2024

Proposing coarse to fine grained prediction and hard negative mining for open set 3D object detection

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Thesis

PROPOSING COARSE TO FINE GRAINED PREDICTION AND HARD NEGATIVE MINING FOR OPEN SET 3D OBJECT DETECTION

by

HARSH KHATRI

B.Tech., Indian Institute of Technology Bombay, 2019

Submitted in partial fulfillment of the

requirements for the degree of

Master of Science

2024

Approved by

First Reader

Bryan A. Plummer, PhD
Assistant Professor of Computer Science

Second Reader

Eshed Ohn-Bar, PhD
Assistant Professor of Electrical and Computer Engineering

Acknowledgments

I am grateful to my advisor, Professor Bryan Plummer, whose unwavering support and guidance have been instrumental throughout my journey in completing this thesis. Since the spring semester of 2023, Professor Plummer has been an invaluable mentor, particularly during my early days as a beginner in Machine Learning. His expertise, patience, and encouragement have been pivotal in shaping my research endeavors and academic growth during my Masters program. I extend my deepest appreciation for his dedication and mentorship.

I also extend my heartfelt thanks to Professor Eshed Ohn-Bar for serving as a reader for my defense, contributing his time and expertise to enhance the quality of my work.

To my dear friends, your constant support and companionship have been a source of strength and inspiration, especially during challenging times. As an international student, your unwavering encouragement has been motivating me to persevere and stay focused.

Lastly, I wish to express my boundless gratitude and love to my parents, whose support and sacrifices have been the cornerstone of my journey. Their encouragement and belief in me have been the driving force behind my achievements, and I am eternally grateful for their support that made this journey possible.

Harsh Khatri

Department of Computer Science

Boston University

PROPOSING COARSE TO FINE GRAINED PREDICTION AND HARD NEGATIVE MINING FOR OPEN SET 3D OBJECT DETECTION

HARSH KHATRI

ABSTRACT

In recent years, there has been a remarkable advancement in robotics, autonomous vehicles, and augmented reality technologies, leading to a surge of interest and research activities in 3D learning. Among the 3D recognition research, a significant portion focuses on closed-set detection, overlooking the inherently open nature of real-world scenarios. Furthermore, the scarcity of large-scale 3D datasets, compounded by the high cost of data collection poses a substantial challenge for researchers in this domain. Motivated by these limitations, our work centers on enhancing 3D object detection within an open-set setting.

Our work addresses two key research questions within the realm of open-set 3D object recognition, which has not been addressed in prior literature. Firstly, we investigate the efficacy of employing a coarse to fine-grained prediction strategy. This approach aims to enhance the performance of visually similar categories while maintaining the performance of other categories within the dataset. Secondly, we explore the utilization of offline hard negative mining, specifically targeting challenging samples within the text and image modalities to align with the point cloud encoder. This methodology leads to robust performance, particularly on syntactically similar categories. Through these approaches, our study contributes to the advancement of open-set object detection in 3D learning, thereby addressing critical gaps in current research efforts.

Contents

List of Tables

List of Figures

List of Abbreviations

AR	.....	Augmented Reality
AV	.....	Autonomous Vehicles
CLIP	.....	Contrastive Language-Image Pretraining
GDANet	.....	Geometry-Disentangled Attention Network
LiDAR	.....	Light Detection and Ranging
ULIP	.....	Unified Representation of Language, Image and Point Cloud
VR	.....	Virtual Reality

Chapter 1

Introduction

Fueled by the increasing demands in practical applications like augmented reality (AR), autonomous driving, and robotics, there has been an upsurge of interest in 3D vision. Among many, one of the crucial 3D task is 3D object detection, recognizing objects within 3D point cloud space. Majority of the research in 3D perception like Qi et al. (2017a,b); Xu et al. (2021); Xiang et al. (2021) are focused on supervised learning. These are state-of-the-art approaches achieving accuracy rates exceeding 90%. However, these methods lack generalization due to the limited availability of data, with datasets like ModelNet40 (Wu et al., 2015), ShapeNet Chang et al., 2015, and ScanObjectNN (Uy et al., 2019) containing fewer than 60 object categories, in contrast to the expansive classes found in 2D image datasets. The dataset scale issue of 2D images has been notably mitigated due to the abundance of image data, as demonstrated in recent work such as CLIP (Radford et al., 2021). Consequently, numerous studies aim to utilize pre-trained 2D image-language models to improve 3D understanding. Some of the multi modal methods to assist 3D tasks, such as 3D object detection (Yan et al., 2022; Xu et al., 2023; Xue et al., 2023; Zhang et al., 2022), 3D generation (Hong et al., 2022; Jain et al., 2022; Lee and Chang, 2022; Canfes et al., 2023) and 3D scene-level segmentation (Xie et al., 2020; Liu et al., 2023; Jatavallabhula et al., 2023; Ha and Song, 2022; Ding et al., 2023). These multi-modal approaches predominantly operate within a supervised framework, presenting limitations in generalization, particularly in real-world scenarios.

In response to these limitations, investigating open-set detection becomes imperative for enhancing generalization capabilities. Open-set approaches like Zhu et al. (2023) enrich 3D scene understanding by incorporating objects from diverse sources into 3D scenes, while Lu et al. (2023) localizes 3D objects using pre-trained 2D detectors, employing triplet cross-modal contrastive learning to connect image, point cloud, and text modalities. Notably, OpenShape (Liu et al., 2024) stands out as the state-of-the-art 3D object detection method in open-set settings, achieving an accuracy of 85% on ModelNet40 and 44% on the challenging Objavese-LVIS benchmark.

Despite its impressive results, OpenShape exhibits room for improvement, particularly in performance on the Objavese-LVIS benchmark, which comprises numerous semantically similar categories, such as shopping bags and shoulder bags. Our work addresses this gap in the research by proposing two novel methodologies: coarse to fine-grained recognition and offline hard negative mining. The coarse to fine recognition method introduces a hierarchical prediction approach, emphasizing fine-grained recognition of visually similar classes. Our second method leverages hard negative examples from text and image modalities, this method enhances the alignment of point cloud detectors, mitigating confusion between similar categories.

Our code is available at `https://github.com/harsh242bu/OpenShape3D.git`

Chapter 2

Related Work

2.1 Supervised object detection

In the realm of 3D object detection and point cloud analysis, several architectures have emerged, each addressing distinct challenges and introducing novel approaches to enhance point cloud understanding. PointNet Qi et al. (2017a), a pioneer method, revolutionized the field by directly processing unordered 3D point clouds without the need for voluminous transformations into regular grids, voxels or set of images. However, its limitation in recognizing fine-grained patterns led to the development of PointNet++ Qi et al. (2017b). This hierarchical extension recursively applies PointNet on nested partitions, enabling the model to learn local features with increasing contextual scales, thereby improving its performance on challenging benchmarks. Whereas, GDANet Xu et al. (2021), or Geometry-Disentangled Attention Network, tackles the limitations of prior networks by introducing a disentanglement strategy. By dynamically separating point clouds into contour and flat components, GDANet captures distinct geometric information. Its unique approach of attentively fusing these components refines holistic 3D geometric semantics, outperforming existing methods with fewer parameters. CurveNet Xiang et al. (2021), although motivated to extract gemetric features similar to GDANet, introduces a distinctive method by aggregating continuous sequences of points, termed curves. By grouping and augmenting pointwise features through guided walks in point clouds, CurveNet achieves state-of-the-art results on various tasks, including object classification, normal estimation, and object

part segmentation.

Comparatively, PointNet and PointNet++ prioritize direct processing of point clouds, with the latter addressing limitations through a recursive hierarchical approach. GDANet focuses on disentangling geometric components, refining semantic understanding, while CurveNet utilizes continuous sequence of points through guided walks in the point cloud for improved feature aggregation and geometric representation. While these methods demonstrate impressive performance within supervised settings on specific datasets, their ability to generalize beyond the confines of their data distribution remains limited. Despite advancements in architecture for 3D point cloud understanding, the efficacy of these methods has largely plateaued since they rely solely on the information from the point cloud. Consequently, our research shifts its focus towards open-set detection, addressing the pressing need for improved generalization capabilities in the field.

2.2 Multi-modal detection

Multi-modal methodologies in 3D object detection have attracted considerable interest owing to their ability to exploit diverse data sources, leading to enhanced accuracy and robustness. Notably, these approaches demonstrate superior generalization compared to traditional supervised methods. By integrating various modalities such as point clouds, images, text data, and even audio, these methodologies aim to elevate object detection performance. PointCMT (Point Cloud Cross-Modal Training) Yan et al. (2022), is one such method which leverages both point cloud and image data. The fundamental concept is to harness additional information such as texture, color, and shading from object images, thereby facilitating more precise shape discrimination. Employing a teacher-student framework, PointCMT involves a sophisticated model (the teacher) trained on both modalities guiding a simpler model (the student)

solely utilizing point clouds. This process is formulated as a knowledge distillation task, with the objective of transferring expertise from the teacher model to the student, thus enhancing object detection capabilities.

Another avenue of research focuses on integrating point cloud data with textual information. The paper PointLLM Xu et al. (2023) explores this approach by empowering large language models to comprehend point clouds through textual descriptions. By encoding textual descriptions alongside point cloud features, the model gains semantic understanding. Consequently, the system becomes adept at grasping intricate details even in scenarios where parts of objects are obscured from view, thereby mitigating the impact of occlusions. Furthermore, this methodology facilitates user interaction through natural language, enabling seamless querying and analysis of 3D shapes.

Advancing further, some methodologies incorporate a combination of point cloud, text, and image data for 3D object detection. Two noteworthy methods in this domain are ULIP Xue et al. (2023) and PointCLIP Zhang et al. (2022), both of which leverage CLIP (Contrastive Language-Image Pretraining). CLIP Radford et al. (2021), renowned for its pioneering work in aligning visual and textual features, is pre-trained on extensive image-text pairs, thus establishing a common space for visual and textual data. Building upon this foundation, PointCLIP methodically analyzes point clouds by projecting them into multiple depth maps, effectively translating them into formats comprehensible to CLIP. This approach not only enhances the interpretability of point cloud data but also facilitates zero-shot learning for point cloud classification.

In contrast, ULIP adopts a slightly different strategy, utilizing the CLIP model for initial representation learning. However, ULIP diverges by necessitating a dataset comprising triplets of image, text description, and corresponding point cloud for fine-

tuning the 3D representation. While PointCLIP offers a simpler and faster approach, ULIP provides a more powerful and nuanced understanding of 3D objects. Nonetheless, ULIP's efficacy is contingent upon the availability of substantial training data and may entail greater computational costs. Thus, the choice between these methodologies hinges on the specific requirements and constraints of the application at hand.

2.3 Open-set detection

Numerous multimodal methods, such as Point-BERT Yu et al. (2022), operate within a supervised framework. As discussed previously, supervised approaches are susceptible to limitations in generalization, performing optimally only within the confines of their training data distribution. Thus, investigating open-set detection becomes imperative for enhancing generalization capabilities. Unlike in the realm of 2D data, the creation of 3D datasets involves costly setups, resulting in a scarcity of available 3D data. Consequently, the exploration of open-set research becomes more significant in this context.

The paper 3DOS Alliegro et al. (2022) introduces the concept of semantic novelty detection in 3D data and presents the 3DOS benchmark, providing a framework for evaluating algorithms' ability to handle novel signals in 3D point clouds. By comparing advancements in 3D data analysis with existing 2D open set literature, the paper underscores the importance of developing models capable of discriminating known classes while rejecting unknown categories. It emphasizes the need for effective representation learning approaches to mitigate limitations such as miscalibration and over-confidence in existing deep models.

While 3DOS works on evaluation framework, OpenIns3D and Object2Scene works on scene understanding in open vocabulary setting. OpenIns3D Huang et al. (2023) focuses on generating 3D mask proposals and mask scores using the Mask Proposal

Module. The Snap module renders synthetic scene-level images, which are then processed by a 2D open-world detector along with text queries. The Lookup module refines mask proposals and unlocks their semantic meaning through Mask2Pixel Guided Lookup and Local Enforced Lookup. Challenges addressed by OpenIns3D include low quality of 3D instances, lack of context information, and domain gap between projected and natural images. The framework incorporates Mask Scoring to eliminate low-quality masks and achieves significant improvements in 3D open-vocabulary instance segmentation results. Object2Scene Zhu et al. (2023) presents L3Det, a novel framework for open-vocabulary 3D object detection that enriches existing 3D scene datasets using large-scale 3D object datasets. By unifying 3D detection and visual grounding, L3Det aims to bridge the domain gap between inserted objects and original scenes. Key components of the approach include object normalization, data augmentation, and cross-domain category-level contrastive learning, which enhance detection performance.

The next two methods OV-3DET and OpenShape work specifically on 3D point cloud object detection in Open-set setting. OV-3DET Lu et al. (2023) proposes a method for Open-Vocabulary 3D Point-Cloud Object Detection without 3D Annotation. This approach employs a dividing-and-conquering strategy, utilizing pre-trained models and de-biased triplet cross-modal contrastive learning. The method achieves superior performance on datasets like ScanNet and SUN RGB-D, highlighting the potential of leveraging pre-trained models and cross-modal learning in open set detection. OpenShape Liu et al. (2024) on the other hand tackles the challenge of recognizing and understanding 3D shapes, even those entirely new (unseen during training). Building upon PointBERT, it aligns 3D shape embeddings with CLIP's text and image spaces, enabling the utilization of textual descriptions and image recognition for shape understanding. It utilizes textual descriptions from the Obja-

verse dataset and BLIP-generated captions for 2D thumbnails, processed by CLIP to obtain text features. OpenShape excels in open-world recognition, achieving notably higher accuracy on a wide benchmark. Its learned embeddings capture diverse concepts, supporting tasks like text-to-3D interactions. Integration with CLIP facilitates tasks such as point cloud captioning.

Despite of this remarkable performance of OpenShape, there's still a lot of room for improvement. Majority of the work in open set object detection revolves around the idea of aligning point cloud with text and image modalities, distilling knowledge from pre-trained image and language models. But this approach works great as a starting foundation and the work so far lacks in accurately distinguishing confusing categories. Our work addresses this research gap by implementing coarse to fine grained prediction. Where we are aiming to bridge the gap in performance by improving on the fine grained predictions. Our second contribution is leveraging hard negative mining for text and image modality. This approach provides better distinction between the categories in the text and image modality.

Despite the remarkable performance achieved by OpenShape, there remains ample opportunity for advancement. Predominantly, existing efforts in open-set object detection have centered around aligning point cloud data with text and image modalities, often relying on the fusion of knowledge extracted from pre-trained models in these domains. While this approach serves as a fundamental starting point, it falls short in accurately distinguishing between similar categories, presenting a notable research gap. Our work introduces two approachs: coarse-to-fine-grained prediction and offline hard negative mining. Coarse to fine prediction aims to improve performance by refining the granularity of predictions, thereby addressing the challenge of accurately discerning nuanced distinctions between object categories. Secondly, our research contributes by incorporating hard negative mining technique across both text

and image modalities. By leveraging this approach, we enhance the differentiation between categories within each modality.

Chapter 3

Coarse to Fine Grained Recognition

In this chapter, we delve into the concept of coarse-to-fine-grained recognition, a hierarchical approach within computer vision aimed at object recognition across multiple levels of detail. Initially, this method identifies broad categories before progressively refining categorizations to finer levels of detail, thereby enhancing recognition accuracy. For instance, Pavlakos et al. (2017) tackles the challenge of 3D human pose estimation through a two-step solution. Initially, it performs fine discretization of the 3D space around the subject to predict the likelihood of each joint per voxel. Subsequently, it iteratively refines and processes image features to improve accuracy. Similarly, Zheng et al. (2020) focuses on cross-domain object detection. At the coarse-grained stage, it utilizes attention mechanisms to extract foreground regions. Then, at the fine-grained stage, it aligns conditional distributions of foregrounds by minimizing the distance between global prototypes from the same category but across different domains. Furthermore, Liu and Chen (2023) proposes a cluster-based coarse-to-fine object detection framework tailored for high-resolution images. This method enhances detection accuracy for small objects while maintaining precision for large objects, ultimately reducing computational costs associated with high-resolution imagery.

Inspired by these advancements, our work implements coarse-to-fine recognition with the aim of distinguishing visually similar categories more effectively. By leveraging this approach, we endeavor to contribute to the refinement and precision of

Table 3.1: Top five confused categories and their confusion pairs predicted by baseline model PointBERT (Liu et al., 2024).

Worst category	Confused category
vodka	bottle(17), wine_bottle(14), beer_bottle(6)
bridal_gown	dress(17), robe(3), skirt(2)
boiled_egg	egg(17), egg_yolk(2)
cabinet	armoire(17), bookcase(9), dresser(7)
sherbert	icecream(17), popsicle(2)

object recognition tasks within our domain

3.1 Method

The purpose of coarse to fine-grained recognition is integral in the field of computer vision, aiming to distinguish between classes that share striking similarities, thereby enhancing the precision and granularity of object classification systems. Through a comprehensive analysis of the performance of the OpenShape model, it becomes apparent that a significant portion of misclassifications originates from categories that exhibit visual and semantic resemblance. This observation underscores the critical need for robust fine-grained recognition methodologies. Table 3.1 presents a detailed breakdown of the top five confused categories predicted by the model, offering insights into the specific challenges encountered. Upon meticulous scrutiny of these confusion pairs, a discernible pattern emerges, shedding light on the underlying reasons behind the model's struggles with classification. It becomes evident that the primary source of confusion lies in the striking similarity between certain categories, evident in both their point cloud representations and language embeddings.

The convergence of similarity in both visual and semantic spaces poses a formidable obstacle for accurate classification, as exemplified in Figure 3·2, where even subtle differences in point clouds prove elusive for human observers. This observation underscores the intricacies involved in fine-grained categorization tasks, where the distinc-

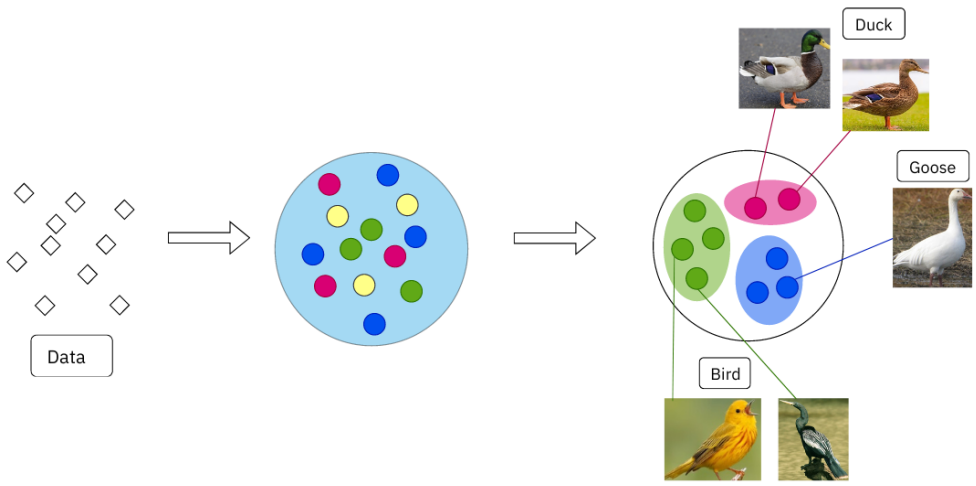

Figure 3·1: Coarse to fine grained recognition. Beginning with a broad categorization, such as birds, we then delve into fine-grained recognition within the identified cluster, discerning between various bird species.

tion between visually and semantically similar categories becomes increasingly challenging. Moreover, our analysis reveals that while certain categories exhibit notable similarities in both point cloud and language spaces, they remain distinguishable. Notable examples include cabinet-armoire, sherbert-icecream, and bridal gown-dress. These findings highlight the nuanced nature of fine-grained recognition, where the ability to discern subtle differences between categories is paramount for accurate classification.

In light of these challenges, our work endeavors to address the inherent complexities of fine-grained categorization. By acknowledging and confronting these obstacles head-on, our subsequent experiments in fine-grained recognition aim to refine the model's capabilities in discerning intricate differences between visually and semantically similar categories.

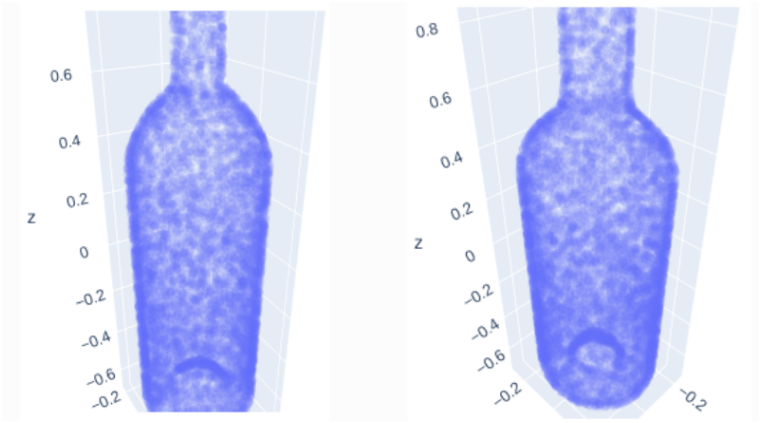

Figure 3·2: Plot of point cloud data of 2 categories.
Left: Vodka category **Right**: Wine bottle category

3.2 Experiments

The objective of this experiment is to explore the efficacy of coarse to fine-grained recognition in separating similar categories. To achieve this, we designed an experiment wherein we selectively focused on a subset of closely related categories and refined the model's performance specifically on these categories. As discussed in the last section, we identified and analyzed the top confusing categories. Subsequently, we visualized the point cloud data corresponding to a subset of these categories, from which we manually identified 10 distinct categories discernible within the point cloud space. This visual inspection confirmed that these categories could indeed be manually distinguished based on their point cloud representations. Following this visual assessment, we partitioned samples from these 10 categories into training and testing datasets. Subsequently, we proceeded to fine-tune the final linear layer of our model exclusively on these shortlisted categories, augmenting the training process with an additional 20 epochs.

The performance of the fine-tuned model exhibits a slight decrement compared to the baseline, indicating an unexpected outcome. This discrepancy prompts a

thorough investigation into the underlying reasons for the suboptimal performance of the coarse-to-fine recognition approach. Upon scrutinizing the results, it becomes apparent that while the confusion decreases for a subset of 10 handpicked categories, there is a simultaneous decrease in correct predictions. This observation suggests a shifting performance trend of the model, where its efficacy appears to fluctuate between different sets of categories.

In response to this phenomenon, we explore ensemble methods to potentially mitigate the shortcomings of individual models. Initially, we implement a naive ensemble technique, which involves averaging the predictions from both the baseline and fine-tuned models. However, this ensemble fails to surpass the performance of the baseline model, albeit outperforming the fine-tuned model in isolation. Subsequently, we employ a weighted ensemble strategy, assigning different weights to the predictions of each model based on their relative performance. Remarkably, this weighted ensemble demonstrates superior performance compared to the baseline model, indicating its efficacy in leveraging the strengths of both models.

The improved performance of the weighted ensemble lends further credence to our observation that the fine-tuned model's performance fluctuates between different categories. By leveraging the complementary strengths of both models, the ensemble approach effectively mitigates the shortcomings of individual models, resulting in enhanced overall performance. This finding underscores the importance of ensemble techniques in addressing the inherent complexities and uncertainties associated with fine-grained recognition tasks.

In summary, our experiment underscores the potential of fine-grained recognition techniques in mitigating category confusion. However, it also highlights the inherent trade-offs between specificity and generalization in model performance, emphasizing the need for nuanced strategies to balance these competing objectives effectively.

Table 3.2: Model Performance: The table presents the results of a PointBERT model fine-tuned on a subset of classes alongside an ensemble model composed of the baseline and the fine-tuned model. The performance was evaluated on Objaverse-LVIS benchmark

Model	Accuracy	Class accuracy
ULIP-PointBERT (Liu et al., 2024)	26.80	24.40
SparseConv (Liu et al., 2024)	43.40	40.86
PointBERT (Liu et al., 2024)	44.44	42.19
Fine tuned (Ours)	43.70	41.14
Standard Ensemble (Ours)	44.08	41.40
Weighted Ensemble (Ours)	45.19	43.02

3.3 Results

In this chapter, we delved into coarse to fine-grained recognition for 3D object detection using the Objaverse dataset. As evidenced by the results presented in Table 3.2, the overall performance of the fine-tuned model degrade over the course of 10 epochs. However, upon ensembling both the base and fine-tuned models, a slight improvement in performance was observed. The improved performance of the ensemble model provides additional validation to the notion that fine-tuning tends to shift the model's accuracy towards specific confused categories. This suggests that more ensembling techniques like mixture of experts can be explored to improve the distinction capabilities of the model.

In conclusion fine-grained recognition yields limited performance improvement. While there were improvements observed in the performance of the few categories, the overall model performance diminished. This observation indicates that distinguishing between similar categories proves challenging.

3.4 Conclusion

In summary, our exploration delved into the realm of coarse to fine-grained recognition, aimed at effectively delineating between closely related classes within the

dataset. In this experiment, we endeavored to fine-tune a model over a limited number of epochs focusing on a subset of categories. Despite our efforts, the performance of the fine-tuned model failed to surpass that of the baseline. However, upon implementing a weighted ensemble approach combining the fine-tuned model with the baseline model, we observed notable enhancements in performance.

Thus far, our analysis has revealed the model's struggles in accurately categorizing similar classes. These challenges appear to stem from both geometric and semantic similarities inherent within the dataset. To address this issue comprehensively, we propose the utilization of offline hard negative mining to discern and isolate semantically similar categories across both text and image embedding spaces. The primary objective of our forthcoming experiment is to identify and address categories that exhibit confusion within these embedding spaces, thereby augmenting the model's capability to accurately classify similar classes.

Chapter 4

Offline Hard Negative Mining

The concept of hard negative mining was initially introduced by Sung (1996) for training face detection models. It has been a fundamental technique in machine learning for over two decades. The key idea involves prioritizing or specifically training on negative examples that are misclassified as positives by the model. Among these negative examples, hard negatives are particularly challenging as they closely resemble misclassified examples, hence the term "hard". The idea of hard negative mining appears in the training of many successful object detectors such as Malisiewicz et al. (2011); Gidaris and Komodakis (2015); Shrivastava et al. (2016); Felzenszwalb et al. (2009).

In this chapter, we delve into the application of offline hard negative mining and its impact on model performance. Given the prevalence of confusion among similar-looking categories in our work, the exploration of hard negative mining emerges as a promising avenue. Coarse to fine-grained recognition, as demonstrated in the preceding chapter, exhibited limited efficacy in distinguishing between these visually similar categories. Hard negative mining, by contrast, offers a targeted approach tailored to addressing such challenges.

4.1 Method

In order to implement offline hard negative mining, we first construct a comprehensive text and image feature matrix by traversing the Objaverse test dataset. Subsequently,

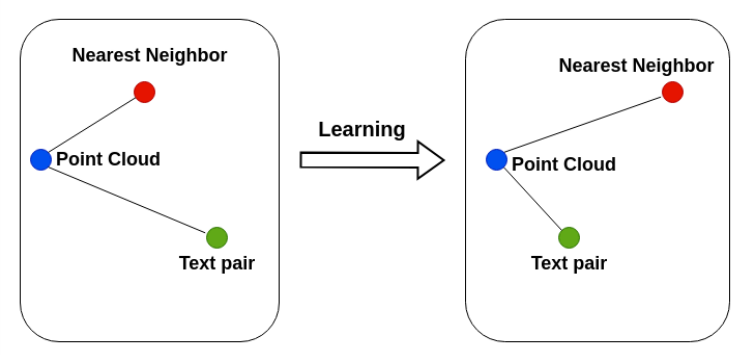

Figure 4·1: Image depicting learning using triplet loss for text features, where triplets are created using nearest neighbor (from offline feature matrix) as negative pairs and text embedding from training batch as positive pair

we establish a nearest neighbor classifier based on the generated feature matrix. During the training phase, the nearest neighbor of each data point is computed for both text and image features. This nearest neighbor then serves as a negative pair for subsequent triplet creation. To elaborate, the training process involves the formation of triplets comprising a positive pair obtained from the dataloader and a negative pair sourced from the nearest neighbor. This construction facilitates the calculation of triplet loss, leveraging the relative distances between positive and negative pairs. By incorporating this approach, we aim to enhance the model's ability to discern between similar categories, thereby addressing the challenges encountered with coarse to fine-grained recognition.

Our dataset comprises a total of 1156 categories, encompassing 875,665 corresponding samples. Notably, the Objaverse training dataset lacks hard labels for individual data points. Instead, each sample is contains point cloud data along with corresponding text features, and image features. The model aims to align point cloud features with text and image features. Given the extensive breadth of categories

under consideration, it is assumed during training that all samples within a batch possess unique category labels—a presumption largely substantiated by the model's favorable performance.

However, this inherent assumption precludes the feasibility of employing online hard negative mining, which necessitates labeling the entire dataset. Consequently, we proceed with an offline approach. Offline hard negative mining entails traversing through the test dataset to construct image feature and text feature matrices, which serve as the foundation for identifying and calculating hard negative examples during training.

To effectively distinguish between similar categories, we employ triplet loss as a pivotal component of our approach. Triplet loss necessitates the generation of triplets comprising a positive pair and a negative pair. However, given the absence of hard labels in our training data, our training set only provides positive pairs for both text and image modalities. To address this challenge, we leverage an offline feature matrix derived from the test dataset to obtain negative pairs. Specifically, we employ a nearest neighbor classifier trained on the feature matrix to identify the nearest neighbor for both text and image features corresponding to each sample in the training data. These nearest neighbors serve as our negative pairs.

Subsequently, utilizing these negative and positive pairs, we compute triplet loss, which captures the relative distances between positive and negative pairs within each triplet. This triplet loss is then incorporated into our training process, augmenting the existing cross-entropy loss.

4.2 Experiments

The methodology outlined in OpenShape Liu et al. (2024) involves sampling 10,000 points from the mesh surface for each category and interpolating point colors based

on mesh textures. Additionally, they generate 12 color images from preset camera poses, uniformly covering the entire shape. For datasets providing thumbnails, they include them as part of image candidates, since they typically capture the shape from a better camera view. During training, one of the 12 images from preset camera poses is selected, or the thumbnail image feature matrix is utilized with a 50-50 probability. We adopt a similar approach to create an offline feature matrix for images. As previously described, positive pairs are sourced from the dataloader, and a nearest neighbor classifier is employed to obtain negative pairs. Triplet loss is then computed using these triplets of positive and negative pairs for both text and image features, with the final triplet loss being the sum of text and image triplet losses.

Our training objective is to minimize the total loss, which is a weighted sum of contrastive loss and triplet loss. Initially, triplet loss is calculated for both text and image features. Given the point cloud, text feature, and image feature of a sample, the negative and positive pair logits are computed for both text and image features. These logits are then used to calculate text and image triplet loss.

$$|f(A) - f(P)|^2 \leq |f(A) - f(N)|^2 \tag{4.1}$$

$$|f(A) - f(P)|^2 - |f(A) - f(N)|^2 \leq 0 \tag{4.2}$$

$$d(a_i, t_p) - d(a_i, t_n) \leq 0 \tag{4.3}$$

$$d(a_i, t_p) - d(a_i, t_n) + \alpha \leq 0 \tag{4.4}$$

$$L(A, T) = \sum_{i,p,n} max[d(a_i, t_p) - d(a_i, t_n) + \alpha, 0] \tag{4.5}$$

In this experiment setup, we finetune the last linear layer of the PointBERT model with triplet loss. The dataset consists of a total of 875,665 samples spanning across 1,156 categories. Given the considerable time required to train the entire dataset, a single epoch necessitating 2.6 hours, we opted to train on a subset comprising 1/4 of

Table 4.1: Comparison of Model Performance: The table showcases performance of a model evaluated on Objaverse-LVIS benchmark and trained using triplet loss. Where triplets are created by picking first nearest neighbor, second nearest neighbor and random neighbors from offline text and image feature matrix.

Model	Accuracy	Class accuracy
PointBERT (Liu et al., 2024)	46.84	34.00
Random neighbor (ours)	46.60	33.90
Second neighbor (ours)	46.07	34.02
First neighbor (ours)	46.10	34.46

the training data. Consequently, the model was finetuned on 218,916 samples, with each epoch taking approximately 50 minutes. Subsequently, we trained our updated architecture with triplet loss for an additional 30 epochs. The training process was conducted on an Nvidia A40 GPU equipped with 48 GB of GPU memory.

The triplet loss function, as represented by Eq 4.5, comprises two essential parameters: the margin and the weight. Consequently, hyperparameter tuning was conducted to determine the optimal values for these parameters, aiming to enhance the model's performance.

4.3 Results

In this experiment, we conducted offline hard negative mining on the Objaverse dataset with the aim of distinguishing similar categories and challenging the assumption that each batch during training contains samples from unique categories. Despite our efforts, this approach yielded only marginal improvements over the baseline performance. Although the overall accuracy did not exhibit significant enhancement, there was a slight improvement observed in the class-wise accuracy, also known as balanced accuracy.

These results suggest that the hard negative mining may be more effective in

improving the accuracy of some of the rarer categories within the dataset. However, it remains evident from this experiment, akin to the preceding one, that separating these similar categories presents a formidable challenge. Further exploration and refinement of methodologies may be necessary to achieve more substantial improvements in the classification of such categories.

Chapter 5

Conclusions

In this work we aimed to bridge the current research gap in the landscape of 3D open-set object detection. We propose two key ideas Coarse to fine grained prediction and offline hard negative mining. We evaluate the performance of our approach on the Objaverse-LVIS benchmark.

5.1 Contribution

Through an exhaustive analysis of the OpenShape method, we identified a performance bottleneck arising from the model's inability to distinguish between similar categories. In response, we introduced a coarse to fine prediction approach aimed at effectively separating these confounding categories. However, despite our efforts, the observed performance improvements were marginal, indicating that the existing method, PointBERT, struggles to adequately address this challenge.

Upon analyzing the confusion pairs and performing binary classification, we proposed another approach of offline hard negative mining. Leveraging triplet loss, we effectively isolated nearest neighbors within the text and image modalities, enhancing the model's discriminative capabilities. This approach leads to a slight improvement in performance. Although the overall accuracy has not improved, we see an improvement in class accuracy. This suggests that the triplet loss is performing better on some of the rare categories compared to the baseline.

After thorough analysis of confusion pairs and subsequent binary classification,

we introduce offline hard negative mining. By employing triplet loss, we effectively identify and isolate nearest neighbors within both the text and image modalities. Our experimentation reveals a noteworthy outcome: while the overall accuracy of the model remains relatively unchanged, there is a slight improvement in class accuracy. This improvement suggests that the triplet loss mechanism outperforms the baseline, particularly in distinguishing rare categories.

This observation underscores the efficacy of our approach in tackling challenging classification scenarios, wherein traditional metrics such as overall accuracy may not fully capture the model's enhanced performance. By focusing on improving accuracy within specific classes, we demonstrate the potential of triplet loss in bolstering the robustness and effectiveness of our classification framework.

5.2 Future work

As previously discussed, a notable limitation in this study lies in the absence of hard labels within the training dataset. Future research should prioritize the creation of hard labels, which would facilitate a more in-depth analysis of existing methods, pinpointing areas of improvement, and refining these methodologies accordingly. Moreover, our analysis has unveiled instances of noise within the dataset, characterized by numerous mislabeled samples. By incorporating hard labels, we can explore techniques for learning from noisy data, thereby enhancing the robustness of the model.

Additionally, our experimentation revealed promising results in binary classification utilizing solely point cloud data for a subset of categories. However, this success implies that the other modalities lack sufficient richness to effectively distinguish between similar categories. To address this, augmenting the text and image data represents a viable avenue for future research. Notably, augmenting image data holds particular promise, given the limited availability of 12 images for each category.

Therefore, augmenting image datasets could significantly improve the discriminative capabilities of our models.

5.3 Limitations

In this work, we introduced two key ideas and assessed their efficacy using the Objaverse-LVIS benchmark. However, it is imperative to acknowledge the limitations inherent in our work to ensure a nuanced interpretation of the results.

Firstly, one of the primary limitations lies in the exclusive testing of our method on the Objaverse dataset. While this benchmark provides valuable insights, the performance of our approach may vary across different datasets. Therefore, to obtain a comprehensive understanding of its effectiveness and generalizability, future studies should evaluate our method across multiple datasets representing diverse contexts and scenarios.

Secondly, our approach involves the computation of an offline feature matrix to facilitate the calculation of nearest neighbors. This process, while effective, can be resource-intensive, particularly when dealing with large datasets. As such, the feasibility and scalability of our method may be constrained by computational costs. To address this limitation, one potential avenue for improvement is the exploration of online hard negative mining techniques. However, this approach necessitates the creation of hard labels for the dataset, which is currently unavailable. Thus, further research is warranted to develop strategies for generating these hard labels efficiently and effectively. By acknowledging and addressing these limitations, we can ensure a more comprehensive evaluation of our approach and pave the way for its broader applicability and impact in the field of 3D object detection.

References

Alliegro, A., Cappio Borlino, F., and Tommasi, T. (2022). 3dos: Towards 3d open set learning-benchmarking and understanding semantic novelty detection on point clouds. *Advances in Neural Information Processing Systems*, 35:21228–21240.

Canfes, Z., Atasoy, M. F., Dirik, A., and Yanardag, P. (2023). Text and image guided 3d avatar generation and manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4421–4431.

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al. (2015). Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., and Qi, X. (2023). Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7010–7019.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2009). Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645.

Gidaris, S. and Komodakis, N. (2015). Object detection via a multi-region and semantic segmentation-aware cnn model. In *Proceedings of the IEEE international conference on computer vision*, pages 1134–1142.

Ha, H. and Song, S. (2022). Semantic abstraction: Open-world 3d scene understanding from 2d vision-language models. *arXiv preprint arXiv:2207.11514*.

Hong, F., Zhang, M., Pan, L., Cai, Z., Yang, L., and Liu, Z. (2022). Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*.

Huang, Z., Wu, X., Chen, X., Zhao, H., Zhu, L., and Lasenby, J. (2023). Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation. *arXiv preprint arXiv:2309.00616*.

Jain, A., Mildenhall, B., Barron, J. T., Abbeel, P., and Poole, B. (2022). Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 867–876.

Jatavallabhula, K. M., Kuwajerwala, A., Gu, Q., Omama, M., Chen, T., Maalouf, A., Li, S., Iyer, G., Saryazdi, S., Keetha, N., et al. (2023). Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*.

Lee, H.-H. and Chang, A. X. (2022). Understanding pure clip guidance for voxel grid nerf models. *arXiv preprint arXiv:2209.15172*.

Liu, J. and Chen, J. (2023). A coarse to fine framework for object detection in high resolution image. *arXiv preprint arXiv:2303.01219*.

Liu, M., Shi, R., Kuang, K., Zhu, Y., Li, X., Han, S., Cai, H., Porikli, F., and Su, H. (2024). Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in Neural Information Processing Systems*, 36.

Liu, M., Zhu, Y., Cai, H., Han, S., Ling, Z., Porikli, F., and Su, H. (2023). Partslip: Low-shot part segmentation for 3d point clouds via pretrained image-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21736–21746.

Lu, Y., Xu, C., Wei, X., Xie, X., Tomizuka, M., Keutzer, K., and Zhang, S. (2023). Open-vocabulary point-cloud object detection without 3d annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1190–1199.

Malisiewicz, T., Gupta, A., and Efros, A. A. (2011). Ensemble of exemplar-svms for object detection and beyond. In *2011 International conference on computer vision*, pages 89–96. IEEE.

Pavlakos, G., Zhou, X., Derpanis, K. G., and Daniilidis, K. (2017). Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7025–7034.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. (2017a). Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660.

Qi, C. R., Yi, L., Su, H., and Guibas, L. J. (2017b). Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Shrivastava, A., Gupta, A., and Girshick, R. (2016). Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769.

Sung, K.-K. (1996). Learning and example selection for object and pattern detection.

Uy, M. A., Pham, Q.-H., Hua, B.-S., Nguyen, T., and Yeung, S.-K. (2019). Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1588–1597.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., and Xiao, J. (2015). 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920.

Xiang, T., Zhang, C., Song, Y., Yu, J., and Cai, W. (2021). Walk in the cloud: Learning curves for point clouds shape analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 915–924.

Xie, S., Gu, J., Guo, D., Qi, C. R., Guibas, L., and Litany, O. (2020). Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 574–591. Springer.

Xu, M., Zhang, J., Zhou, Z., Xu, M., Qi, X., and Qiao, Y. (2021). Learning geometry-disentangled representation for complementary understanding of 3d object point cloud. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3056–3064.

Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J., and Lin, D. (2023). Pointllm: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*.

Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J. C., and Savarese, S. (2023). Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1179–1189.

Yan, X., Zhan, H., Zheng, C., Gao, J., Zhang, R., Cui, S., and Li, Z. (2022). Let images give you more: Point cloud cross-modal training for shape analysis. *Advances in Neural Information Processing Systems*, 35:32398–32411.

Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., and Lu, J. (2022). Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19313–19322.

Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., and Li, H. (2022). Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8552–8562.

Zheng, Y., Huang, D., Liu, S., and Wang, Y. (2020). Cross-domain object detection through coarse-to-fine feature adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13766–13775.

Zhu, C., Zhang, W., Wang, T., Liu, X., and Chen, K. (2023). Object2scene: Putting objects in context for open-vocabulary 3d detection. *arXiv preprint arXiv:2309.09456*.

CURRICULUM VITAE