Uncertainty Meets Diversity: A Comprehensive Active Learning Framework for Indoor 3D Object Detection

Jiangyi Wang, Na Zhao*
Singapore University of Technology and Design (SUTD)
wangjiangyi0519@gmail.com, na_zhao@sutd.edu.sg

Abstract

Active learning has emerged as a promising approach to reduce the substantial annotation burden in 3D object detection tasks, spurring several initiatives in outdoor environments. However, its application in indoor environments remains unexplored. Compared to outdoor 3D datasets, indoor datasets face significant challenges, including fewer training samples per class, a greater number of classes, more severe class imbalance, and more diverse scene types and intra-class variances. This paper presents the first study on active learning for indoor 3D object detection, where we propose a novel framework tailored for this task. Our method incorporates two key criteria - uncertainty and diversity - to actively select the most ambiguous and informative unlabeled samples for annotation. The uncertainty criterion accounts for both inaccurate detections and undetected objects, ensuring that the most ambiguous samples are prioritized. Meanwhile, the diversity criterion is formulated as a joint optimization problem that maximizes the diversity of both object class distributions and scene types, using a new Class-aware Adaptive Prototype (CAP) bank. The CAP bank dynamically allocates representative prototypes to each class, helping to capture varying intra-class diversity across different categories. We evaluate our method on SUN RGB-D and ScanNetV2, where it outperforms baselines by a significant margin, achieving over 85% of fully-supervised performance with just 10% of the annotation budget. Our code is available at: https://github.com/JoeWang-0519/CVPR25_UMD.

1. Introduction

3D object detection, which aims to localize and identify objects within a 3D scene, is crucial for various applications, including robotics [10], autonomous driving [20, 58], and augmented/virtual reality [51]. While deep learning techniques have achieved significant success in 3D object detection [31, 41, 48], most existing approaches are data-

*Corresponding Author: na_zhao@sutd.edu.sg

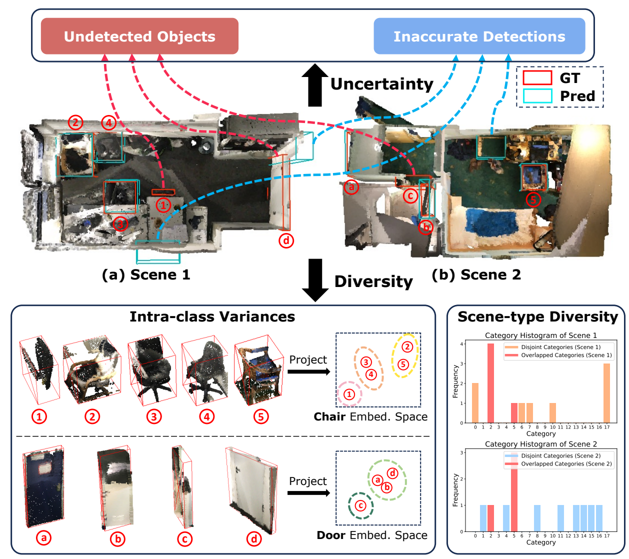

Figure 1. **Challenges in active learning for indoor 3D object detection, including both uncertainty and diversity aspects.** We show two indoor scenes on ScanNetV2, with red boxes for ground truths and cyan for predictions (10% data using CAGroup3D detector), displaying only 'chair' and 'door' categories for clarity. For *uncertainty*, the presence of *undetected objects* and *inaccurate detections* undermines the quality of uncertainty estimation. For *diversity*, high *scene-type diversity* and varying *intra-class variances* in indoor environments complicate diverse sample selection. Our work presents a hybrid approach to address these challenges.

hungry [41], requiring large amounts of annotated data for training. However, annotating objects in a 3D scene is a labor-intensive process. For example, annotating a single scene in the SUN RGB-D dataset takes over **1.9 minutes** per object, with an average of approximately **14 objects** per scene [46]. To mitigate the high costs of 3D annotation, active learning (AL) [30, 34, 38] offers a promising solution. AL seeks to maximize the utility of limited labeling resources by actively selecting the most informative samples for human annotation. Through this process, AL can achieve performance comparable to fully-supervised methods, but with significantly fewer labeled samples.

Although active learning has proven effective in image-based tasks such as image classification and object detection, using uncertainty [5, 16, 37] or/and diversity [34, 38, 60] as selection criteria, its application to 3D object detection remains largely underexplored. To date, only a few studies [12, 24, 25] have applied AL techniques to outdoor 3D object detection. For instance, CRB [25] uses label conciseness to capture uncertainty, along with feature representativeness and geometric balance to enhance diversity for sample selection. KECOR [24] favors samples whose latent features have the highest coding length, thereby ensuring diversity. While these methods demonstrate effectiveness in outdoor scenarios, they are suboptimal for indoor environments, as illustrated in Tab. 2, due to the unique challenges posed by 3D object detection in indoor settings.

On the one hand, a 3D indoor dataset [6] could be **6.2x smaller** in scale but contain **3.3x more** number of categories, with more severe class imbalance, compared to an outdoor dataset [11], as shown in [14]. This scarcity of training samples, coupled with pronounced class imbalance, impairs 3D detection accuracy in indoor scenarios, leading to more frequent occurrences of *inaccurate detection* and *undetected objects*, as illustrated in Fig. 1. In this figure, three detections are inaccurate, incorrectly localizing background objects, while three other objects remain undetected. These challenges underscore the need to account for both inaccurately detected proposals and undetected objects when selecting informative samples from unlabeled data.

On the other hand, unlike outdoor scenes, which primarily feature road environments with a limited set of object classes, indoor environments exhibit greater *scene-type diversity* and *intra-class variances*, caused by diverse layouts and variations in object shapes within each class. For example, as shown in Fig. 1, five 'chair' instances can form three clusters in the embedding space due to the variations in shape and occlusion. This phenomenon in indoor environments complicate the selection of diverse samples.

In this work, we present the *first solution* to apply active learning in indoor 3D object detection, utilizing both uncertainty and diversity as selection criteria. For *uncertainty*, we design a two-pronged epistemic uncertainty estimation strategy to address *inaccurate detections* and *undetected objects*, which first assesses uncertainty separately for each case, then integrates them into a unified criterion. Specifically, we introduce a localization-aware uncertainty score that combines the IoU score – measured between the original detected object and its perturbed version – with the Shannon entropy. This helps mitigate the negative effects of invalid proposals, such as background or partial detections, when estimating the classification uncertainty of detected proposals. To quantify the epistemic uncertainty arising from undetectability, we develop a lightweight network that estimates the number of undetected objects. For *diversity*, we frame the diverse sample selection problem as an optimization task that jointly maximizes *scene-type diversity* and *intra-class variances*, leveraging a novel Class-aware Adaptive Prototype (CAP) bank. This CAP bank is designed to capture varying intra-class variances by generating an adaptive number of representative prototypes per class. Observing that different scene types exhibit distinct histograms of object categories as shown in Fig. 1, each composed of various prototypes, we use the CAP bank to model both class-level and scene-level diversity.

We evaluate the proposed method on SUN RGB-D [46] and ScanNet [6] datasets. Using CAGroup3D [48] as base detection model, our approach achieves 56.11%, 61.77% mAP@0.25 on SUN RGB-D and ScanNet, respectively, reaching 87.24% and 85.02% of the fully-supervised performance with only 10% of the annotation budget.

2. Related Work

Generic Active Learning. The majority of existing works have explored active learning (AL) strategies on image classification task, which can be categorized into uncertainty-based [5, 9, 16, 37, 50] and diversity-based [1, 7, 34, 38, 54, 60] methods based on the selection criteria used. Uncertainty-based approaches select ambiguous samples for annotation by maximizing uncertainty scores, which can be computed in various ways, such as entropy [9, 17, 39], minimum posterior probability [50], or multi-class margin [5, 16, 37]. Shannon entropy [39] is a commonly used metric in classification that quantifies uncertainty in the probability distribution, while multi-class margin [37] measures the uncertainty through the difference between the top two predicted probabilities. MC Dropout [9] employs Monte Carlo Dropout to approximate the posterior predicted probability and use Shannon entropy to assess the underlying uncertainty. Diversity-based methods [1, 38, 60] select a compact subset of samples that best represent the global data distribution. For example, Core-set [38] formulates active learning as a core-set selection problem and solves it using the k-centroid algorithm. Recent works [4, 22, 45] combine both uncertainty and diversity through hybrid approaches. A representative approach, BADGE [4], balances uncertainty and diversity by applying the k-means++ [52] algorithm in the gradient space, which considers both magnitude and direction of gradient vectors.

3D Object Detection. We focus on indoor 3D object detection [15, 63, 64, 66] and exclude outdoor methods [3, 42–44], which differ significantly. Prior indoor 3D object detection methods can be categorized into two lines in terms of underlying representation, *i.e.*, voxel-based methods [13, 28, 40], and point-based methods [36, 49, 55, 56, 67]. Voxel-based methods convert point cloud to voxel grids and use 3D convolution networks to address the irregularity of point cloud. In contrast, point-based methods extract point-

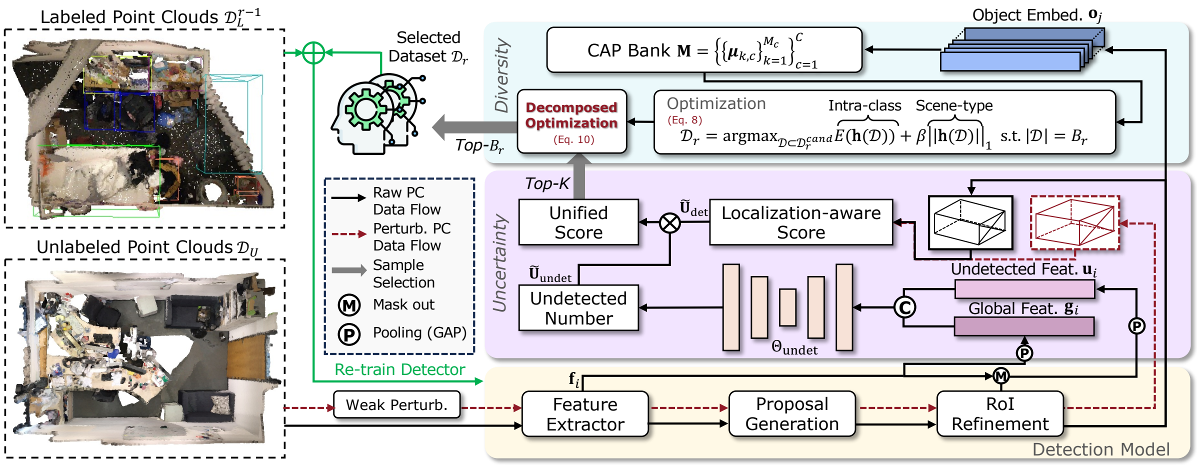

Figure 2. **Overview of our proposed AL framework for indoor 3D object detection, exploiting both uncertainty and diversity**. In the r-th round, we first construct a candidate pool of size $\delta \cdot B_r$ using the Top-K unified uncertainty score, which accounts for both inaccurate detections and undetections. Then, we jointly optimize the intra-class and scene-type diversity to select the r-th selected dataset $\mathcal{D}_r$. Lastly, along with the previous labeled point clouds $\mathcal{D}_L^{r-1}$, we retrain the model until the total labeled dataset reaches the annotation budget B.

level features and use a voting mechanism [36, 56] to encourage points to vote towards their corresponding object centers, followed by grouping the vote points into clusters and then generating object proposals. In this work, we adopt the state-of-the-art method in indoor 3D object detection, CAGroup3D [48], which takes advantage of the two lines of work through a two-stage fully sparse convolutional 3D detection framework. It extracts voxel-wise features from raw point clouds and leverages 'voting and grouping' mechanism to generate object proposals.

Active Learning for Object Detection. Despite extensive studies on AL in image classification, applying it to image-based object detection is challenging due to the task's complexity, which involves both localization and classification of multiple objects. Several works [1, 18, 30, 53, 59, 61] have investigated active object detection by adapting generic AL techniques to this task. For example, CDAL [1] designs contextual diversity to capture spatial and semantic context diversity within a dataset, while DivProto [53] and PPAL [59] incorporate both diversity and uncertainty. While these methods yield promising results in 2D object detection, they cannot be directly applied to 3D object detection due to the significant modality gap. Recently, driven by the need to reduce annotation burden in 3D object detection for autonomous driving, a few studies [8, 12, 21, 24, 25, 27, 29, 32] have explored active learning for 3D object detection in outdoor environments. Notably, CRB [25] selects active samples with concise labels, representative features, and balanced density progressively. KECOR [24] leverages kernel coding rate [26] to measure the richness of latent features. However, these outdoor methods are suboptimal for indoor environments, where 3D object detection faces unique challenges. To address these, this paper presents the *first solution* for applying active learning to indoor 3D object detection, overcoming these challenges by incorporating designs to estimate uncertainty in both inaccurate detections and undetections, while simultaneously optimizing diversity with respect to scene-type and intra-class variances.

3. Methodology

Given a point cloud $\mathcal{P}_i = \{(x, y, z)\}$, a 3D object detection model aims to predict the locations and class labels for a set of objects $\mathfrak{B}_i = \{(\mathbf{b}_k, y_k)\}$, where $\mathbf{b}_k \in \mathbb{R}^7$ represents the parameter of the k-th bounding box, including the center position, box size and yaw angle, and y_k indicates its predicted class label. In the active learning setting, our objective is to train the 3D detection model by constructing a small labeled dataset $\mathcal{D}_L$, which starts with a small portion of random samples $\mathcal{D}_L^0$ and is progressively expanded over R rounds from a large pool $\mathcal{D}$. In each active learning round, $r \in \{1, .., R\}$, B_r samples are selected from the unlabeled dataset $\mathcal{D}_U = \mathcal{D}/\mathcal{D}_L^{r-1}$ by an active learning policy, and labeled by an oracle, which constitute the r-th selected dataset $\mathcal{D}_r$. Then, $\mathcal{D}_L$ will be updated as $\mathcal{D}_L^r \leftarrow \mathcal{D}_L^{r-1} \cup \mathcal{D}_r$ until the size of $\mathcal{D}_L$ reaches the final budget B, *i.e.*, $|\mathcal{D}_L^0| + \sum_{r=1}^{R} B_r = B$ where $B \ll |\mathcal{D}|$.

To tackle active 3D object detection in indoor scenarios, we propose a novel framework, illustrated in Fig. 2, that leverages both uncertainty and diversity as selection criteria. Following previous work [21, 53, 59] that considers both uncertainty and diversity, we adopt a budget expansion ratio $\delta \in \mathbb{Z}^+$ to first select a candidate pool $\mathcal{D}_r^{\text{cand}}$ of size $K = \delta \cdot B_r$ based on the uncertainty criterion for r-th active learning round. Then, from this pool, the $\mathcal{D}_r$ of size

Dataset	mAP@.25	mAP@.50	mAR@.25	mAR@.50
SUN RGB-D [46]	28.91	12.48	67.12	26.09
ScanNetV2 [6]	35.75	20.40	76.65	42.48

Table 1. **Motivation for designing epistemic uncertainty estimation.** We report the mean average precision (mAP) and mean average recall (mAR) at two IoU thresholds (0.25 and 0.50) for the initial detection model, which is trained using 2% labeled data.

B_r is chosen based on the diversity criterion. For the uncertainty criterion, we propose a two-pronged epistemic uncertainty estimation strategy that generates a unified score, accounting for uncertainty in both detected and undetected objects. This is achieved through a localization-aware uncertainty score and an estimate of the undetected object count (Sec. 3.1). For the diversity criterion, we frame the diverse sample selection as an optimization problem and propose an efficient solution that jointly optimizes intra-class variances and scene-type diversity (Sec. 3.2).

3.1. Epistemic Uncertainty Estimation

In Tab. 1, we evaluate the detection performance of a 3D object detection model trained on the initial random samples D_L^0 (2% labeled data). As shown, the model achieves 20.40% mAP@.50 and 42.48% mAR@.50 on ScanNet [6], indicating that approximately 80% of the predictions are inaccurate and 60% of ground-truth objects remain undetected. The results are even worse on the more challenging SUN RGB-D [46] dataset. These findings highlight the need to address the two types of epistemic uncertainty in 3D object detection: *inaccurate detections* and *undetected objects*. Given that the detection task involves both localization and classification, we estimate the epistemic uncertainty of inaccurate detections by considering both aspects. Specifically, we introduce a localization-aware uncertainty score that exploits an IoU score for localization reliability to combine with Shannon entropy. To handle undetections, we design a lightweight network to estimate the number of undetected objects, which serves as the uncertainty measure for undetections. These two uncertainty scores are then integrated into a unified uncertainty score, which is used for Top-K selection to generate the candidate pool $\mathcal{D}_r^{\text{cand}}$.

Localization-aware Uncertainty Score. To estimate the uncertainty of detected objects from a point cloud $\mathcal{P}_i$, we first compute the classification uncertainty using the Shannon entropy [39] of the predicted class probabilities for all detected object proposals. For the j-th proposal, with predicted class probabilities over C classes denoted as $\mathbf{p}_j = \{p_{j,c}\}_{c=1}^{C} \in \mathbb{R}^C$, the Shannon entropy is calculated as:

$$E(\mathbf{p}_j) = -\sum_{c=1}^{C} p_{j,c} \log(p_{j,c}). \tag{1}$$

However, not all detected proposals are valid, as some may correspond to background or partial objects. To mitigate the impact of these invalid proposals, we incorporate an IoU score that measures the similarity between the original detection and its perturbed version, obtained by applying weak augmentations (*e.g.*, random rotation and scaling) to the input point cloud. The intuition is that if the model is confident about an object, small perturbations should not significantly affect its predicted position, while background or partial proposals will likely produce inconsistent predictions. Consequently, the localization-aware uncertainty score for detected objects in $\mathcal{P}_i$ is defined as:

$$\mathbf{U}_{\text{det}}(\mathcal{P}_i) := \frac{\sum_{j=1}^{|\mathfrak{B}_i|} \text{IoU}_j \cdot E(\mathbf{p}_j)}{\sum_{j=1}^{|\mathfrak{B}_i|} \text{IoU}_j}. \tag{2}$$

Our IoU scores serve as soft weights, adaptively reducing the contribution of invalid objects in the epistemic uncertainty estimation.

Undetected Number Prediction Network. When estimating uncertainty for active object detection, previous works [25, 53, 59] typically focus only on the uncertainty within the detected objects. However, the uncertainty of undetected objects is equally important, especially in the context of active learning for indoor 3D object detection, where model performance is often hindered by the scarcity of training samples and severe class imbalance. As shown by the mean average recall (mAR) scores in Tab. 1, more than half of the ground truth objects remain undetected in both 3D indoor datasets. This observation highlights the need to account for the uncertainty of undetected objects. To address this, we design a lightweight network to estimate the number of undetected objects.

As depicted in Fig. 2, we first extract the feature map $\mathbf{f}_i$ from the point cloud $\mathcal{P}_i$ using the feature extractor. We then apply global average pooling (GAP) on $\mathbf{f}_i$ to obtain the global feature $\mathbf{g}_i$. Simultaneously, we create a masked feature map by masking out the regions corresponding to the detected bounding boxes produced by RoI refinement. This masked feature map is passed through GAP to yield the undetected feature $\mathbf{u}_i$. We concatenate these two features, $\mathbf{g}_i$ and $\mathbf{u}_i$, and use the concatenation as input to our undetected object number prediction network Θ_{undet}, a lightweight network consisting of three-layer MLPs. The output of the network, *i.e.* $\Theta_{\text{undet}}([\mathbf{g}_i; \mathbf{u}_i])$ represents the uncertainty score for undetections, denoted as $\mathbf{U}_{\text{undet}}$. This output is compared with the true number of undetected objects, computed based on the detected objects $\mathfrak{B}_i$ and the ground-truth objects $\mathfrak{B}_i^*$, to compute the loss for training Θ_{undet}:

$$\mathcal{L}_{\text{undet}} = \sum_{i=1}^{|\mathcal{D}_L|} (\Theta_{\text{undet}}([\mathbf{g}_i; \mathbf{u}_i]) - \mathcal{N}(\mathfrak{B}_i, \mathfrak{B}_i^*))^2, \tag{3}$$

where $\mathcal{N}(\cdot, \cdot)$ is a function that computes the number of undetected objects, with details provided in the supplementary material. Additionally, to enhance the generalizability of the network Θ_{undet}, we employ random masking of the predicted boxes during the training phase, encouraging the network to better capture the undetectability in point clouds.

Unified Uncertainty Score. After obtaining the two uncertainty scores, $\mathbf{U}_{\text{det}}$ and $\mathbf{U}_{\text{undet}}$, we aim to combine them into a unified uncertainty score. However, these two scores are on different scales, making their direct product infeasible. To address this issue, we introduce a normalization step, inspired by [33], using a hyperparameter k. Specifically, we normalize each score as follows:

$$\tilde{\mathbf{U}}_{\star}(\mathcal{P}_i) = \max(0, \frac{\mathbf{U}_{\star}(\mathcal{P}_i) - \mu(\mathbf{U}_{\star})}{k \cdot \sigma(\mathbf{U}_{\star})} + 0.5), \tag{4}$$

where $\star \in \{\text{det}, \text{undet}\}$, and $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation, respectively. Once the scores are normalized, we compute their product to yield the unified uncertainty score.

3.2. Prototype-based Diversity Optimization

As illustrated in Fig. 1, indoor environments exhibit significant *intra-class variances* and *scene-type diversity*, both of which motivate our approach to the diversity criterion. In 3D indoor scenes, the intra-class variances can vary greatly across object categories due to the differences in shape and occlusion. For example, the variance in 'chairs' is typically high due to the wide variety of shapes and designs, whereas 'doors' tend to have lower variance due to their more standardized geometry. This variability makes methods that model each class with a fixed number of prototypes [57, 65] insufficient to capture the intra-class distributions. To address this, we propose a novel Class-aware Adaptive Prototype (CAP) bank, which dynamically allocates a varying number of prototypes for each class, inferred from the predicted object proposals. Additionally, we observe that indoor scenes naturally form distinct clusters based on the distribution of object categories. For instance, bathrooms typically contain toilets and bathtubs, while bedrooms feature beds and dressers. Through this observation, we represent each scene as a histogram of the prototypes corresponding to the predicted object categories within the scene. As both intra-class variances and scene-type diversity can be modeled using the CAP bank, we frame diverse sample selection as an optimization problem that jointly maximizes both aspects, and propose an efficient solution.

Class-aware Adaptive Prototype Bank. The design of our Class-aware Adaptive Prototype (CAP) bank is inspired by infinite mixture prototypes (IMP) [2], which was originally proposed for few-shot learning. IMP allows for inferring the number of clusters from the data, rather than relying on a pre-defined number of clusters. To build the CAP bank, we first iterate over the initial labeled set D_L^{r-1} to initialize one prototype per class. Specifically, we leverage the ground-truth annotations to extract features for all objects belonging to the c-th class in D_L^{r-1}, and compute their average as to form the initial prototype $\boldsymbol{\mu}_{1,c}$ for class c. Subsequently, we iterate over the unlabeled dataset $\mathcal{D}_U$ and update the prototypes for each class, resulting in our CAP bank, denoted as $\mathbf{M} = \{\{\boldsymbol{\mu}_{k,c}\}_{k=1}^{M_c}\}_{c=1}^{C}$, where $\boldsymbol{\mu}_{k,c}$ represents the k-th prototype for class c, M_c is the number of prototypes for class c, which may vary between classes, and C is the total number of classes. In each iteration (one batch), we use the detection model from the previous round to generate object proposals and extract their features and predicted class labels. For each predicted object, we compute the cosine similarity between its feature embedding $\mathbf{o}_j$ and each of the existing prototypes in its predicted class. This similarity score is formulated as:

$$\text{Sim}(\mathbf{o}_j, \boldsymbol{\mu}_{k,c}) = \frac{\mathbf{o}_j^T \cdot \boldsymbol{\mu}_{k,c}}{||\mathbf{o}_j||_2 ||\boldsymbol{\mu}_{k,c}||_2} \in [-1, 1]. \tag{5}$$

After calculating the similarity scores, each object is either assigned to an existing prototype or used to form a new prototype, according to the following assignment rule:

$$\mathcal{A}(\mathbf{o}_j) = \begin{cases} M_c + 1, & \text{if } \max\limits_k \text{Sim}(\mathbf{o}_j, \boldsymbol{\mu}_{k,c}) < \tau_{\text{sim}}, \\ \arg\max\limits_k \text{Sim}(\mathbf{o}_j, \boldsymbol{\mu}_{k,c}), & \text{otherwise.} \end{cases} \tag{6}$$

Here, τ_{sim} is similarity threshold, which is set to 0.3 by default. Finally, the prototypes - whether assigned or newly formed - are updated using the new object proposals in each batch. However, since many predicted bounding boxes might be invalid, especially in the early stages of training (*e.g.*, as shown in Tab. 1), we seek to mitigate the influence of these invalid proposals when updating the prototypes. To this end, we incorporate the IoU score from Eq. 2 into the prototype update process, effectively weighing each object's contribution to the prototype updates as follows:

$$\boldsymbol{\mu}_{k,c} = \frac{\sum_{j=1}^{|\mathfrak{B}_i|} \mathbb{1}\{\mathcal{A}(\mathbf{o}_j), k\} \cdot \text{IoU}_j \cdot \mathbf{o}_j}{\sum_{j=1}^{|\mathfrak{B}_i|} \mathbb{1}\{\mathcal{A}(\mathbf{o}_j), k\} \cdot \text{IoU}_j}, \tag{7}$$

where the indicator function $\mathbb{1}(m, n)$ equals to 1 if $m = n$, and 0 otherwise. The complete CAP bank creation procedure is outlined in Algo. 1 of the supplementary material.

Prototype-based Diversity Optimization. With the generated CAP bank, we can characterize each class through a set of prototypes, and further represent objects as histograms of these prototypes. These histograms allow us to capture the intra-class variance for each class, as objects within the same class are reflected in the variation of their corresponding histograms. Furthermore, based on the observation that each scene can be characterized by the distribution of object categories, which distinguishes one scene type from another, we also model scenes as histograms of prototypes. Therefore, this histogram-based representation implicitly captures the scene type. Since both intra-class variance and scene-type diversity are encoded in the prototype histograms, we design a novel and efficient optimization method to quantify the diversity criterion by considering both these two aspects. Specifically, we formulate the optimization for diverse sample selection as follows:

$$\begin{aligned} \max_{\mathcal{D}_r \subset \mathcal{D}_r^{\text{cand}}} \quad & E(\mathbf{h}(\mathcal{D}_r)) + \beta ||\mathbf{h}(\mathcal{D}_r)||_1, \\ \text{s.t.} \quad & \mathbf{h}(\mathcal{D}_r) = \sum_{\mathcal{P}_i \in \mathcal{D}_r} \mathbf{h}(\mathcal{P}_i) = \sum_{\mathcal{P}_i \in \mathcal{D}_r} \sum_{\mathbf{o}_j \in \mathcal{P}_i} \mathbf{h}(\mathbf{o}_j), \\ & \mathbf{h}(\mathbf{o}_j) = \sum_{c=1}^{C} \sum_{k=1}^{M_c} [\text{one\_hot}(\mathcal{A}(\mathbf{o}_j))]^{c,k}, \end{aligned} \tag{8}$$

where $\mathbf{h}(\mathbf{o}_j)$ represents the assignment histogram of j-th object *w.r.t.* the CAP bank $\mathbf{M}$, while $\mathbf{h}(\mathcal{P}_i)$ and $\mathbf{h}(\mathcal{D}_r)$ represent the histograms of all objects in the individual point cloud $\mathcal{P}_i$ and the selected labeled samples $\mathcal{D}_r$ at the r-th round, respectively. The two terms in this objective function are designed to select samples that are both **diverse** and **holistic** across all prototypes in the CAP bank. The first term captures the intra-class variance for all classes, ensuring that the selected samples represent the diversity within each class as well as across different classes. The second term promotes scene-level diversity by encouraging coverage across different object categories, which implicitly increases the diversity in scene types. Given that the first term (entropy) and the second term (l_1-norm) have different magnitudes, we introduce a hyperparameter β to rescale the second term. In our experiments, we set $\beta = 0.01$.

However, directly solving the optimization in Eq. 8 is computationally expensive due to the combinatorial search over $|\mathcal{D}_r^{\text{cand}}|$ samples. To efficiently narrow the search space while maintaining solution quality, we leverage one key *property* of the objective function over S disjoint partitions $\{\mathcal{D}_{r,s}\}_{s=1}^{S}$ of $\mathcal{D}_r$, *i.e.*, $\cup_s \mathcal{D}_{r,s} = \mathcal{D}_r$:

$$\begin{aligned} & E(\mathbf{h}(\mathcal{D}_r)) + \beta ||\mathbf{h}(\mathcal{D}_r)||_1 \\ \overset{c}{=} & \sum_{s=1}^{S} \{ \frac{N_s}{N_{\text{total}}} E(\mathbf{h}(\mathcal{D}_{r,s})) + \beta ||\mathbf{h}(\mathcal{D}_{r,s})||_1 \} + \mathcal{O}(\frac{N_{\text{inter}}}{N_{\text{total}}}), \end{aligned} \tag{9}$$

where $\mathbf{h}(\mathcal{D}_{r,s})$ represents the histograms of all objects in s-th partition $\mathcal{D}_{r,s}$, N_s and N_{total} are the number of objects in the s-th partition $\mathcal{D}_{r,s}$ and selected labeled samples $\mathcal{D}_r$, N_{inter} is the number of objects that share the same prototypes but belong to different partitions, and $\mathcal{O}(\cdot)$ is the Big O notation from mathematics. To align with the S disjoint partitions of $\mathcal{D}_r$, we use k-means++ [52] to group the candidate pool $\mathcal{D}_r^{\text{cand}}$ into S clusters, denoted as $\{\mathcal{D}_{r,s}^{\text{cand}}\}_{s=1}^{S}$, each of which is approximately disjoint with each other *w.r.t.* prototype assignments. This guarantees that $\frac{N_{\text{inter}}}{N_{\text{total}}} \rightarrow 0$, and then allows us to decompose the original optimization problem into S sub-problems within each cluster $\mathcal{D}_{r,s}^{\text{cand}}$:

$$\begin{aligned} \max_{\mathcal{D}_{r,s} \subset \mathcal{D}_{r,s}^{\text{cand}}} \quad & E(\mathbf{h}(\mathcal{D}_{r,s})) + \beta \frac{N_{\text{total}}}{N_s} ||\mathbf{h}(\mathcal{D}_{r,s})||_1, \\ \text{s.t.} \quad & \mathbf{h}(\mathcal{D}_{r,s}) = \sum_{\mathcal{P}_i \in \mathcal{D}_{r,s}} \mathbf{h}(\mathcal{P}_i) = \sum_{\mathcal{P}_i \in \mathcal{D}_{r,s}} \sum_{\mathbf{o}_j \in \mathcal{P}_i} \mathbf{h}(\mathbf{o}_j), \\ & \mathbf{h}(\mathbf{o}_j) = \sum_{c=1}^{C} \sum_{k=1}^{M_c} [\text{one\_hot}(\mathcal{A}(\mathbf{o}_j))]^{c,k}. \end{aligned} \tag{10}$$

We use greedy search to solve the S sub-problems to obtain the diverse samples $\mathcal{D}_{r,s}$ for each cluster, and combine them together as the selected dataset $\mathcal{D}_r$ in the r-th active learning round. We provide the detailed derivation of Eq. 9 in the supplementary material.

4. Experiments

4.1. Experimental Setups

3D Detection Datasets. We evaluate our proposed method on two 3D indoor object detection benchmarks: SUN RGB-D [46] and ScanNet [6]. **SUN RGB-D** is an indoor dataset for 3D object detection, which contains 10,335 RGB-D images with 3D room layouts from four different sensors. Officially, they are split into 5,285 training samples and 5,050 testing samples. Following the standard evaluation protocol [19, 35, 36], we evaluate the performance on 10 most common categories. **ScanNetV2** is a large 3D dataset with 1,513 RGB-D scans of 707 indoor scenes. These scans are split into 1,201 samples for training and 312 samples for testing, each of which is annotated with dense semantic segmentation masks. We follow [36] to derive the axis-aligned bounding boxes without considering the orientation, and adopt 18 object categories for evaluation.

Implementation Details. We use the state-of-the-art CAGroup3D [48] as the detection model, and implement our code base using OpenPCDet [47] toolkit. Following the training strategies in [48], we set the voxel size as 0.02m. For both datasets, the network is trained by AdamW optimizer [23] with an initial learning rate and weight decay rate of 0.001 and 0.0001 respectively. The gradnorm clip [62] is applied to stabilize the training process. For each active learning round, SUN RGB-D takes 48 epochs to train on the updated labeled dataset, and the learning rate decays on [32, 44] epochs with decay rate of 0.1. ScanNet requires 120 epochs for training and decays the learning rate on [80, 110] epochs by the same decay rate. All experiments are conducted using one 3090Ti NVIDIA GPU. We set the similarity threshold τ_{sim} to 0.3, balance ratio β to 0.01, and budget expanding ratio δ to 6 empirically.

Evaluation. Following the evaluation protocol in [31, 36, 48], we report mean average precision (mAP) with different IoU thresholds, *i.e.*, 0.25 and 0.50. Additionally, to investigate the effectiveness of different active learning strategies, we record the mAP with 2%, 4%, 6%, 8%, 10% of the train-

	Methods	Presened at	SUN RGB-D		ScanNetV2	
			mAP@0.25 (%)	mAP@0.50 (%)	mAP@0.25 (%)	mAP@0.50 (%)
2D-Generic	RAND	-	42.71	25.14	55.09	39.74
	MC-DROPOUT [9]	ICML'17	47.48	28.36	45.71	28.72
	CORESET [38]	ICLR'18	46.99	28.12	55.34	40.08
	BADGE [4]	ICLR'20	49.12	30.54	57.10	41.97
2D-OD	CDAL [1]	ECCV'20	50.86	30.75	57.99	42.33
	DIVPROTO [53]	CVPR'22	52.31	32.54	59.01	43.10
	PPAL [59]	CVPR'24	52.55	33.51	59.44	43.15
3D-OD	CRB [25]	ICLR'23	50.89	31.30	58.73	42.71
	KECOR [24]	ICCV'23	51.07	31.41	58.54	42.56
	OURS	-	**56.11**	**35.42**	**61.77**	**45.93**
	ORACLE†	-	64.31	47.43	72.65	58.81

Table 2. mAP@0.25 (%) and mAP@0.50 (%) scores of the proposed method and AL baselines on the SUN RGB-D and ScanNetV2 benchmarks with 10% queried point clouds. CAGroup3D is used as the detection model for all approaches. †: indicates the detection results using fully-annotated data, and serves as the upper bound for all active learning strategies.

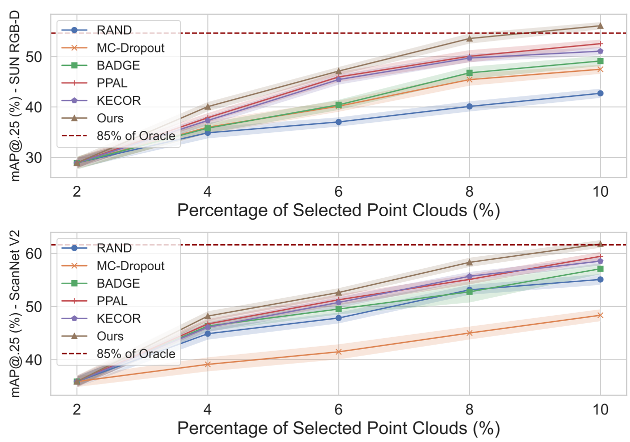

Figure 3. mAP@0.25 (%) score of the proposed method and AL baselines on SUN RGB-D and ScanNetV2 benchmarks.

ing set for both datasets, where the first 2% samples are selected randomly. In each active learning round, 2% of the training point clouds are acquired from the rest unlabeled dataset via active learning strategies for annotation.

Baselines. We evaluate *nine* AL baselines: random sampling (**RAND**), *three* 2D generic AL strategies (**MC-DROPOUT**, **CORESET**, **BADGE**) representing uncertainty-based, diversity-based, and hybrid methods, *three* 2D AL detection strategies (**CDAL**, **DIVPROTO**, **PPAL**) adapted to 3D detection tasks by us, and *two* state-of-the-art 3D outdoor AL detection strategies (**CRB** and **KECOR**).

4.2. Experimental Results

SUN RGB-D. In Fig. 3, we compare our proposed method with different baselines through an increasing percentage of selected point clouds [25, 59]. As demonstrated by the results, our proposed strategy consistently outperforms all competing baselines by significant margins for different percentages of selected point clouds. Notably, with only 10% annotated data, our approach achieves over **85%** of the fully-annotated performance (red dashed line in Fig. 3). Specifically, as detailed in Tab. 2, it surpasses the state-of-the-art 2D active learning strategy, **PPAL**, by 3.56% in mAP@0.25 and 1.91% in mAP@0.50, with a 10% annotation budget. To elaborate further, it is observed that AL methods for 2D object detection generally outperform those for 3D outdoor counterpart. We hypothesize that current 3D outdoor approaches overlook the intra-class variances across different categories, leading to redundancy in selecting active samples. This redundancy may hinder the capability of the learned detection model.

ScanNetV2. To study the effectiveness of our proposed approach on point clouds with complicated room layouts, we conduct experiments using the ScanNetV2 dataset, a more comprehensive benchmark for indoor environments. The results, illustrated in Fig. 3 and Tab. 2, demonstrate the performance comparison across varying annotation budgets. Our method surpasses all existing AL approaches by a huge margin, achieving approximately **85%** of the fully-annotated performance with 10% annotation budget. In Fig. 4, we present a qualitative comparison with random sampling. Random sampling captures only partial object characteristics (*e.g.*, 'bookshelf') due to limited training data. In contrast, our approach successfully detects most of the target objects in challenging scenarios by effectively selecting informative unlabeled samples based on uncertainty and diversity criteria. Additional qualitative results can be found in the supplementary materials.

4.3. Ablation Study

We ablate some critical designs for an in-depth exploration.

Design Choices. We start by exploring the effectiveness of different designs in our proposed method, as shown in Tab. 3. When equipped with the IoU score in Eq. 2 and the uncertainty score for undetections $\mathbf{U}_{\text{undet}}$, the mAP@0.25 increases by 1.1% and 0.7% on SUN RGB-D respectively,

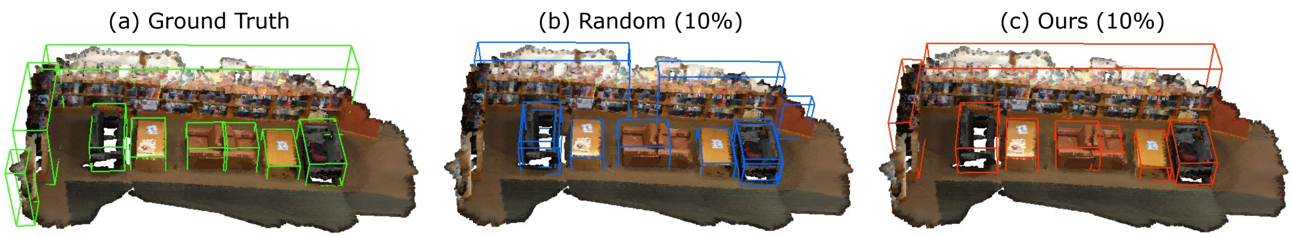

Figure 4. Qualitative comparison between random sampling and our proposed active learning method on ScanNetV2 val set.

Uncertainty	Diversity	SUN RGB-D mAP@.25	SUN RGB-D mAP@.50	ScanNetV2 mAP@.25	ScanNetV2 mAP@.50
✗	✗	42.7	25.1	55.1	39.7
✗	✓	54.5	33.9	58.7	42.9
w/ IoU score	✓	55.6	34.7	60.8	43.8
w/ $\mathrm{U}_{\mathrm{undet}}$	✓	55.2	34.3	60.1	43.3
✓	✗	54.3	33.5	58.4	42.8
✓	w/ K-means	55.3	34.1	60.4	43.5
✓	w/ Greedy-O	54.5	33.3	58.7	43.0
✓	✓	**56.1**	**35.4**	**61.8**	**45.9**

Table 3. Ablation studies of different designs in our proposed method on SUN RGB-D and ScanNetV2 datasets.

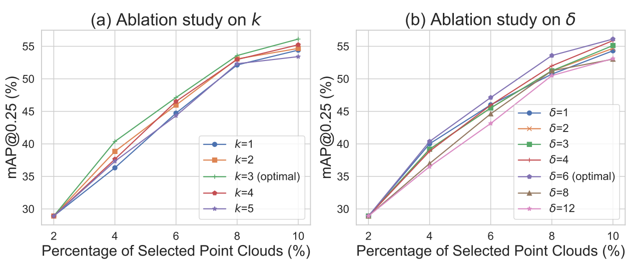

Figure 5. Effects of hyper-parameters in our proposed method on SUN RGB-D dataset.

validating its effectiveness as an uncertainty-based strategy. Additionally, we ablate the diversity branch through two alternatives: (1) Use K-means [52] to select a fixed number of prototypes to replace CAP Bank, denoted as '*K-means*'. (2) Apply greedy search to the original optimization problem as defined in Eq. 8, denoted as '*Greedy-O*'. It demonstrates that the mAP@0.50 of '*K-means*' decreases by 2.4% for ScanNetV2 dataset, which suggests that a fixed prototype number leads to a sub-optimal outcome. Furthermore, '*Greedy-O*' solution only achieves minor performance improvements. A possible reason is that we restrict the searching space of the original optimization problem (Eq. 8) to several regularized sub-spaces (Eq. 10), eliminating a vast number of sub-optimal solutions.

Hyper-parameters. In Fig. 5, we report the ablation study of two hyper-parameters in our proposed method, which are the scale k in Eq. 4 and the budget expanding ratio δ for determining the candidate pool [59]. Overall, the proposed method is relatively insensitive to the choices of hyper-parameters. As depicted in Fig. 5 (a), the best performance is reached with a moderate scale value of $k = 3$, which effectively suppresses extreme values while preserving essential information from different sources of uncertainty. Additionally, we observe that excessively large values of δ will

Measures	SUN RGB-D mAP@.25	SUN RGB-D mAP@.50	ScanNetV2 mAP@.25	ScanNetV2 mAP@.50
Prob. Margin [37]	55.8	35.0	61.5	45.5
MC-Dropout [9]	56.0	35.2	61.6	45.6
Entropy [39]	**56.1**	**35.4**	**61.8**	**45.9**

Table 4. Ablation studies of different uncertainty measures in our proposed method on SUN RGB-D and ScanNetV2 datasets.

significantly degrade model performance in the early stages, which is shown in Fig. 5 (b). We hypothesize that for big δ, the candidate pool contains many samples the model is already confident about, thereby harming the quality of the final selected samples. Conversely, when δ is very small, the selection process is dominated by uncertainty, which restricts the impact of diversity branch. We find that setting $\delta = 6$ achieves a better tradeoff between uncertainty- and diversity-based selection, optimizing overall performance.

Uncertainty Measures. We replace the Shannon entropy in Eq. 2 with two alternative uncertainty measures: MC-Dropout [9] and probability margin [37], and compare their performance in Tab. 4. The results show that all three measures yield similar performance, with entropy performing slightly better. Therefore, we select entropy as our uncertainty measure, as it is both more efficient and effective.

5. Conclusion

This paper presents a novel active learning approach for 3D indoor object detection, improving annotation efficiency by combining uncertainty estimation with diversity-driven sampling. Specifically, our approach estimates epistemic uncertainty, accounting for both inaccurate detections and undetected objects, and selects diverse samples through an optimization method based on a Class-aware Adaptive Prototype bank. Our method significantly outperforms state-of-the-art baselines, achieving over 85% of fully-supervised performance with just 10% of the annotation budget.

6. Acknowledgment

This research is supported by the Ministry of Education, Singapore, under its MOE Academic Research Fund Tier 2 (MOE-T2EP20124-0013), it is also supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021).

References

[1] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 137–153. Springer, 2020. 2, 3, 7

[2] Kelsey Allen, Evan Shelhamer, Hanul Shin, and Joshua Tenenbaum. Infinite mixture prototypes for few-shot learning. In *International conference on machine learning*, pages 232–241. PMLR, 2019. 5

[3] Eduardo Arnold, Omar Y Al-Jarrah, Mehrdad Dianati, Saber Fallah, David Oxtoby, and Alex Mouzakitis. A survey on 3d object detection methods for autonomous driving applications. *IEEE Transactions on Intelligent Transportation Systems*, 20(10):3782–3795, 2019. 2

[4] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019. 2, 7

[5] Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. Batch active learning at scale. *Advances in Neural Information Processing Systems*, 34:11933–11944, 2021. 2

[6] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 2, 4, 6

[7] Ehsan Elhamifar, Guillermo Sapiro, Allen Yang, and S Shankar Sasrty. A convex optimization framework for active learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 209–216, 2013. 2

[8] Di Feng, Xiao Wei, Lars Rosenbaum, Atsuto Maki, and Klaus Dietmayer. Deep active learning for efficient training of a lidar 3d object detector. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 667–674. IEEE, 2019. 3

[9] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017. 2, 7, 8

[10] Mike Gashler and Tony Martinez. Temporal nonlinear dimensionality reduction. In *The 2011 International Joint Conference on Neural Networks*, pages 1959–1966. IEEE, 2011. 1

[11] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 2

[12] Ahmed Ghita, Bjørk Antoniussen, Walter Zimmer, Ross Greer, Christian Creß, Andreas Møgelmose, Mohan M Trivedi, and Alois C Knoll. Activeanno3d–an active learning framework for multi-modal 3d object detection. *arXiv preprint arXiv:2402.03235*, 2024. 2, 3

[13] JunYoung Gwak, Christopher Choy, and Silvio Savarese. Generative sparse detection networks for 3d single-shot object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 297–313. Springer, 2020. 2

[14] Yucheng Han, Na Zhao, Weiling Chen, Keng Teck Ma, and Hanwang Zhang. Dual-perspective knowledge enrichment for semi-supervised 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2049–2057, 2024. 2

[15] Pengkun Jiao, Na Zhao, Jingjing Chen, and Yu-Gang Jiang. Unlocking textual and visual wisdom: Open-vocabulary 3d object detection enhanced by comprehensive guidance from text and image. In *European Conference on Computer Vision*, pages 376–392. Springer, 2024. 2

[16] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *2009 ieee conference on computer vision and pattern recognition*, pages 2372–2379. IEEE, 2009. 2

[17] Yoon-Yeong Kim, Kyungwoo Song, JoonHo Jang, and Il-Chul Moon. Lada: Look-ahead data acquisition via augmentation for deep active learning. *Advances in Neural Information Processing Systems*, 34:22919–22930, 2021. 2

[18] Seong Min Kye, Kwanghee Choi, Hyeongmin Byun, and Buru Chang. Tidal: Learning training dynamics for active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22335–22345, 2023. 3

[19] Jean Lahoud and Bernard Ghanem. 2d-driven 3d object detection in rgb-d images. In *Proceedings of the IEEE international conference on computer vision*, pages 4622–4630, 2017. 6

[20] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 1

[21] Jinpeng Lin, Zhihao Liang, Shengheng Deng, Lile Cai, Tao Jiang, Tianrui Li, Kui Jia, and Xun Xu. Exploring diversity-based active learning for 3d object detection in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2024. 3

[22] Zhuoming Liu, Hao Ding, Huaping Zhong, Weijia Li, Jifeng Dai, and Conghui He. Influence selection for active learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9274–9283, 2021. 2

[23] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[24] Yadan Luo, Zhuoxiao Chen, Zhen Fang, Zheng Zhang, Mahsa Baktashmotlagh, and Zi Huang. Kecor: Kernel coding rate maximization for active 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18279–18290, 2023. 2, 3, 7

[25] Yadan Luo, Zhuoxiao Chen, Zijian Wang, Xin Yu, Zi Huang, and Mahsa Baktashmotlagh. Exploring active 3d object detection from a generalization perspective. *arXiv preprint arXiv:2301.09249*, 2023. 2, 3, 4, 7

[26] Yi Ma, Harm Derksen, Wei Hong, and John Wright. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE transactions on pattern analysis and machine intelligence*, 29(9):1546–1562, 2007. 3

[27] Roberta Maisano, Valeria Tomaselli, Alessandro Capra, Francesco Longo, and Antonio Puliafito. Reducing complexity of 3d indoor object detection. In *2018 IEEE 4th International Forum on Research and Technology for Society and Industry (RTSI)*, pages 1–6. IEEE, 2018. 3

[28] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 922–928. IEEE, 2015. 2

[29] Michael Meyer and Georg Kuschk. Automotive radar dataset for deep learning based 3d object detection. In *2019 16th european radar conference (EuRAD)*, pages 129–132. IEEE, 2019. 3

[30] Peng Mi, Jianghang Lin, Yiyi Zhou, Yunhang Shen, Gen Luo, Xiaoshuai Sun, Liujuan Cao, Rongrong Fu, Qiang Xu, and Rongrong Ji. Active teacher for semi-supervised object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14482–14491, 2022. 1, 3

[31] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2906–2917, 2021. 1, 6

[32] Aimee Moses, Srikanth Jakkampudi, Cheryl Danner, and Derek Biega. Localization-based active learning (local) for object detection in 3d point clouds. In *Geospatial Informatics XII*, pages 44–58. SPIE, 2022. 3

[33] Yuzuru Nakamura, Yasunori Ishii, and Takayoshi Yamashita. Active domain adaptation with false negative prediction for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28782–28792, 2024. 5

[34] Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 79, 2004. 1, 2

[35] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 918–927, 2018. 6

[36] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 2, 3, 6

[37] Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *Machine Learning: ECML 2006: 17th European Conference on Machine Learning Berlin, Germany, September 18-22, 2006 Proceedings 17*, pages 413–424. Springer, 2006. 2, 8

[38] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. 1, 2, 7

[39] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. 2, 4, 8

[40] Xiaoke Shen and Ioannis Stamos. Frustum voxnet for 3d object detection from rgb-d or depth images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1698–1706, 2020. 2

[41] Yichao Shen, Zigang Geng, Yuhui Yuan, Yutong Lin, Ze Liu, Chunyu Wang, Han Hu, Nanning Zheng, and Baining Guo. V-detr: Detr with vertex relative position encoding for 3d object detection. *arXiv preprint arXiv:2308.04409*, 2023. 1

[42] Hualian Sheng, Sijia Cai, Na Zhao, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, Min-Jian Zhao, and Gim Hee Lee. Rethinking iou-based optimization for single-stage 3d object detection. In *European Conference on Computer Vision*, pages 544–561. Springer, 2022. 2

[43] Hualian Sheng, Sijia Cai, Na Zhao, Bing Deng, Min-Jian Zhao, and Gim Hee Lee. Pdr: Progressive depth regularization for monocular 3d object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7591–7603, 2023.

[44] Hualian Sheng, Sijia Cai, Na Zhao, Bing Deng, Qiao Liang, Min-Jian Zhao, and Jieping Ye. Ct3d++: Improving 3d object detection with keypoint-induced channel-wise transformer. *International Journal on Computer Vision*, 2025. 2

[45] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5972–5981, 2019. 2

[46] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 1, 2, 4, 6

[47] OpenPCDet Development Team. Openpcdet: An open-source toolbox for 3d object detection from point clouds. `https://github.com/open-mmlab/OpenPCDet`, 2020. 6

[48] Haiyang Wang, Lihe Ding, Shaocong Dong, Shaoshuai Shi, Aoxue Li, Jianan Li, Zhenguo Li, and Liwei Wang. Cagroup3d: Class-aware grouping for 3d object detection on point clouds. *Advances in Neural Information Processing Systems*, 35:29975–29988, 2022. 1, 2, 3, 6

[49] Jiangyi Wang, Zhongyao Cheng, Na Zhao, Jun Cheng, and Xulei Yang. On-the-fly point feature representation for point clouds analysis. *arXiv preprint arXiv:2407.21335*, 2024. 2

[50] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016. 2

[51] Jamie Watson, Mohamed Sayed, Zawar Qureshi, Gabriel J. Brostow, Sara Vicente, Oisin Mac Aodha, and Michael Firman. Virtual occlusions through implicit depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9053–9064, 2023. 1

[52] Junjie Wu. *Advances in K-means clustering: a data mining thinking*. Springer Science & Business Media, 2012. 2, 6, 8

[53] Jiaxi Wu, Jiaxin Chen, and Di Huang. Entropy-based active learning for object detection with progressive diversity constraint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9397–9406, 2022. 3, 4, 7

[54] Tsung-Han Wu, Yueh-Cheng Liu, Yu-Kai Huang, Hsin-Ying Lee, Hung-Ting Su, Ping-Chia Huang, and Winston H Hsu. Redal: Region-based and diversity-aware active learning for point cloud semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15510–15519, 2021. 2

[55] Qian Xie, Yu-Kun Lai, Jing Wu, Zhoutao Wang, Yiming Zhang, Kai Xu, and Jun Wang. Mlcvnet: Multi-level context votenet for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10447–10456, 2020. 2

[56] Qian Xie, Yu-Kun Lai, Jing Wu, Zhoutao Wang, Dening Lu, Mingqiang Wei, and Jun Wang. Venet: Voting enhancement network for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3712–3721, 2021. 2, 3

[57] Yating Xu, Conghui Hu, Na Zhao, and Gim Hee Lee. Generalized few-shot point cloud segmentation via geometric words. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21506–21515, 2023. 5

[58] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17830–17839, 2023. 1

[59] Chenhongyi Yang, Lichao Huang, and Elliot J Crowley. Plug and play active learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17784–17793, 2024. 3, 4, 7, 8

[60] Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113:113–127, 2015. 2

[61] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5330–5339, 2021. 3

[62] Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. *Advances in Neural Information Processing Systems*, 33:15511–15521, 2020. 6

[63] Na Zhao and Gim Hee Lee. Static-dynamic co-teaching for class-incremental 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3436–3445, 2022. 2

[64] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Sess: Self-ensembling semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11079–11087, 2020. 2

[65] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Few-shot 3d point cloud semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8873–8882, 2021. 5

[66] Na Zhao, Peisheng Qian, Fang Wu, Xun Xu, Xulei Yang, and Gim Hee Lee. Sdcot++: Improved static-dynamic co-teaching for class-incremental 3d object detection. *IEEE Transactions on Image Processing*, 2024. 2

[67] Qiang Zhou and Chaohui Yu. Point rcnn: An angle-free framework for rotated object detection. *Remote Sensing*, 14 (11):2605, 2022. 2