Enhancing Visual Encoder Self-Supervised Pre-training With OOD Detector

Lake Wang
yxw209@nyu.edu

Ziyue (Tom) Zhou
tz1307@nyu.edu

Patrick Su
zs1512@nyu.edu

Abstract

Self-supervised learning (SSL) often fails to address the challenges posed by long-tailed datasets, where rare but critical objects are overshadowed by frequent classes. We propose a novel framework that integrates Out-of-Distribution (OOD) detection into SSL to dynamically identify and prioritize rare samples using a flexible distance metric. These samples are incorporated into training via a memory buffer, ensuring balanced representation learning. Evaluated on an autonomous driving dataset, our approach significantly improves feature quality for rare objects and downstream task performance, outperforming standard SSL methods. Ablation studies will be conducted in future to demonstrate the impact of key components, such as memory buffer size and OOD thresholds, and showcase the potential of adaptive SSL pipelines for real-world imbalanced datasets.

1 Introduction

Self-supervised learning (SSL) has emerged as a powerful paradigm for visual representation learning, particularly in domains where labeled data is scarce. However, traditional SSL methods struggle in scenarios with long-tailed data distributions, such as autonomous driving datasets. These datasets are dominated by frequent but less critical objects (e.g., roads, skies) while rare objects (e.g., traffic signs, pedestrians) receive limited attention, leading to suboptimal representations for safety-critical tasks.

To address this, we propose a novel pipeline that integrates out-of-distribution (OOD) detection into SSL to prioritize rare and critical objects during training. Our approach uses a distance-based OOD detector to identify rare samples, which are dynamically incorporated into training via a memory buffer. This ensures that the model consistently learns from underrepresented samples without disrupting computational efficiency. We demonstrate the effectiveness of our method on a driving recorder dataset, showing improved feature representations for rare objects and enhanced downstream task performance.

Through this work, we aim to bridge the gap between SSL and robust feature learning in long-tailed distributions, paving the way for more reliable and efficient vision systems in safety-critical domains such as autonomous driving.Our contributions are as follows:

- **Integration of OOD Detection with SSL**: We present a novel pipeline that incorporates OOD detection to identify and prioritize rare samples in the training process.
- **Dynamic Memory Buffer for Long-Tailed Distributions**: We design a memory buffer mechanism that adaptively integrates rare samples into the training pipeline while maintaining computational efficiency.

This study utilizes the BDD10K dataset[7], a large-scale and naturalistic benchmark for autonomous driving research. The dataset comprises video clips captured from diverse driving scenarios, spanning various times of day, weather conditions, and geographic locations. Figure1 displays the flow chart of our proposed framework. For this research, individual frames such as Figure2 are extracted from

Preprint. Under review.

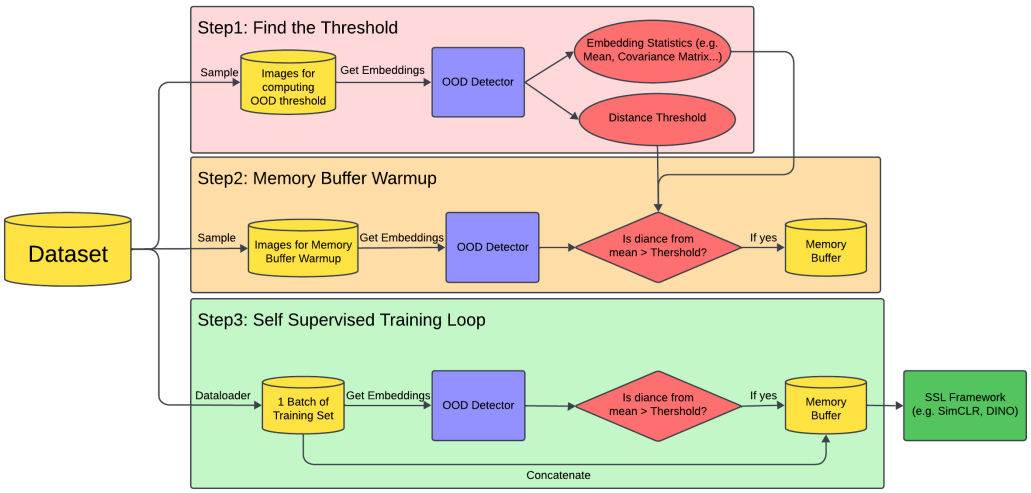

Figure 1: OOD Detection and SSL Training Pipeline. (1) Threshold Calculation: Embedding statistics determine the OOD distance threshold. (2) Memory Buffer Warm-up: Rare samples exceeding the threshold are stored in a FIFO buffer. (3) SSL Training: Batches are processed, rare samples are added to the buffer, and the combined data is used for SSL training (e.g., DINO).

the videos and used as input data. Figure3 displays the long-tail distribution of objects, accurately reflecting the natural frequency of occurrences in real-world driving environments, making the dataset ideal for addressing challenges in data imbalance and object detection in complex urban settings.

2 Related Work

OOD detection has been a significant focus in machine learning for identifying rare or unknown instances in datasets. Sehwag et al. [5] introduced framework that leverages SSL to detect outliers in unlabeled data, improving anomaly detection in computer vision tasks. Wong et al. [6] proposed a method emphasizing the critical role of OOD detection in safety-critical real-time systems. While these works focus on detection, our approach extends this research by directly integrating OOD detection into the SSL pipeline to prioritize rare instances during training.

SSL methods such as SimCLR [3] and DINO [2] have achieved remarkable success in learning feature representations without labels. However, these methods often overlook the challenges of imbalanced datasets, where rare classes are underrepresented. Cao et al. [1] addressed such challenges with a label-distribution-aware margin loss to improve representation learning in long-tailed datasets. Similarly, Kang et al. [4] proposed decoupling representation learning from classifier training for long-tailed recognition.

Inspired by these approaches, our method incorporates a dynamic memory buffer that prioritizes rare samples detected via OOD mechanisms during SSL training, ensuring balanced representation learning in imbalanced and long-tailed datasets.

3 Methodology

3.1 Out-of-Distribution Detection

Our OOD detection module leverages a pretrained visual encoder (e.g., ResNet with DINO weights) to extract robust feature embeddings from images. To identify rare samples, we randomly sample a subset of the dataset and compute a mean feature vector that represents the typical data distribution. A distance threshold is then established, often using the 95th percentile of distances from the mean. The distance metric used for this computation is modular, allowing for flexibility; while Euclidean distance works effectively in our experiments, the framework supports alternatives such as cosine

Figure 2: Original Training Image

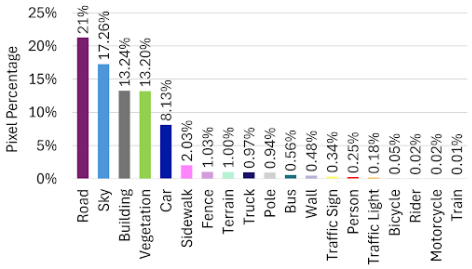

Figure 3: Long Tail Distribution

similarity or other domain-specific metrics. This flexibility ensures adaptability across datasets with varying characteristics.

In the training loop, the OOD module processes batch features in real-time and compares them against the precomputed mean. Samples that exceed the threshold are flagged as rare and considered for inclusion in the memory buffer. By identifying these rare samples dynamically, the OOD module ensures that critical but underrepresented instances are emphasized during the self-supervised learning process. This integration of OOD detection not only enhances the model's focus on challenging instances but also ensures the framework remains lightweight and computationally efficient.

3.2 Detection Algorithm Proof of Concept

To validate the proposed Out-of-Distribution (OOD) detection module, we implemented a pipeline leveraging embeddings extracted from pretrained models such as ResNet-50 with DINO weights. Below are the key steps and observations from the proof-of-concept implementation:

1. **Embedding Extraction**: Cropped images from the dataset were mapped to a high-dimensional embedding space using a pretrained encoder. To establish a baseline for "in-distribution" samples, we randomly selected 5% of the dataset and computed the centroid of their embeddings. This centroid served as a reference point for identifying OOD samples.
2. **Distance-Based OOD Detection**: A flexible distance metric, such as Euclidean distance, was used to calculate the distance of each sample to the centroid. Samples with distances exceeding a dynamic threshold (e.g., the 95th percentile of baseline distances) were flagged as OOD. This approach effectively highlighted rare samples that deviated from the dominant in-distribution patterns.
3. **Cluster Visualization for Validation**: Dimensionality reduction techniques, such as t-SNE, were employed to examine the clustering behavior of embeddings. As shown in Figure 5, common classes (e.g., roads, vehicles) formed distinct, tight clusters, reflecting their prevalence in the dataset. This clustering behavior validated the suitability of embedding-based methods for detecting OOD samples.
4. **Testing with Alternative Distance Metrics**: Experiments with various distance metrics demonstrated that Euclidean distance was particularly effective for distinguishing rare samples. While Mahalanobis distance accounted for data covariance, its numerical instability in high-dimensional spaces limited its practicality. The framework's modularity supports adapting to other metrics as needed.
5. **Integration for Training**: The detection mechanism was integrated into the self-supervised training loop. For each batch, embeddings were compared to the precomputed centroid in real time. Rare samples flagged by the detection algorithm were stored in the memory buffer and reintroduced into subsequent training iterations. This ensured consistent exposure to underrepresented samples.

3.3 Memory Buffer

Building on the OOD detection pipeline introduced earlier, rare samples identified as outliers are systematically managed through a dedicated memory buffer. This buffer plays a pivotal role in

ensuring consistent exposure to underrepresented instances during training, bridging the gap between identification and effective utilization of rare samples.

The memory buffer operates as a First-In-First-Out (FIFO) queue with a fixed size, striking a balance between computational efficiency and representational diversity. As new rare samples are detected during each training iteration, they are added to the buffer, replacing the oldest samples to maintain its fixed capacity. This dynamic update mechanism ensures that the buffer reflects the most current dataset characteristics without overwhelming memory resources.

During training, the buffer's contents are concatenated with the original batch samples to create a combined dataset. This approach integrates rare samples into the training loop, amplifying their representation while preserving the overall distribution of the dataset. By consistently exposing the model to rare and safety-critical instances, the memory buffer facilitates robust feature representation learning, particularly for rare or domain-specific objects. This design aligns seamlessly with the OOD detection module by ensuring that flagged samples actively contribute to the learning process rather than being ignored.

3.4 Self-Supervised Training

The self-supervised training phase leverages the combined dataset (original batch and memory buffer samples) to refine feature representation learning. Following the methodology outlined in the proof-of-concept pipeline, the training process integrates the DINO framework, which employs a cross-view consistency loss to maximize agreement across diverse views of the same instance.

Multi-crop augmentation generates global and local views, which are processed by both the student and teacher networks. The inclusion of rare samples from the memory buffer introduces underrepresented instances into this cross-view consistency learning, encouraging the model to focus on learning features from critical yet sparse examples.

By systematically amplifying the contribution of rare samples through the memory buffer and enforcing cross-view consistency across augmented views, the self-supervised training process ensures improved generalization, especially for long-tailed datasets. This design empowers the model to prioritize critical features while maintaining robustness across diverse distributions, enhancing its utility for downstream tasks.

4 Experiments

4.1 Dataset and Preprocessing

We evaluate our framework on a challenging driving dataset characterized by a long-tailed distribution, where frequent objects such as roads and skies dominate, while critical rare objects like pedestrians, traffic signs, and poles are underrepresented. The dataset is divided into training, validation, and test sets in a 70-15-15 ratio. To prepare the data, we apply augmentations including random cropping, resizing, and normalization, ensuring compatibility with the DINO framework. Each original image is augmented to produce 8 crops: 2 large global crops and 6 smaller local crops. If a global crop includes a rare object, the entire batch of 8 crops is treated as positive pairs and saved to the memory buffer. This approach upweights rare objects during DINO training, mitigating their underrepresentation.

4.2 Experimental Setup

The experiments involve three key components: an OOD detector, a memory buffer, and DINO training. For out-of-distribution (OOD) detection, we use a pretrained ResNet50 model initialized with DINO-ImageNet weights to extract feature embeddings from a randomly sampled 5% subset of the training data. The mean embedding vector is computed, and the Euclidean distance of each embedding to the mean is calculated. A distance threshold (e.g., the 95th percentile) is established to identify samples containing rare objects.

The memory buffer stores up to 32 samples, and training is conducted with a batch size of 1024. For each batch, the OOD detector identifies global crops with distances exceeding the threshold, marking them as rare object samples and updating them to the memory buffer. During each training epoch, the model processes both the standard batch and all rare samples from the memory buffer. This approach

Table 1: Detection Results Evaluation with ResNet-50 architecture and DINO framework

Dataset	Training Set Size	Threshold Percentile	Precision	Recall	F1
ImageNet	256	0.99	**0.2000**	0.0018	0.0035
ImageNet	256	0.95	0.1911	0.0051	0.0099
ImageNet	256	0.90	0.1817	0.0094	0.0178
ImageNet	256	0.80	0.1836	0.0189	0.0343
ImageNet	1024	0.99	0.1842	0.0014	0.0028
ImageNet	1024	0.95	**0.1966**	0.0053	0.0103
ImageNet	1024	0.90	0.1937	0.0094	0.0178
ImageNet	1024	0.80	0.1851	0.0174	0.0319
BDD100K	256	0.99	**0.1946**	0.0043	0.0084
BDD100K	256	0.95	0.1818	0.0061	0.0118
BDD100K	256	0.90	0.1705	0.0102	0.0193
BDD100K	256	0.80	0.1764	0.0190	0.0344

enhances the model's robustness in recognizing rare objects during testing. The DINO model is trained for 100 epochs using a cosine learning rate scheduler. We employ the AdamW optimizer with a learning rate of 5×10^{-4}.

4.3 Results and Discussion

Figure 4: Rare Crop Example

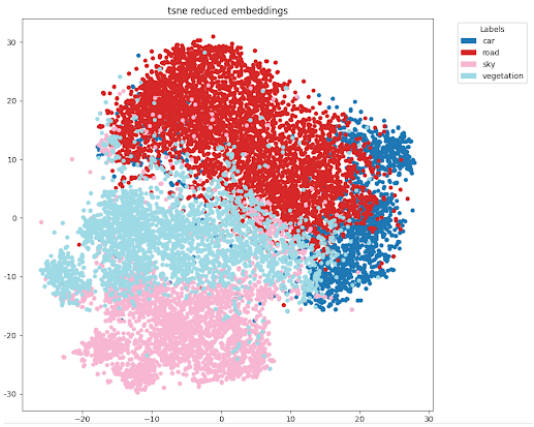

Figure 5: Embedding Cluster of Common Classes

The performance of the proposed framework for rare object detection is evaluated across different training configurations. Ground-truth labels in the corresponding pooled label matrix are used to verify the rarity of the crops. Figure4 shows an example of memory buffer crop.

Table1 shows detection results with different configurations. Our experiments demonstrate that detection performance varies with the threshold percentile choice. Using a training set size of 256 samples on ImageNet, the framework achieves a precision of 0.20 at the 99th percentile threshold, while maintaining lower but more balanced performance at lower thresholds. The BDD100K dataset shows similar trends, with a peak precision of 0.19 at the 99th percentile. Increasing the training set size to 1024 samples for ImageNet yields comparable precision values, suggesting that performance is relatively stable across different training set sizes.

The results demonstrate that the framework captures a small but consistent portion of rare objects, with recall and F1 scores varying based on the threshold setting. These variations align with the expected trade-off between precision and recall, which is inherent in the challenging task of rare object detection. For the downstream task of enhanced training with rare samples, precision is the most critical metric as it directly impacts the quality of the memory buffer. Therefore, our efforts prioritize optimizing precision to ensure the reliability and relevance of the selected samples.

Comparing the datasets, models using ImageNet weights consistently outperform those using BDD100K weights in precision. This is likely due to the broader diversity and focused representation of ImageNet pretraining, whereas BDD100K, as a challenging driving dataset, introduces complexities specific to its domain that may impact detection performance.

Due to time constraints, a full evaluation of retrained models with enhanced training detection strategy was not conducted, leaving future opportunities to explore enhanced training strategies and further assess their impact.

5 Conclusion

In this work, we addressed the challenge of learning robust feature representations from long-tailed datasets, particularly in the domain of autonomous driving. By integrating an Out-of-Distribution (OOD) detection module with a self-supervised learning (SSL) framework, we demonstrated the ability to dynamically identify and prioritize rare samples during training. The proposed memory buffer mechanism further ensured consistent exposure to these critical samples, resulting in improved representation learning for underrepresented objects.

Although we did not perform a full evaluation of retrained models, this work represents an in-development detection framework aimed at enhancing rare object identification. As we continue to refine the approach, we face significant challenges in improving detection reliability. Achieving more consistent and reliable detection is a prerequisite before we can meaningfully evaluate the framework's impact on downstream tasks. This study lays the groundwork for future advancements in adaptive training pipelines for imbalanced datasets.

5.1 Future Work

While our framework effectively enhances feature learning for rare objects, several directions remain open for exploration:

- **Advanced OOD Detection Techniques**: Future work could explore more sophisticated approaches for feature extraction and rare object detection. For example, attention-based models might better capture context, while dynamic distance thresholds or gradually increasing resampling rates based on distance could further enhance detection accuracy. These techniques may refine the identification and prioritization of rare objects.

- **Online Adaptation**: Currently, the distance threshold is computed beforehand to reflect a meaningful quantile from a representative sample of crops. Incorporating online learning methods could make the threshold dynamic, adapting in real-time to batch-level variations during data acquisition. Such an approach would enable both the OOD detector and memory buffer to be updated continuously, enhancing the system's responsiveness to evolving data distributions.

- **Robustness to Noisy Data**: Addressing the challenge of distinguishing between true rare objects and rare environmental features remains an important avenue. For instance, our current system occasionally misidentifies rare terrains, such as snowy scenes, as rare objects. Developing methods to differentiate between object-based rarity and scene-based rarity would improve detection precision and overall model robustness.

- **Dataset pixel label adjustment**: One limitation of the current dataset is the presence of labeling inaccuracies, where common objects in the vicinity of rare objects are sometimes mislabeled, creating challenges for learning meaningful representations of rare objects. Future work could address this issue by leveraging a more robust network to refine pixel-level labels, ensuring greater accuracy and consistency in the dataset.

By addressing these challenges, we aim to further advance the utility of self-supervised learning frameworks in tackling real-world problems, particularly in safety-critical applications like autonomous driving.

References

[1] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss, 2019.

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

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[4] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition, 2020.

[5] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection, 2021.

[6] Kelvin Wong, Shenlong Wang, Mengye Ren, Ming Liang, and Raquel Urtasun. Identifying unknown instances for autonomous driving, 2019.

[7] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2636–2645, 2020.