BRADD: Balancing Representations with Anomaly Detection and Diffusion

Filipe Laitenberger
University of Amsterdam
Netherlands
filipe.laitenberger@student.uva.nl

Nesta Midavaine
University of Amsterdam
Netherlands

Ioana Simion
University of Amsterdam
Netherlands

Stefan Vasilev
University of Amsterdam
Netherlands

Hirokatsu Kataoka
AIST / University of Oxford
Japan / United Kingdom

Cees G. M. Snoek
University of Amsterdam
Netherlands

Yuki M Asano
University of Technology Nuremberg
Germany

Mohammadreza Salehi
University of Amsterdam
Netherlands

Abstract

Self-supervised learning (SSL) has allowed for advancements in language processing and computer vision, as unlabelled data is available in large quantities. However, imbalances in training datasets can lead to strong biases in the learned features of pre-trained models. Previous results show that pre-training using imbalanced data can also hurt downstream performance. We propose a data-centric approach: our method trains on the data, finds underrepresented samples, and uses diffusion to generate novel data complementing the underrepresented images. Our proposed method, BRADD (Balancing Representations through Anomaly Detection and Diffusion), utilizes distance-based outlier detection to identify regions of the embedding space that are underrepresented in each training cycle. Experimental results on ImageNet-100-LT demonstrate that BRADD consistently outperforms both balanced and imbalanced baselines, with significant improvements on fine-grained classification tasks. Detailed ablation studies confirm that both out-of-distribution sample selection and diffusion-based generation contribute substantially to the effectiveness of our approach, offering a promising alternative to model-centric solutions for addressing imbalance in self-supervised learning.

1. Introduction

Self-supervised learning (SSL) methods have emerged as powerful techniques for learning transferable features across diverse tasks [2, 3, 17]. However, their performance is significantly affected by dataset imbalance [1, 12, 23]. For instance, SimCLR underperforms on long-tailed datasets due to insufficient negative samples [23], while joint-embedding methods like VICReg assume uniform clustering, hampering performance on imbalanced data [1].

Existing solutions typically adopt model-centric approaches, such as ensemble learning [23], incorporating arbitrary feature priors [1], or modifying training dynamics [12]. These approaches, however, often require prior knowledge of dataset distributions or extensive hyperparameter tuning, contradicting the unsupervised nature of SSL.

We propose BRADD (**B**alancing **R**epresentations with **A**nomaly **D**etection and **D**iffusion), a data-centric approach to address imbalance in SSL. BRADD divides training into cycles, identifies underrepresented samples via OOD detection after each cycle, and augments them using diffusion models for subsequent training. This approach (1) eliminates the need for prior dataset knowledge and (2) dynamically balances the latent space to avoid suboptimal local minima.

Experiments on ImageNet-100-LT across multiple architectures (ResNet-50, ViT-S, ViT-B) and SSL methods (Sim-

CLR, DINO, MoCo) demonstrate that BRADD consistently outperforms both balanced and imbalanced baselines on diverse downstream tasks. BRADD achieves significant improvements on fine-grained tasks (up to 11.7% on Oxford Flowers) and surpasses state-of-the-art methods on CIFAR-10 (73.3%) and CIFAR-100 (45.4%), showing that our data-centric approach offers a promising alternative to model-centric solutions.

2. Related Work

Self-Supervised Learning (SSL) leverages unlabeled data to learn meaningful representations for downstream tasks. In computer vision, three main approaches are: (1) masked prediction [7], (2) contrastive learning like MoCo [6] and SimCLR [3], and (3) self-distillation methods such as DINO [2]. These techniques have enabled large-scale training of models with emergent abilities [24].

Training on Imbalanced Datasets often leads to inferior performance and bias toward majority classes [10]. While SSL methods are more robust to imbalance than supervised approaches [13], they still exhibit diminished performance on imbalanced data. Current solutions are primarily model-centric, attempting to learn arbitrary feature priors [1]. Data-centric approaches that directly complement imbalanced datasets remain underexplored.

Out-Of-Distribution (OOD) Detection is crucial for enhancing model robustness by maintaining high-quality datasets [25]. Distance-based approaches [21] identify OOD samples by measuring their distance from in-distribution samples in the embedding space. These non-parametric methods offer flexibility without distributional assumptions, making them suitable for imbalanced, unlabeled data.

Diffusion-Based Image Generation models systematically add and then remove noise to generate high-quality images [8]. Stable Diffusion [20], particularly Stable Diffusion 2 UnClip, leverages CLIP embeddings to generate semantically similar images to the input context. Its ability to perform image-to-image generation while preserving semantic content makes it well-suited for augmenting self-supervised learning datasets.

3. Method: BRADD

We propose BRADD (Balancing Representations with Anomaly Detection and Diffusion), a data-centric approach to address imbalance in self-supervised learning. Unlike model-centric approaches, BRADD identifies and augments underrepresented concepts in the data distribution itself.

BRADD alternates between self-supervised pre-training and dataset augmentation phases (Algorithm 1). After training for N_E epochs, we identify underrepresented data points by computing k-nearest neighbor distances in the embedding space. Rather than using a percentile threshold, we select the top-N_{OOD} samples with highest k-NN distances, providing precise control over augmentation.

Algorithm 1 BRADD: Balancing Representations through Anomaly Detection and Diffusion

```
Input: model $m$, SSL algorithm $a$, diffusion model $m_d$, cycles $N_C$, epochs per cycle $N_E$, dataset $D$, samples per point $N_{Aug}$, OOD samples $N_{OOD}$
repeat
    Train $m$ using $a$ for $N_E$ epochs on $D$
    Compute embeddings $E = \{m(x_i) | x_i \in D\}$
    Find OOD points $O = \{x_i | \text{rank}(d_{kNN}(m(x_i), E)) \leq N_{OOD}\}$
    for $x \in O$ do
        $D = D \cup \{m_d(x, \epsilon_j) | j = 1, ..., N_{Aug}\}$
    end for
until $N_C$ cycles completed
```

For each identified OOD point, we generate N_{Aug} new samples using Stable Diffusion 2 UnCLIP [19], which preserves semantic content while introducing sufficient variation. We use k=5 for k-NN computation, $N_{OOD} = 500$, and $N_{Aug} = 5$, adding 2,500 new images per cycle across 5 cycles with 20 epochs each (100 total epochs).

4. Experiments

To evaluate the performance of our proposed method BRADD (Balancing Representations through Automated Detection and Diffusion), we conduct experiments on ImageNet-100-LT. We test various backbone architectures, SSL methods, and implementation details through comprehensive ablation studies, and compare against state-of-the-art approaches.

4.1. ImageNet-100 Experiments

ImageNet-100 is a subset of ImageNet with 100 randomly selected classes [22]. Following previous work [14], we introduce imbalance by creating a long-tailed distribution (ImageNet-100-LT) that follows a Pareto distribution with $\alpha = 6$. The dataset contains around 15 thousand images.

Models. We evaluate our method across multiple architectures: ResNet-50 (25.6M parameters), ViT-Small (22.1M parameters), and ViT-Base (86.6M parameters).

Self-Supervised Learning Methods. Our primary experiments use SimCLR [3] with a temperature of 0.5, and we later compare with DINO [2] and MoCo [5] in our ablation studies.

Downstream Evaluation. We evaluate the learned features using both linear probing and K-nearest neighbor (KNN) classification across multiple datasets: CIFAR-10, CIFAR-100 [11], Stanford Cars [4], FGVC Aircraft [15], Oxford

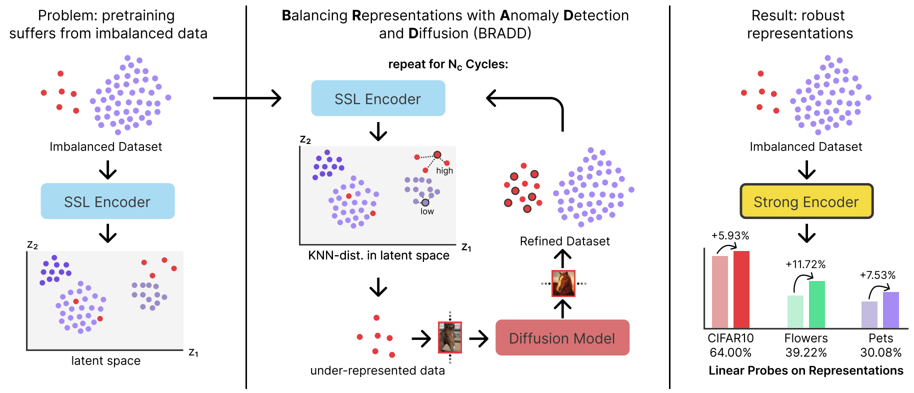

Figure 1. The BRADD algorithm augments the most underrepresented datapoints in a dataset based on OOD detection.

Table 1. **Comparison to balanced and imbalanced baselines**
(ViT-B, SimCLR, ImageNet-100-LT, 100 epochs)

	C-10		**C-100**		**Cars**		**Aircraft**		**Flowers**		**Pets**	
Setting	Lin.	KNN	Lin.	KNN	Lin.	KNN	Lin.	KNN	Lin.	KNN	Lin.	KNN
Balanced	59.90	57.13	33.63	28.34	4.29	3.31	9.14	4.83	31.25	26.04	23.64	13.04
Imbalanced	59.89	57.25	33.02	27.78	3.80	3.43	8.75	5.37	27.50	25.42	22.55	11.96
BRADD (ours)	**64.00**	**60.73**	**38.42**	**31.56**	**6.13**	**4.54**	**11.55**	**6.72**	**39.22**	**30.89**	**30.08**	**14.69**

Flowers [16], and Oxford-IIIT Pets [18].

Baselines. We train two baselines: (1) a model trained on a balanced subset of ImageNet-100 with as much data uniformly removed as in ImageNet-100-LT and (2) a model trained on the imbalanced ImageNet-100-LT.

Proposed Method. Our proposed method BRADD starts with ImageNet-100-LT and trains for multiple cycles, where each cycle consists of training epochs followed by OOD detection and generation steps.

4.2. Experimental Results

Table 1 compares our method against balanced and imbalanced baselines using a ViT-B backbone with SimCLR. The results demonstrate that:

- While imbalance causes only slight performance degradation compared to the balanced setting on some datasets, our method consistently outperforms both baselines across all datasets.
- BRADD achieves substantial improvements in linear probing, with gains of up to 11.72% on Oxford Flowers and 7.53% on Oxford-IIIT Pets compared to the imbalanced baseline.
- Our method also shows consistent improvements in KNN classification, demonstrating the enhanced quality of the learned feature space.

Table 2 shows BRADD compared to state-of-the-art methods using ResNet-50 with SimCLR (500 epochs):

- Our method achieves superior linear probing performance on CIFAR-10 (73.34%) and CIFAR-100 (45.38%) compared to previous methods.
- While TS [12] performs better on KNN classification for most datasets, BRADD shows competitive performance across all benchmarks.

4.3. Ablation Studies

To analyze the effectiveness of different components in our approach, we conducted extensive ablation studies:

SSL Method. Table 3a demonstrates that SimCLR consistently outperforms DINO and MoCo across all datasets when using our method, with substantial margins particularly on fine-grained classification tasks.

Backbone Architecture. As shown in Table 3b, we compared ViT-S, ViT-B, and ResNet-50 backbones. ResNet-50 achieves the best linear probing performance on CIFAR-10 (71.17%), while ViT architectures perform better on fine-grained datasets, with ViT-B showing the strongest performance on Oxford-IIIT Pets (30.08%).

Sample Selection Strategy. In Table 3c, we compare uniform sampling versus our OOD-based selection. The results confirm that OOD-based selection provides consistent performance gains, validating our hypothesis that targeting underrepresented regions of the feature space is more effective than random augmentation.

Sample Generation Method. Table 3d compares re-

Table 2. **Comparison to SOTA**
(ResNet50, SimCLR, ImageNet-100-LT, 500 epochs)

	C-10		**C-100**		**Cars**		**Aircraft**		**Flowers**		**Pets**	
Method	Lin.	KNN	Lin.	KNN	Lin.	KNN	Lin.	KNN	Lin.	KNN	Lin.	KNN
SDCLR [9]	68.72	65.16	38.71	35.53	7.84	5.21	11.24	10.11	11.87	31.14	40.22	21.98
TS [12]	71.26	**66.76**	43.90	**35.64**	**10.91**	**5.58**	**12.95**	**10.85**	32.05	3.36	**47.01**	**23.20**
BRADD (ours)	**73.34**	61.76	**45.38**	32.26	8.95	4.73	10.95	7.92	26.14	24.94	34.68	16.05

Table 3. **Ablations of the key components of our method.**

(a) Pretraining
(ViT-B, ImageNet-100-LT, 100 epochs)

	C-10		**Flowers**		**Pets**	
backbone	Lin.	KNN	Lin.	KNN	Lin.	KNN
SimCLR	**63.40**	**60.73**	**39.22**	**30.89**	**30.08**	**14.69**
DINO	47.97	37.36	20.92	12.46	10.30	6.00
MoCo	49.13	42.86	15.03	17.63	14.36	7.50

(b) Backbone
(SimCLR, ImageNet-100-LT)

	C-10		**Flowers**		**Pets**	
backbone	Lin.	KNN	Lin.	KNN	Lin.	KNN
ViT-S	64.48	**62.13**	**44.44**	**34.78**	27.91	**15.94**
R50	**71.17**	59.06	26.80	21.07	26.29	15.43
ViT-B	63.40	60.73	39.22	30.89	**30.08**	14.69

(c) Sample Selection
(ViT-B, SimCLR, ImageNet-100-LT)

		C-10		**Flowers**		**Pets**	
SSL	Method	Lin.	KNN	Lin.	KNN	Lin.	KNN
SimCLR	uniform	62.01	58.81	31.37	29.56	23.85	13.25
	OOD	**63.40**	**60.73**	**39.22**	**30.89**	**30.08**	**14.69**
MoCoV3	uniform	51.38	41.08	21.79	16.16	12.90	7.14
	OOD	52.42	42.13	19.87	17.12	11.83	11.56

(d) Sample Generation
(ViT-B, SimCLR, ImageNet-100-LT)

		C-10		**Flowers**		**Pets**	
SSL	Method	Lin.	KNN	Lin.	KNN	Lin.	KNN
SimCLR	Re-Popul.	63.36	59.68	30.07	29.24	26.02	14.12
	Stable-Diff.	**63.40**	**60.73**	**39.22**	**30.89**	**30.08**	**14.69**
MoCoV3	Re-Popul.	48.76	38.68	24.56	15.56	19.23	15.50
	Stable-Diff.	52.42	42.13	19.87	17.12	11.83	11.56

(e) Number of Cycles
(ViT-S, SimCLR, ImageNet-100-LT, 100 Epochs, 10k created images total)

	C-10		**Flowers**		**Pets**	
#Cycles	Lin.	KNN	Lin.	KNN	Lin.	KNN
2	64.08	61.15	**39.87**	**31.63**	26.02	**14.83**
5	63.40	60.73	39.22	30.89	**30.08**	14.69
10	64.46	60.63	31.38	31.28	27.91	14.23
20	**68.68**	**62.02**	35.95	31.56	26.56	14.72

(f) Number of most OOD samples selected for aug. per cycle
(ViT-B, SimCLR, ImageNet-100-LT, 100 Epochs, 2500 created images per cycle)

	C-10		**Flowers**		**Pets**	
#Samples	Lin.	KNN	Lin.	KNN	Lin.	KNN
5	62.09	59.04	33.33	28.55	25.47	13.79
500	**63.40**	**60.73**	**39.22**	**30.89**	**30.08**	**14.69**
1250	58.09	56.10	28.10	24.21	27.10	11.50
2500	57.03	55.75	24.18	23.15	23.57	11.42

population (adding back removed samples) versus generation using Stable Diffusion. While both approaches improve over the baseline, Stable Diffusion generation yields better results, particularly on fine-grained datasets like Oxford Flowers.

Number of Cycles. As shown in Table 3e, we tested different numbers of cycles (2, 5, 10, 20) while keeping the total number of generated images fixed. Performance generally improves with more cycles, with the best results at 20 cycles for CIFAR-10 (68.68% linear probing accuracy) and 5 cycles for Oxford-IIIT Pets (30.08%).

Number of OOD Samples. Table 3f analyzes the effect of selecting different numbers of samples per cycle for augmentation. We find that 500 samples per cycle provides the optimal balance, while selecting too many samples (2500) degrades performance, suggesting that focusing on the most out-of-distribution samples is important.

5. Conclusion

Our comprehensive experiments demonstrate that BRADD effectively mitigates the negative impact of dataset imbalance in self-supervised learning. By strategically detecting regions of the feature space that are underrepresented and augmenting them with generated samples, our method achieves consistent improvements across different architectures, SSL methods, and downstream tasks. The ablation studies confirm that both OOD-based sample selection and diffusion-based generation contribute significantly to the effectiveness of our approach.

References

[1] M Assran, R Balestriero, Q Duval, F Bordes, I Misra, P Bojanowski, P Vincent, M Rabbat, and N Ballas. The hidden uniform cluster prior in self-supervised learning. *URL https://arxiv. org/abs/2210.07277*, 19:26, 2022a. 1, 2

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *CoRR*, abs/2104.14294, 2021. 1, 2

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2

[4] Afshin Dehghan, Syed Zain Masood, Guang Shu, and Enrique. G. Ortiz. View independent vehicle make, model and color recognition using convolutional neural network, 2017. 2

[5] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *CoRR*, abs/1911.05722, 2019. 2

[6] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 2

[9] Ziyu Jiang, Tianlong Chen, Bobak Mortazavi, and Zhangyang Wang. Self-damaging contrastive learning, 2021. 4

[10] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6 (1):1–54, 2019. 2

[11] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2

[12] Anna Kukleva, Moritz Böhle, Bernt Schiele, Hilde Kuehne, and Christian Rupprecht. Temperature schedules for self-supervised contrastive methods on long-tail data, 2023. 1, 3, 4

[13] Hong Liu, Jeff Z HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. *arXiv preprint arXiv:2110.05025*, 2021. 2

[14] 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. 2

[15] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 2

[16] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics Image Processing*, pages 722–729, 2008. 3

[17] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[18] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 3

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 2

[21] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022. 2

[22] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020. 2

[23] Yonglong Tian, Olivier J Henaff, and Aäron Van den Oord. Divide and contrast: Self-supervised learning from uncurated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10063–10074, 2021. 1

[24] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022. 2

[25] Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. Data collection and quality challenges in deep learning: A data-centric ai perspective. *The VLDB Journal*, 32(4): 791–813, 2023. 2