Efficient Long-Tail Learning in Latent Space by sampling Synthetic Data

Nakul Sharma
Independent Researcher
0xnakul.sh@gmail.com

Abstract

Imbalanced classification datasets pose significant challenges in machine learning, often leading to biased models that perform poorly on underrepresented classes. With the rise of foundation models, recent research has focused on the full, partial, and parameter-efficient fine-tuning of these models to deal with long-tail classification. Despite the impressive performance of these works on the benchmark datasets, they still fail to close the gap with the networks trained using the balanced datasets and still require substantial computational resources, even for relatively smaller datasets. Underscoring the importance of computational efficiency and simplicity, in this work we propose a novel framework that leverages the rich semantic latent space of Vision Foundation Models to generate synthetic data and train a simple linear classifier using a mixture of real and synthetic data for long-tail classification. The computational efficiency gain arises from the number of trainable parameters that are reduced to just the number of parameters in the linear model. Our method sets a new state-of-the-art for the CIFAR-100-LT benchmark and demonstrates strong performance on the Places-LT benchmark, highlighting the effectiveness and adaptability of our simple and effective approach.

1. Introduction

Long-tail classification addresses the reality that real-world data often follow highly skewed distributions: a few head classes have abundant samples, whereas many tail classes are represented by only a handful of examples. The importance of long-tail learning cannot be overstated, as models trained on such imbalanced datasets often exhibit biased performance, excelling on head classes while struggling with tail classes. This bias can lead to severe consequences in critical applications such as medical diagnosis, autonomous driving, and financial fraud detection, where accurate prediction across all classes is paramount. Despite extensive research spanning data re-balancing (resampling or augmentation) strategies, improved representation learning, and adjusted loss functions, the tail-class performance still lags far behind what is achieved on balanced data.

The field of long-tail learning has witnessed significant advancements in recent years, evolving from traditional approaches like data resampling and loss reweighting to more sophisticated techniques leveraging the power of deep learning. Recently, large pre-trained vision models like CLIP [18] have emerged as powerful visual backbones for long-tail recognition. These models are trained on massive, diverse datasets and produce high-quality feature embeddings, which can be leveraged to mitigate class imbalance. By fine-tuning a foundation model (instead of training from scratch on an imbalanced set), researchers have reported substantial gains in overall accuracy [14, 21, 22, 26]. However, the challenge is how to adapt these models without undoing their benefits for tail classes. Recent studies have revealed that heavy fine-tuning of CLIP-based ViT [7] model on imbalanced data can distort the feature space and actually degrade tail-class accuracy [21] – to avoid this pitfall, the authors advocate for lightweight fine-tuning.

In this work, we look at the problem of long-tail learning from the perspective of generating novel data samples for minority classes to balance the training set, and to effectively utilize the latent space of foundation models like CLIP [4]. To this end, we propose a novel long-tail learning approach that operates entirely in the latent feature space of a frozen vision foundation model. Specifically, we use a pre-trained CLIP ViT encoder to embed all images into a high-dimensional feature space rich in semantic information. Rather than fine-tuning the backbone on the imbalanced data, we keep the feature extractor fixed and address the imbalance by generating synthetic features for all the classes so that their cardinality becomes equal. For each minority class, we perform a simple kernel density estimation (KDE) on its few available feature vectors, using von Mises-Fisher kernels to account for the directional nature of the normalized CLIP embeddings. This KDE effectively models a smooth manifold of plausible feature vectors around the known samples. We then sample additional feature points from this estimated distribution, yielding "synthetic data" for the tail class without ever gener-

ating images. Finally, we train a linear classifier on the augmented dataset composed of the original real features plus the synthetic features for the tail classes. The classifier training is a lightweight optimization, and by construction, it sees a much more balanced training set.

Our novelty lies in generating synthetic training examples in the latent space of a powerful vision model and using an extremely simple learning algorithm (linear classification) to achieve high accuracy on the CIFAR-100-LT and the Places-LT benchmark. This stands in contrast to prior works that either fine-tune large networks or train complex generative models to augment data for class-balancing. By avoiding any heavy network updates and repeated forward passes, our approach is efficient in computation and memory. It requires neither external data nor multi-stage training; in fact, once features are extracted (which can be done in a single forward pass per image), the rest of the training pipeline is akin to training a logistic regression on an augmented feature set. Furthermore, our approach is model-agnostic, allowing it to be seamlessly integrated with future vision foundation models and to deliver ongoing, incremental gains across a wide range of imbalanced datasets. To summarize, our contributions are as follows: i) We introduce a simple approach to efficiently utilize the pre-trained embedding space of vision foundation models, ii) We empirically validate our approach on the CIFAR-100-LT and Places-LT benchmarks, achieving state-of-the-art and competitive performance, respectively, and iii) We ablate the choice of encoders and the latent sampling to quantify the components that make our approach work better.

2. Methodology

Our method leverages the pre-trained vision encoder of OpenCLIP [4] to obtain the latent embeddings of all images. Since empirical evidence suggests that normalized visual embeddings for similar images obtained from this contrastive model lie close to each other on a unit hypersphere [8, 16], we estimate the latent distribution of each class using a mixture of von Mises-Fisher (vMF) distribution kernels. The vMF distribution is one of the simplest parametric distributions for hyperspherical data, and has properties analogous to those of the multi-variate Gaussian distribution for data in $\mathbb{R}^d$ [15, 19]. Once estimated, we sample new data points in latent space from the estimated distribution such that the number of samples in all classes of the data becomes equal. These newly sampled latents serve as novel synthetic data for each class, and are used to train a logistic regression model along with the original latents for classification. The following sections detail the extraction of latent embeddings, estimation of the latent distribution for each class, and generation of synthetic data.

Latent Embedding Extraction. For a given dataset with $|C|$ classes, let I^k denote the set of images that belong to class k, where $k \in \{0, 1, \ldots, C-1\}$. For each image $\mathbf{x} \in I^k$, we compute its latent embedding $\mathbf{z} \in \mathbb{R}^d$ using the frozen vision encoder f_θ and ℓ_2–normalize it:

$$\mathbf{z} = \frac{f_\theta(\mathbf{x})}{\|f_\theta(\mathbf{x})\|_2} \in \mathbb{S}^{d-1}, \quad \mathbf{z} \in Z^k = \{\mathbf{z}_i^k\}_{i=1}^{N_k}. \tag{1}$$

Here $\mathbb{S}^{d-1}$ is the unit hypersphere in $\mathbb{R}^d$ and $N_k = |I^k|$.

Class–Conditional Density Estimation. Since all embeddings lie on $\mathbb{S}^{d-1}$, we approximate the density of the latent embeddings of class k using the Kernel Density Estimation [23] with kernels given by von Mises-Fisher distributions centered at each of the observed latent embeddings:

$$\hat{p}_k(\cdot) = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathrm{vMF}\big(\mathbf{z}; \boldsymbol{\mu} = \mathbf{z}_i^k, \kappa = \kappa_{\mathbf{z}_i^k}\big), \tag{2}$$

where the von Mises-Fisher (vMF) probability density function is defined as:

$$\mathrm{vMF}(\mathbf{z}; \boldsymbol{\mu}, \kappa) = C_d(\kappa) \exp\big(\kappa \boldsymbol{\mu}^\top \mathbf{z}\big), \tag{3}$$

where $C_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}$, and $\kappa > 0$ controls the concentration of the distribution, and $I_\nu(\cdot)$ denotes the modified Bessel function of order ν. To estimate vMF KDE in Equation (2), we need to estimate the concentration $\kappa_{\mathbf{z}_i^k}$ at each embedding $\mathbf{z}_i^k$.

Estimating the Concentration $\kappa_{\mathbf{z}_i^k}$. Banerjee et al. [1] derives that the concentration κ_k for the observations Z^k can be approximated as $\kappa_k \approx \frac{\bar{R}_k(d-\bar{R}_k^2)}{1-\bar{R}_k^2}$ where $\bar{R}_k$ is the sample resultant length given by $\bar{R}_k = \left\| \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbf{z}_i^k \right\|_2$. This closed-form solution for κ_k yields accurate estimates with an almost negligible computational cost. However, we are interested in estimating $\kappa_{\mathbf{z}_i^k}$ in Equation (2) which is entirely different from κ_k, the concentration estimate for the whole distribution of the k-th class.

To estimate the local concentration parameter $\kappa_{\mathbf{z}_i^k}$ at each latent embedding $\mathbf{z}_i^k$, we exploit the empirical observation that latent embeddings of visually similar images exist closely on the unit hypersphere. Specifically, for each embedding $\mathbf{z}_i^k$, we first identify its nearest neighbor $\mathbf{z}_j^k$ from the same class k in the latent space using cosine similarity such that $\mathbf{z}_j^k = \arg\max_{\mathbf{z} \in \mathbf{Z}^k \setminus \mathbf{z}_i^k} (\mathbf{z}_i^k)^\top \mathbf{z}$.

Since these embeddings belong to same class k and $\mathbf{z}_i^k$ is closest to $\mathbf{z}_j^k$, we assume that both of these embeddings originate from the same underlying concentrated latent vMF distribution. Under this assumption, we form the local dataset $\tilde{Z}_i^k = \{\mathbf{z}_i^k, \mathbf{z}_j^k\}$ and estimate the concentration parameter $\kappa_{\mathbf{z}_i^k}$ using the method proposed by Banerjee et

al. [1] as follows:

$$\kappa_{\mathbf{z}_i^k} \approx \frac{\tilde{R}_i^k(d-(\tilde{R}_i^k)^2)}{1-(\tilde{R}_i^k)^2}, \quad \text{where} \quad \tilde{R}_i^k = \left\| \frac{1}{2}(\mathbf{z}_i^k + \mathbf{z}_j^k) \right\|_2. \tag{4}$$

The calculated $\kappa_{\mathbf{z}_i^k}$ thus reflects the local concentration at the embedding $\mathbf{z}_i^k$, and helps in effectively capturing variations in density across different regions of the class-conditional latent space. These locally estimated concentrations are then used in Equation (2) to accurately model the class-specific distribution of latent embeddings.

Synthetic Latent Generation After estimating the density using Equation (2), we generate synthetic latent embeddings to balance each class. Specifically, we draw $\tilde{N}_k = N_{\max} - N_k$ synthetic samples per class, where $N_{\max} = \max_c N_c$, ensuring all the classes reach equal cardinality. We utilize Wood's rejection sampling method [27] to efficiently sample from the estimated mixture of vMF distributions for each class k. The augmented and balanced latent set $\mathcal{Z}^k$ for class k is expressed as follows:

$$\mathcal{Z}^k = Z^k \cup \{\tilde{\mathbf{z}}_j^k\}_{j=1}^{\tilde{N}_k}, \quad \tilde{\mathbf{z}}_j^k \sim \hat{p}_k(\cdot), \tag{5}$$

where $\hat{p}_k(\cdot)$ is defined in Equation (2). After generating synthetic samples for each minority class, we merge the balanced embedding sets into $\mathcal{Z}$ and fit a multinomial logistic-regression classifier W_ϕ with L–BFGS. At inference, we encode an image I_t, ℓ_2-normalize its embedding $z = f_\theta(\mathbf{I}_t)/\|f_\theta(\mathbf{I}_t)\|_2$, and predict $\hat{\mathbf{y}} = \arg\max W_\phi(\mathbf{z})$.

3. Experiments

We conduct comprehensive experiments to demonstrate the effectiveness, robustness, and versatility of our proposed method. Our evaluation is focused on widely used long-tail benchmarks, specifically CIFAR-100-LT [2] and Places-LT [13], chosen due to their prevalence and challenging nature in the domain of imbalanced classification tasks. Performance is primarily assessed using top-1 accuracy, offering a clear and direct measure of classification effectiveness. Additionally, we conduct detailed ablation studies to gain deeper insights into the influence of encoders and synthetic data generation methodologies.

Experimental Setup. For extracting latent embeddings, we employ the Vision Transformer ViT-L/14 encoder from the OpenCLIP [4], due to its proven ability to capture discriminative visual features across diverse datasets. To model the latent embeddings, we adopt a mixture of vMF distributions, implementing our approach by extending the base `vonmises_fisher` class available in the SciPy library [24]. For classifier training, we utilize logistic regression optimized through the L-BFGS algorithm [11], using the scikit-learn library [17].

Table 1. Comparison with state-of-the-art methods on CIFAR-100-LT with various imbalance ratios.

Method	Learnable Params.	Imbalance Ratio 100	50	10
Training from scratch				
LDAM [2]	0.46M	42.0	46.6	58.7
BBN [33]	0.46M	42.6	47.0	59.1
DiVE [9]	0.46M	45.4	51.1	62.0
MiSLAS [32]	0.46M	47.0	52.3	63.2
BS [20]	0.46M	50.8	54.2	63.0
PaCo [5]	0.46M	52.0	56.0	64.2
BCL [34]	0.46M	51.9	56.6	64.9
Fine-tuning pre-trained model				
LiVT [28]	85.80M	58.2	-	69.2
BALLAD [14]	149.62M	77.8	-	-
LIFT [21]	0.10M	80.3	82.0	83.8
LIFT+ [22]	0.10M	81.7	83.1	84.7
Utilizing frozen pre-trained model				
Ours	0.10M	**89.0**	**90.3**	**91.4**

Table 2. Comparison with state-of-the-art methods on Places-LT.

Method	Learnable Params.	Overall	Head	Medium	Tail
Training from scratch (init. from ImageNet-1K backbone)					
OLTR [12]	58.14M	35.9	44.7	37.0	25.3
cRT [10]	58.14M	36.7	42.0	37.6	24.9
LWS [10]	58.14M	37.6	40.6	39.1	28.6
MiSLAS [32]	58.14M	40.4	39.6	43.3	36.1
DisAlign [30]	58.14M	39.3	40.4	42.4	30.1
ALA [31]	58.14M	40.1	43.9	40.1	32.9
PaCo [5]	58.14M	41.2	36.1	47.9	35.3
LiVT [28]	85.80M	40.8	48.1	40.6	27.5
Fine-tuning foundation model					
BALLAD [14]	149.62M	49.5	49.3	50.2	48.4
Decoder [25]	21.26M	46.8	-	-	-
LPT [6]	1.01M	50.1	49.3	**52.3**	46.9
LIFT [21]	0.18M	**51.5**	**51.3**	52.2	50.5
LIFT+ [22]	0.18M	**51.5**	50.8	52.0	**51.6**
Utilizing frozen pre-trained model					
Ours	0.37M	48.3	49.3	48.4	47.0

Comparison with State-of-the-Art. We benchmark our method against recent works that address long-tail learning using vision foundation models, as well as traditional approaches known for low computational demands on CIFAR-100-LT [2] and Places-LT [13] benchmarks.

Table 3. Comparison of training strategy for ViT on CIFAR-100-LT, trained across different imbalance ratios.

Method	Size	Imbalance Ratio 10	50	100
SigLIP	B/16	80.84	77.52	75.22
	L/14	87.65	85.61	83.96
OpenCLIP	B/16	87.67	85.43	83.49
	L/14	**91.44**	**90.31**	**89.05**

Table 4. Ablation study of synthetic data generation methods on CIFAR-100-LT trained across different imbalance ratios.

Method	Imbalance Ratio 10	50	100
No Synth. Data	90.73	86.66	84.08
OS w/ replacement	91.24	90.13	87.32
SMOTE	91.36	90.17	88.66
Gaussian KDE	91.18	88.99	86.94
vMF-KDE (Ours)	**91.44**	**90.31**	**89.05**

Results for the CIFAR-100-LT benchmark in Table 1 show that our approach achieves a substantial performance leap over existing methods at all imbalance ratios. Notably, under the most challenging imbalance ratio (100), our method reaches an impressive top-1 accuracy of 89%, outperforming the previous best LIFT+ [22] baseline by 7.3%. Even at less severe imbalance ratios (50 and 10), our method maintains a superior accuracy (90.3% and 91.4%, respectively), demonstrating its robustness across varying levels of class imbalance.

The competitive performance of our method on the Places-LT benchmark reported in Table 2 demonstrates that our density-guided synthesis strategy scales gracefully from small-scale to large-scale long-tail settings. Despite the encouraging results, our method still trails the latest adapter-based fine-tuning techniques.We conjecture that this shortfall stems from limitations of the frozen latent embeddings themselves — we use pooled embeddings, which could lead to residual semantic overlap in the latent space in a physical-world dataset like Places-LT where some categories share coarse features which require more fine-grained representation. In future work, we plan to investigate this further and develop strategies that enhance embedding discriminability.

In addition to our strong predictive performance, our method distinguishes itself by its computational efficiency. Unlike fine-tuning-based approaches, which require repeated forward and backward passes through all data across the full network during each epoch, our method needs only a single forward pass over the pre-trained backbone to extract embeddings. Subsequent steps are limited to efficient density modeling and lightweight classifier training, resulting in reduced compute requirements.

Ablations. First, we investigate how different contrastive learning objectives affect performance on the CIFAR-100-LT benchmark. We compare softmax-normalized contrastive learning backbone OpenCLIP [4] against non-softmax alternative SigLIP [29]. As shown in Table 3, softmax-based encoders consistently outperform their non-softmax counterparts across all imbalance ratios. This gap suggests that the softmax normalization's competitive dynamics create more compact and well-separated class manifolds in the latent space—this is particularly advantageous for our vMF-based density estimation, as they reduce inter-class overlap when generating synthetic samples. Moreover, scaling to larger encoder sizes yields consistent improvements, underscoring the importance of model capacity for capturing discriminative features.

Second, we investigate the impact of synthetic data generation techniques. We use OpenCLIP ViT-L/14 encoder to compare our proposed vMF-based KDE with Gaussian KDE, SMOTE [3], random oversampling, and a baseline that does not generate synthetic data. Results on the CIFAR-100-LT benchmark presented in Table 4 clearly indicate that our proposed vMF KDE method outperforms all other synthetic data generation strategies, validating its effectiveness in accurately modeling latent distributions on a hypersphere and generating high-quality synthetic embeddings. Notably, our proposed method achieves the greatest performance improvement on CIFAR-100 with an imbalance ratio (IR) of 100, outperforming both the baseline and other competing methods. This result further highlights the effectiveness of our approach in handling highly imbalanced datasets.

4. Conclusion

In this preliminary work, we propose a novel and efficient approach to long-tail learning by generating synthetic samples in the latent space of powerful vision foundation models. Our method uses kernel density estimation with the vMF kernel to jointly model and synthesize high-quality synthetic latent samples for minority classes, enabling the training of a simple linear classifier on a balanced and enriched feature set. This approach achieves strong empirical results on challenging long-tail benchmarks such as CIFAR-100-LT and Places-LT, while significantly reducing computational overhead compared to existing state-of-the-art methods. Building on our promising results, future work will focus on (i) reducing the latent semantic overlap in the pre-trained representations to learn better classifiers, and (ii) systematically evaluating and, where possible, extending the transferability of the proposed approach across different modalities and application domains.

References

[1] Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(46):1345–1382, 2005. 2, 3

[2] 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*, 32, 2019. 3

[3] 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. 4

[4] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 1, 2, 3, 4

[5] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 715–724, 2021. 3

[6] Bowen Dong, Pan Zhou, Shuicheng Yan, and Wangmeng Zuo. LPT: Long-tailed prompt tuning for image classification. In *International Conference on Learning Representations*, 2023. 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1

[8] Stephanie Fu, Netanel Y. Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: learning new dimensions of human visual similarity using synthetic data. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023. 2

[9] Yin-Yin He, Jianxin Wu, and Xiu-Shen Wei. Distilling virtual examples for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 235–244, 2021. 3

[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*, 2020. 3

[11] Dong C. Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Math. Program.*, 45:503–528, 1989. 3

[12] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. 3

[13] 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. 3

[14] Teli Ma, Shijie Geng, Mengmeng Wang, Jing Shao, Jiasen Lu, Hongsheng Li, Peng Gao, and Yu Qiao. A simple long-tailed recognition baseline via vision-language model. *arXiv preprint arXiv:2111.14745*, 2021. 1, 3

[15] Kanti V. Mardia. *Statistical Distributions in Scientific Work*, chapter Characteristics of directional distributions, pages 365–385. Reidel, Dordrecht, 1975. 2

[16] Prasanna Mayilvahanan, Thaddäus Wiedemer, Evgenia Rusak, Matthias Bethge, and Wieland Brendel. Does CLIP's generalization performance mainly stem from high train-test similarity? In *The Twelfth International Conference on Learning Representations*, 2024. 2

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 3

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1

[19] C. Radhakrishna Rao. *Linear Statistical Inference and its Applications*. Wiley, New York, 2 edition, 1973. 2

[20] Jiawei Ren, Cunjun Yu, shunan sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems*, pages 4175–4186, 2020. 3

[21] Jiang-Xin Shi, Tong Wei, Zhi Zhou, Jie-Jing Shao, Xin-Yan Han, and Yu-Feng Li. Long-tail learning with foundation model: Heavy fine-tuning hurts. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. 1, 3

[22] Jiang-Xin Shi, Tong Wei, and Yu-Feng Li. Lift+: Lightweight fine-tuning for long-tail learning. *arXiv preprint arXiv:2504.13282*, 2025. 1, 3, 4

[23] B.W. Silverman. Density estimation for statistics and data analysis. *Routledge*, 1998. 2

[24] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. 3

[25] Yidong Wang, Zhuohao Yu, Jindong Wang, Qiang Heng, Hao Chen, Wei Ye, Rui Xie, Xing Xie, and Shikun Zhang.

Exploring vision-language models for imbalanced learning. *International Journal of Computer Vision*, 132:224–237, 2024. 3

[26] Yidong Wang, Zhuohao Yu, Jindong Wang, Qiang Heng, Hao Chen, Wei Ye, Rui Xie, Xing Xie, and Shikun Zhang. Exploring vision-language models for imbalanced learning. *International Journal of Computer Vision*, 132(1):224–237, 2024. 1

[27] Andrew T. A. Wood. Simulation of the von mises fisher distribution. *Communications in Statistics - Simulation and Computation*, 23:157–164, 1994. 3

[28] Zhengzhuo Xu, Ruikang Liu, Shuo Yang, Zenghao Chai, and Chun Yuan. Learning imbalanced data with vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15793–15803, 2023. 3

[29] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 4

[30] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2361–2370, 2021. 3

[31] Yan Zhao, Weicong Chen, Xu Tan, Kai Huang, and Jihong Zhu. Adaptive logit adjustment loss for long-tailed visual recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3472–3480, 2022. 3

[32] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16489–16498, 2021. 3

[33] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9719–9728, 2020. 3

[34] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6908–6917, 2022. 3

Efficient Long-Tail Learning in Latent Space by sampling Synthetic Data

Supplementary Material

Extended Discussion

While this preliminary study demonstrates strong potential for leveraging semantic representations from vision foundation models in long-tail recognition, several aspects merit deeper investigation to fully realize this approach.

Domain Transferability. Our empirical success raises fundamental questions about when and why latent-space augmentation succeeds. A systematic investigation of vision foundation model embeddings — particularly quantifying inter-class overlap, analyzing clustering quality across semantic categories, and establishing relationships between pre-training diversity and downstream performance would provide crucial insights. This analysis could explain our contrasting results on the Places-LT benchmark, where scene categories likely exhibit greater semantic ambiguity than object-centric datasets. Understanding these distributional properties would enable principled sampling strategies. Such theoretical grounding would help practitioners predict which visual recognition tasks benefit most from embedding-space augmentation versus traditional fine-tuning approaches.

Trade-offs of the Frozen Encoder Paradigm. Our approach achieves efficiency through single-pass feature extraction and a lightweight linear classifier, yet the quality of synthetic samples remains bounded by the pre-trained backbone's discriminative capacity, as shown in Table 3. Class ambiguities or suboptimal decision boundaries in these representations create an irreversible bottleneck, which is particularly problematic for fine-grained datasets where subtle inter-class distinctions matter. For instance, distinguishing between similar bird species or architectural styles requires feature refinements that frozen general-purpose pooled embedding cannot directly provide. This limitation suggests a promising direction: leveraging full spatial feature maps rather than single pooled embeddings. Patch-wise representations could capture fine-grained distinctions while maintaining computational efficiency, potentially through attention-weighted pooling or learnable aggregation mechanisms that still require minimal parameters.