailed Recognition

Dong-Jin Kim
Hanyang University
Seoul, South Korea
djdkim@hanyang.ac.kr

Tsung-Wei Ke
UC Berkeley / ICSI
Berkeley, CA
twke@berkeley.edu

Stella X. Yu
UC Berkeley / ICSI
Berkeley, CA
stellayu@berkeley.edu

Abstract

We observe that solving the long-tailed distribution problems in real-world image datasets requires a fine-grained understanding of the local parts in an image. We propose a novel self-supervised learning framework with a new concept we call local pseudo-attribute (LPA) that is learned via clustering with local features without any extra human annotation. Note that the pseudo-attributes can be more balanced compared to the image-level class labels. Our final method using local pseudo-attributes achieves state-of-the-art performance through the experiments on various image classification setups with long-tailed distribution, such as CIFAR100-LT, iNaturalist, and ImageNet-LT datasets compared to the current long-tailed recognition methods.

1 Introduction

Real-world datasets mostly exhibit long-tailed data distributions [29]. This motivates the well-studied problem of long-tailed recognition, which has attracted increasing attention especially using deep neural networks [4, 10, 16, 17, 24, 31, 32, 33]. Various methods have been proposed to alleviate the long-tailed distribution problem starting from simple re-sampling [1, 2, 5, 27, 30] and re-weighting [3, 4, 8, 9, 11, 12]. Most of the existing methods can be seen as *machine learning based* approaches, which considers images as data points. In contrast, we propose a *vision perspective* to alleviate long-tailed distribution problem and focus on the local relations within the image.

Unlike small-scale datasets like CIFAR-LT [4, 8], realistic image datasets requires *fine-grained* scene understanding capability, which makes the long-tailed recognition more challenging. Without fine-grained understanding capability, a model can be easily confused among the rare class images ("Robin" or "Hummingbird" in Fig. 2) with similar *global structure*. In such realistic and fine-grained image datasets, the key to distinguish the rare class images from the distractor images is 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 the pairs of rare and non-rare class images that share the same attributes (a long tail of the bird or pointed ears of the dog). Our goal is to leverage such knowledge from non-rare classes to better discriminate the rare class samples.

To this end, we propose a novel concept of *local pseudo-attributes* (LPA) that are obtained in an unsupervised manner. In particular, we run the K-means clustering on the pixel-level feature vectors from a model to obtain the pseudo-attribute prototypes. We conjecture that each prototype contains the specific visual pattern such as "forest-like" or "sky-like." Then, we aggregate the pseudo-attribute scores of the samples with the same class to compute the class-level attribute ground truth with the visual commonality. Fig. 2 shows the examples of the pseudo-attribute ground truth such as "Blue sky, Blunt beak, Long tail" for "Robin" class that contains the visual commonality within the class. The rare class "Robin" is assigned with the more discriminative supervision signals by leveraging the knowledge of the abundant samples for non-rare classes. We name the label as the *pseudo*-attributes because these labels contains the attribute information such as the shape or texture without relying on any extra human labels. Also note that, as the pseudo-attributes are based on the low-level local parts,

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Figure 1: The example of the pairs of rare (top row) and non-rare (bottom row) class samples from the ImageNet-LT dataset [22]. Although the overall appearance is different between the rare and non-rare classes, there are local parts with the same attributes that are shared between the rare and non-rare classes (long tail of the bird, pointed ears of the dogs, white color of the birds, and dotted patterns on the lizards). We propose to leverage such local attributes to better train rare class images.

the pseudo-attributes are less biased compared to the image-level class label. Given a pseudo-attribute labels for each class, we use supervised contrastive learning [7] to force the predictions between the samples with similar pseudo-attribute labels to be similar. Thereby, in Fig. 2, although the global appearance is similar, the image of "Robin" is pushed towards the image of "Bulbul" while the image of "Hummingbird" is not. Such different supervision helps the model to discriminate the two similar rare class images.

Our method achieves the state-of-the-art image classification accuracy on various long-tailed image classification benchmarks such as CIFAR100-LT [4, 8], iNaturalist [29], and ImageNet-LT [22] datasets. Our contributions can be summarized as follows: (1) For a long-tailed recognition problem, we propose a novel concept we call local pseudo-attribute (LPA) that is learned via clustering with local features which does not require any extra human annotation. (2) We leverage the pseudo-attribute labels to train our model via contrastive learning so that the model receive discriminative supervision based on the fine-grained detailed object parts. (3) We evaluate our new method on various data setups with long-tailed distribution including CIFAR100-LT, iNaturalist, and our ImageNet-LT dataset. Our method surpasses the current state-of-the-art methods long-tailed recognition methods.

2 Local Pseudo-Attributes

Given an input image $\mathbf{x}$ and the class label y, a covolutional network consist of an encoder $f(\cdot)$ and the decoder $g(\cdot)$ takes the image $\mathbf{x}$ as input to compute the class probability sequentially, *i.e.*, $Z = f(\mathbf{x}) \in \mathbb{R}^{W \times H \times D}$ and $\hat{\mathbf{y}} = g(GAP(Z)) \in \mathbb{R}^{C}$, where W, H, D, C are the width and height of the latent spatial feature, dimension of the feature vector, and the number of classes. Here, $\mathbf{z} = GAP(Z)$ is the latent feature vector from the global average pooling. Then, the pooled feature vector $\mathbf{z}$ and the class prediction $\hat{\mathbf{y}}$ are used to compute the supervised contrastive loss function $\mathcal{L}_{PaCo}(\hat{\mathbf{y}}, \mathbf{z}, y)$ from [7] to train the model.

In order to improve the model's fine-grained discrimination among rare classes, we introduce a novel concept we call *local pseudo-attribute* (LPA) which are be obtained by clustering the local features. In particular, we cluster the $W \times H$ number of D-dimensional local features from Z along the training data to obtain the cluster prototypes that we name the pseudo-attributes. We call this a *pseudo-attribute* because this can be the an alternative of the attribute labels that does not require human labeling effort. Also note that the pseudo-attributes are less biased compared to the original class label in that the non-rare class and the rare class share similar local parts.

Then, we aggregate the pseudo-attribute scores of the samples with the same class to compute the class-level attribute ground truth $\mathbf{a}(y)$. Given an pseudo-attribute labels for each class, we use the supervised contrastive learning objective [13] to force the predictions between the samples with similar pseudo-attribute labels to be similar. Then, we use the combination of the loss functions to

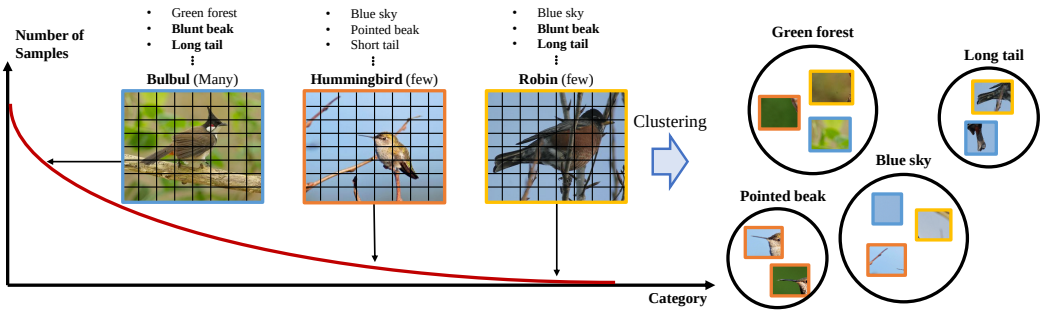

Figure 2: The illustration of the proposed framework with the examples in the ImageNet-LT dataset [22]. We run K-means clustering to obtain a set of cluster of local features, we call them local pseudo-attributes. Then, we assign each training image with the pseudo-attribute labels and use the pseudo-labels to refine the model in a self-supervised way. The 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 the knowledge from the non-rare class samples (Bulbul).

Methods	100	50	10
LDAM-DRW [4]	42.0	46.6	58.7
LWS [10]	42.3	46.0	58.1
BBN [35]	42.6	47.0	59.1
MisLAS [34]	47.0	52.3	63.2
Logit Adj [23]	43.9	-	-
PaCo [7]	52.0	56.0	64.2
TSC [20]	43.8	47.4	59.0
GCL [19]	48.7	53.6	-
CMO [25]	50.0	53.0	60.2
Our Baseline	51.3	55.8	63.2
LPA **(Ours)**	**53.3**	**57.1**	**65.1**

Table 1: Top-1 accuracy on CIFAR100-LT dataset. Our final model consistently outperforms the state-of-the-art method across different imbalance factors.

Methods	Top-1 accuracy
LDAM-DRW [4]	68.0
LWS [10]	65.9
BBN [35]	66.3
MiSLAS [34]	70.7
Logit Adj [23]	66.4
RIDE [31]	72.6
PaCo [7]	73.2
TSC [20]	69.7
CMO [25]	72.8
GCL [19]	72.0
LPA **(Ours)**	**73.6**

Table 2: Top-1 accuracy over all classes on iNaturalist 2018 with ResNet-50. Our method achieves the state-of-the-art performance on iNaturalist dataset as well.

train the model:

$$\mathcal{L}_{PaCo}(\hat{\mathbf{y}}, \mathbf{z}, y) + \alpha \mathcal{L}_{SupCon}(\mathbf{z}, \mathbf{a}(y)), \tag{1}$$

where α is the hyper-parameter to weight the losses.

3 Experiments

We follow the common evaluation protocol [7, 22] in long-tailed recognition. We conduct experiments on long-tailed version of CIFAR-100 [4, 8], iNaturalist 2018 [29], and ImageNet [22] datasets.

CIFAR100-LT. The experimental results on CIFAR100-LT are listed in Table 1. As shown in Table 1, our final model (LPA) consistently outperforms the state-of-the-art methods, PaCo [7] and CMO [26], 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 factor 100, 50 and 10 respectively, which testify the effectiveness of our method. Note that our baseline model shows lower performance than PaCo across different imbalance factors.

iNaturalist and ImageNet-LT. Table 2 lists experimental results on iNaturalist 2018. Under fair training setting, our local pseudo-attribute (LPA) method consistently surpasses recent state-of-the-art methods of TSC, CMO, and GCL as well as the CIFAR100-LT dataset. Also, Table 3 shows the results on the ImageNet-LT. Our method still outperforms the state-of-the-art long-tailed recognition method, CMO [25], in all the metrics, which signifies the general effectiveness of our method.

Methods	Many	Med.	Few	All
Cross Entropy	65.9	37.5	7.7	44.4
OLTR [22]	-	-	-	46.3
cRT [10]	61.8	46.2	27.4	49.6
LWS [10]	60.2	47.2	30.3	49.9
SSD [21]	64.2	50.8	34.5	53.8
TSC [18]	63.5	49.7	30.4	52.4
CMO [25]	66.4	53.9	35.6	56.2
LPA **(Ours)**	**66.7**	**55.4**	**39.0**	**57.5**

Table 3: Top-1 accuracy on ImageNet-LT dataset compared to the state-of-the-art methods with ResNeXt-50 as backbone. Our method outperforms the state-of-the-art methods in all the metrics.

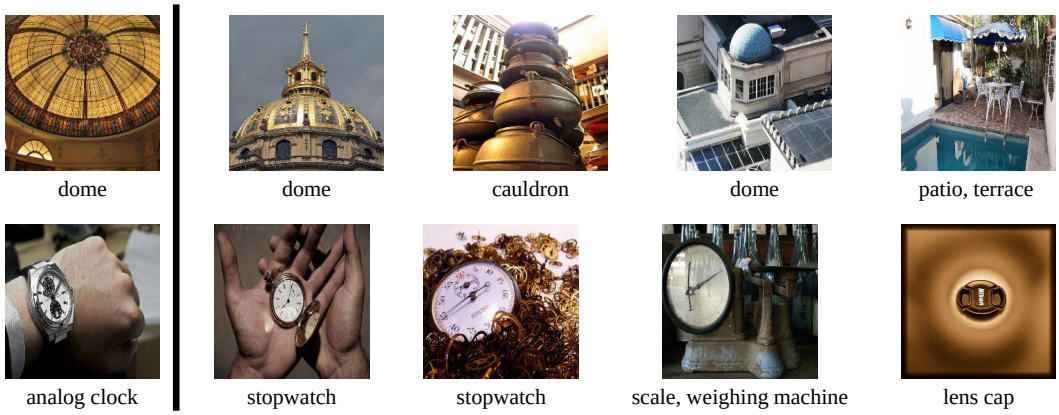

Figure 3: Image retrieval results with the same pseudo-attributes given a rare class sample. The first row shows that the model correctly learns the pseudo-attribute for "dome-like" structure. The second row shows the result of the pseudo-attribute for "watch-like" structure.

Analysis on the pseudo-attributes. Given the the rare class sample as a query, we retrieve images with the same pseudo-attribute labels in Fig. 3. The figure shows that the learned pseudo-attributes which helps the model correctly predict the class of the test images. For example, the first and second rows show the retrieval result of the pseudo-attribute for "dome-like" and "watch-like" structure, respectively. Note that learning this pseudo-attributes is a *self-supervised* learning method and does not require any human labor to label specific attributes.

4 Conclusion

To alleviate the long-tailed distribution in the image classification, we have proposed a self-supervised learning framework with the novel concept of *pseudo-attribute*, which does not require extra annotations. The proposed model achieved the state-of-the-art performance on various long-tailed recognition benchmarks such as CIFAR100-LT, iNaturalist, and ImageNet-LT datasets. The possible future research direction would be applying the pseudo-attributes for other tasks that suffer from the bias problems such as semi-supervised learning [15, 24] or active learning [6, 14, 28].

References

[1] Shin Ando and Chun Yuan Huang. Deep over-sampling framework for classifying imbalanced data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 770–785, 2017.

[2] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

[3] Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning (ICML)*, 2019.

[4] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in Neural Information Processing Systems (NIPS)*, 2019.

[5] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[6] Jae Won Cho, Dong-Jin Kim, Yunjae Jung, and In So Kweon. Mcdal: Maximum classifier discrepancy for active learning. *IEEE transactions on neural networks and learning systems*, 2022.

[7] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[8] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[9] Qi Dong, Shaogang Gong, and Xiatian Zhu. Imbalanced deep learning by minority class incremental rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(6):1367–1381, 2018.

[10] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations (ICLR)*, 2020.

[11] Salman Khan, Munawar Hayat, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Striking the right balance with uncertainty. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[12] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587, 2017.

[13] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems (NIPS)*, 2020.

[14] Dong-Jin Kim, Jae Won Cho, Jinsoo Choi, Yunjae Jung, and In So Kweon. Single-modal entropy based active learning for visual question answering. In *British Machine Vision Conference (BMVC)*, 2021.

[15] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[16] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. In *European Conference on Computer Vision (ECCV)*, 2020.

[17] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Acp++: Action co-occurrence priors for human-object interaction detection. *IEEE Transactions on Image Processing (TIP)*, 30:9150–9163, 2021.

[18] Bolian Li, Zongbo Han, Haining Li, Huazhu Fu, and Changqing Zhang. Trustworthy long-tailed classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[19] Mengke Li, Yiu-ming Cheung, and Yang Lu. Long-tailed visual recognition via gaussian clouded logit adjustment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[20] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[21] Tianhao Li, Limin Wang, and Gangshan Wu. Self supervision to distillation for long-tailed visual recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[22] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[23] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations (ICLR)*, 2021.

[24] Youngtaek Oh, Dong-Jin Kim, and In So 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] Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Young 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.

[26] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. Influence-balanced loss for imbalanced visual classification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 735–744, 2021.

[27] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *European Conference on Computer Vision (ECCV)*, 2016.

[28] Inkyu Shin, Dong-Jin Kim, Jae Won Cho, Sanghyun Woo, KwanYong Park, and In So Kweon. Labor: Labeling only if required for domain adaptive semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[29] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[30] Jason Van Hulse, Taghi M Khoshgoftaar, and Amri Napolitano. Experimental perspectives on learning from imbalanced data. In *International Conference on Machine Learning (ICML)*, 2007.

[31] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations (ICLR)*, 2021.

[32] Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2020.

[33] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*, 2021.

[34] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[35] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.