EXPLORING ACTIVE 3D OBJECT DETECTION FROM A GENERALIZATION PERSPECTIVE

Yadan Luo*, **Zhuoxiao Chen***, **Zijian Wang, Xin Yu, Zi Huang, Mahsa Baktashmotlagh**
The University of Queensland, Australia

ABSTRACT

To alleviate the high annotation cost in LiDAR-based 3D object detection, active learning is a promising solution that learns to select only a small portion of unlabeled data to annotate, without compromising model performance. Our empirical study, however, suggests that mainstream uncertainty-based and diversity-based active learning policies are not effective when applied in the 3D detection task, as they fail to balance the trade-off between point cloud informativeness and box-level annotation costs. To overcome this limitation, we jointly investigate three novel criteria in our framework **CRB** for point cloud acquisition - *label conciseness*, *feature representativeness* and *geometric balance*, which hierarchically filters out the point clouds of redundant 3D bounding box labels, latent features and geometric characteristics (*e.g.*, point cloud density) from the unlabeled sample pool and greedily selects informative ones with fewer objects to annotate. Our theoretical analysis demonstrates that the proposed criteria aligns the marginal distributions of the selected subset and the prior distributions of the unseen test set, and minimizes the upper bound of the generalization error. To validate the effectiveness and applicability of CRB, we conduct extensive experiments on the two benchmark 3D object detection datasets of KITTI and Waymo and examine both one-stage (*i.e.*, SECOND) and two-stage 3D detectors (*i.e.*, PV-RCNN). Experiments evidence that the proposed approach outperforms existing active learning strategies and achieves fully supervised performance requiring 1% and 8% annotations of bounding boxes and point clouds, respectively. Source code: https://github.com/Luoyadan/CRB-active-3Ddet.

1 INTRODUCTION

LiDAR-based 3D object detection plays an indispensable role in 3D scene understanding with a wide range of applications such as autonomous driving (Deng et al., 2021; Wang et al., 2020) and robotics (Ahmed et al., 2018; Montes et al., 2020; Wang et al., 2019). The emerging stream of 3D detection models enables accurate recognition at the cost of large-scale labeled point clouds, where 7-degree of freedom (DOF) 3D bounding boxes - consisting of a position, size, and orientation information- for each object are annotated. In the benchmark datasets like Waymo (Sun et al., 2020), there are over 12 million LiDAR boxes, for which, labeling a precise 3D box takes more than 100 seconds for an annotator (Song et al., 2015). This prerequisite for the performance boost greatly hinders the feasibility of applying models to the wild, especially when the annotation budget is limited.

To alleviate this limitation, active learning (AL) aims to reduce labeling costs by querying labels for only a small portion of unlabeled data. The criterion-based query selection process iteratively selects the most beneficial samples for the subsequent model training until the labeling budget is run out. The criterion is expected to quantify the sample informativeness using the heuristics derived from *sample uncertainty* (Gal et al., 2017; Du et al., 2021; Caramalau et al., 2021; Yuan et al., 2021; Choi et al., 2021; Zhang et al., 2020; Shi & Li, 2019) and *sample diversity* (Ma et al., 2021; Gudovskiy et al., 2020; Gao et al., 2020; Sinha et al., 2019; Pinsler et al., 2019). In particular, uncertainty-driven approaches focus on the samples that the model is the least confident of their labels, thus searching for the candidates with: maximum entropy (MacKay, 1992; Shannon, 1948; Kim et al., 2021b; Siddiqui et al., 2020; Shi & Yu, 2019), disagreement among different experts (Freund et al., 1992; Tran et al., 2019), minimum posterior probability of a predicted class (Wang et al., 2017), or the samples

*Equal contribution. Correspondence to Yadan Luo <y.luo@uq.edu.au>.

with reducible yet maximum estimated error (Roy & McCallum, 2001a; Yoo & Kweon, 2019; Kim et al., 2021a). On the other hand, diversity-based methods try to find the most representative samples to avoid sample redundancy. To this end, they form subsets that are sufficiently diverse to describe the entire data pool by making use of the greedy coreset algorithms (Sener & Savarese, 2018), or the clustering algorithms (Nguyen & Smeulders, 2004). Recent works (Liu et al., 2021; Citovsky et al., 2021; Kirsch et al., 2019; Houlsby et al., 2011) combine the aforementioned heuristics: they measure uncertainty as the gradient magnitude of samples (Ash et al., 2020) or its second-order metrics (Liu et al., 2021) at the final layer of neural networks, and then select samples with gradients spanning a diverse set of directions. While effective, the hybrid approaches commonly cause heavy computational overhead, since gradient computation is required for each sample in the unlabeled pool. Another stream of works apply active learning to 2D/3D object detection tasks (Feng et al., 2019; Schmidt et al., 2020; Wang et al., 2022; Wu et al., 2022; Tang et al., 2021b), by leveraging ensemble (Beluch et al., 2018) or Monte Carlo (MC) dropout (Gal & Ghahramani, 2016) algorithms to estimate the classification and localization uncertainty of bounding boxes for images/point clouds acquisition (more details in Appendix I). Nevertheless, those AL methods generally favor the point clouds with more objects, which have a higher chance of containing uncertain and diverse objects. With a fixed annotation budget, it is far from optimal to select such point clouds, since more clicks are required to form 3D box annotations.

To overcome the above limitations, we propose to learn AL criteria for cost-efficient sample acquisition at the 3D box level by empirically studying its relationship with optimizing the generalization upper bound. Specifically, we propose three selection criteria for cost-effective point cloud acquisition, termed as CRB, *i.e., label conciseness*, *feature representativeness* and *geometric balance*. Specifically, we divide the sample selection process into three stages: (1) To alleviate the issues of label redundancy and class imbalance, and to ensure *label conciseness*, we firstly calculate the entropy of bounding box label predictions and only pick top $\mathcal{K}_1$ point clouds for Stage 2; (2) We then examine the *feature representativeness* of candidates by formulating the task as the $\mathcal{K}_2$-medoids problem on the gradient space. To jointly consider the impact of classification and regression objectives on gradients, we enable the Monte Carlo dropout (MC-DROPOUT) and construct the hypothetical labels by averaging predictions from multiple stochastic forward passes. (3) Finally, to maintain the *geometric balance* property, we minimize the KL divergence between the marginal distributions of point cloud density of each predicted bounding box. This makes the trained detector predict more accurate localization and size of objects, and recognize both close (*i.e.*, dense) and distant (*i.e.*, sparse) objects at the test time, using minimum number of annotations. We base our criterion design on our theoretical analysis of optimizing the upper bound of the generalization risk, which can be reformulated as distribution alignment of the selected subset and the test set. Note that since the empirical distribution of the test set is not observable during training, WLOG, we make an appropriate assumption of its prior distribution.

Contributions. Our work is a pioneering study in active learning for 3D object detection, aiming to boost the detection performance at the **lowest cost of bounding box-level annotations**. To this end, we propose a hierarchical active learning scheme for 3D object detection, which progressively filters candidates according to the derived selection criteria without triggering heavy computation. Extensive experiments conducted demonstrate that the proposed CRB strategy can consistently outperform all the state-of-the-art AL baselines on two large-scale 3D detection datasets irrespective of the detector architecture. To enhance the reproducibility of our work and accelerate future work in this new research direction, we develop a `active-3D-det` toolbox, which accommodates various AL approaches and 3D detectors. The source code is available in the supplementary material, and will be publicly shared upon acceptance of the paper.

2 METHODOLOGY

2.1 PROBLEM FORMULATION

In this section, we mathematically formulate the problem of active learning for 3D object detection and set up the notations. Given an orderless LiDAR point cloud $\mathcal{P} = \{x, y, z, e\}$ with 3D location (x, y, z) and reflectance e, the goal of 3D object detection is to localize the objects of interest as a set of 3D bounding boxes $\mathcal{B} = \{b_k\}_{k\in[N_B]}$ with N_B indicating the number of detected bounding boxes, and predict the associated box labels $Y = \{y_k\}_{k\in[N_B]} \in \mathcal{Y} = \{1, \dots, C\}$, with C being the number of classes to predict. Each bounding box b represents the relative center position (p_x, p_y, p_z) to the object ground planes, the box size (l, w, h), and the heading angle θ. Mainstream 3D object detectors

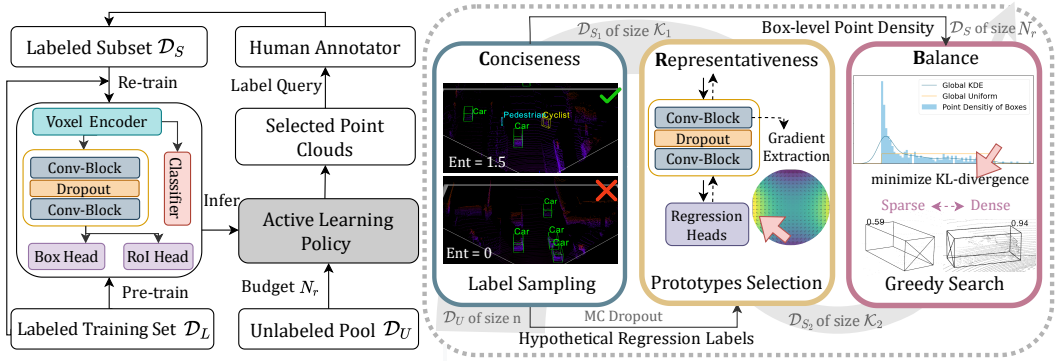

Figure 1: An illustrative flowchart of the proposed CRB framework for active selection of point clouds. Motivated by optimizing the generalization risk, the derived strategy hierarchically selects point clouds that have non-redundant bounding box labels, latent gradients and geometric characteristics to mitigate the gap with the test set and minimize annotation costs.

use point clouds $\mathcal{P}$ to extract point-level features $\boldsymbol{x} \in \mathbb{R}^{W \cdot L \cdot F}$ (Shi et al., 2019; Yang et al., 2019; 2020) or by voxelization (Shi et al., 2020), with W, L, F representing width, length, and channels of the feature map. The feature map $\boldsymbol{x}$ is passed to a classifier $f(\cdot; \boldsymbol{w}_f)$ parameterized by $\boldsymbol{w}_f$ and regression heads $g(\cdot; \boldsymbol{w}_g)$ (*e.g.,* box refinement and ROI regression) parameterized by $\boldsymbol{w}_g$. The output of the model is the detected bounding boxes $\widehat{\mathcal{B}} = \{\hat{b}_k\}$ with the associated box labels $\widehat{Y} = \{\hat{y}_k\}$ from anchored areas. The loss functions ℓ^{cls} and ℓ^{reg} for classification (*e.g.*, regularized cross entropy loss Oberman & Calder (2018)) and regression (*e.g.*, mean absolute error/L_1 regularization Qi et al. (2020)) are assumed to be Lipschitz continuous. As shown in the left half of Figure 1, in an active learning pipeline, a small set of labeled point clouds $\mathcal{D}_L = \{(\mathcal{P}, \mathcal{B}, Y)_i\}_{i \in [m]}$ and a large pool of raw point clouds $\mathcal{D}_U = \{(\mathcal{P})_j\}_{j \in [n]}$ are provided at training time, with n and m being a total number of point clouds and $m \ll n$. For each active learning round $r \in [R]$, and based on the criterion defined by an active learning policy, we select a subset of raw data $\{\mathcal{P}_j\}_{j \in [N_r]}$ from $\mathcal{D}_U$ and query the labels of 3D bounding boxes from an oracle $\boldsymbol{\Omega} : \mathcal{P} \rightarrow \mathcal{B} \times \mathcal{Y}$ to construct $\mathcal{D}_S = \{(\mathcal{P}, \mathcal{B}, Y)_j\}_{j \in [N_r]}$. The 3D detection model is pre-trained with $\mathcal{D}_L$ for active selection, and then retrained with $\mathcal{D}_S \cup \mathcal{D}_L$ until the selected samples reach the final budget B, *i.e.,* $\sum_{r=1}^{R} N_r = B$.

2.2 THEORETICAL MOTIVATION

The core question of active 3D detection is how to design a proper criterion, based on which a fixed number of unlabeled point clouds can be selected to achieve minimum empirical risk $\mathfrak{R}_T[\ell(f, g; \boldsymbol{w})]$ on the test set $\mathcal{D}_T$ and minimum annotation time. Below, inspired by (Mansour et al., 2009; Ben-David et al., 2010), we derive the following **generalization bound** for active 3D detection so that the desired acquisition criteria can be obtained by optimizing the generalization risk.

Theorem 2.1. *Let $\mathcal{H}$ be a hypothesis space of Vapnik-Chervonenkis (VC) dimension d, with f and g being the classification and regression branches, respectively. The $\widehat{\mathcal{D}}_S$ and $\widehat{\mathcal{D}}_T$ represent the empirical distribution induced by samples drawn from the acquired subset $\mathcal{D}_S$ and the test set $\mathcal{D}_T$, and ℓ the loss function bounded by $\mathcal{J}$. It is proven that $\forall\, \delta \in (0, 1)$, and $\forall f, g \in \mathcal{H}$, with probability at least $1 - \delta$ the following inequality holds,*

$$\mathfrak{R}_T[\ell(f, g; \boldsymbol{w})] \leq \mathfrak{R}_S[\ell(f, g; \boldsymbol{w})] + \frac{1}{2} disc(\widehat{\mathcal{D}}_S, \widehat{\mathcal{D}}_T) + \lambda^* + const,$$

where $const = 3\mathcal{J}(\sqrt{\frac{\log \frac{4}{\delta}}{2N_r}} + \sqrt{\frac{\log \frac{4}{\delta}}{2N_t}}) + \sqrt{\frac{2d \log(eN_r/d)}{N_r}} + \sqrt{\frac{2d \log(eN_t/d)}{N_t}}$.

Notably, $\lambda^* = \mathfrak{R}_T[\ell(f^*, g^*; \boldsymbol{w}^*)] + \mathfrak{R}_S[\ell(f^*, g^*; \boldsymbol{w}^*)]$ denotes the joint risk of the optimal hypothesis f^* and g^*, with $\boldsymbol{w}^*$ being the model weights. N_r and N_t indicate the number of samples in the $\mathcal{D}_S$ and $\mathcal{D}_T$. The proof can be found in the supplementary material.

Remark. *The first term indicates the training error on the selected subsets, which is assumed to be trivial based on the zero training assumption (Sener & Savarese, 2018). To obtain a tight upper bound of the generalization risk, the **optimal subset** $\mathcal{D}_S^*$ can be determined via minimizing the discrepancy distance of empirical distribution of two sets, i.e.,*

$$\mathcal{D}_S^* = \underset{\mathcal{D}_S \subset \mathcal{D}_U}{\arg\min}\, disc(\widehat{\mathcal{D}}_S, \widehat{\mathcal{D}}_T).$$

Below, we define the discrepancy distance for the 3D object detection task.

Definition 1. *For any $f, g, f', g' \in \mathcal{H}$, the discrepancy between the distribution of the selected sets $\mathcal{D}_S$ and unlabeled pool $\mathcal{D}_T$ can be formulated as,*

$$disc(\widehat{\mathcal{D}}_S, \widehat{\mathcal{D}}_T) = \sup_{f,f' \in \mathcal{H}} |\mathbb{E}_{\widehat{\mathcal{D}}_S} \ell(f, f') - \mathbb{E}_{\widehat{\mathcal{D}}_T} \ell(f, f')| + \sup_{g,g' \in \mathcal{H}} |\mathbb{E}_{\widehat{\mathcal{D}}_S} \ell(g, g') - \mathbb{E}_{\widehat{\mathcal{D}}_T} \ell(g, g')|,$$

where the bounded expected loss ℓ for any classification and regression functions are symmetric and satisfy the triangle inequality.

Remark. *As 3D object detection is naturally an integration of classification and regression tasks, mitigating the set discrepancy is basically aligning the inputs and outputs of each branch. Therefore, with the detector freezed during the active selection, finding an optimal $\mathcal{D}_S^*$ can be interpreted as enhancing the acquired set's (1) **Label Conciseness**: aligning marginal label distribution of bounding boxes, (2) **Feature Representativeness**: aligning marginal distribution of the latent representations of point clouds, and (3) **Geometric Balance**: aligning marginal distribution of geometric characteristics of point clouds and predicted bounding boxes, and can be written as:*

$$\mathcal{D}_S^* \approx \underset{\mathcal{D}_S \subset \mathcal{D}_U}{\arg\min} \underbrace{d_{\mathcal{A}}(P_{\widehat{Y}_S}, P_{Y_T})}_{\textit{Conciseness}} + \underbrace{d_{\mathcal{A}}(P_{X_S}, P_{X_T})}_{\textit{Representativeness}} + \underbrace{d_{\mathcal{A}}(P_{\phi(\mathcal{P}_S, \widehat{\mathcal{B}}_S)}, P_{\phi(\mathcal{P}_T, \mathcal{B}_T)})}_{\textit{Balance}}. \tag{1}$$

Here, $\mathcal{P}_S$ and $\mathcal{P}_T$ represent the point clouds in the selected set and the ones in the test set. $\phi(\cdot)$ indicates the geometric descriptor of point clouds and $d_{\mathcal{A}}$ distance (Kifer et al., 2004) which can be estimated by a finite set of samples. For latent features X_S and X_T, we only focus on the features that differ from the training sets, since $\mathbb{E}_{\widehat{D}_L} \ell^{cls} = 0$ and $\mathbb{E}_{\widehat{D}_L} \ell^{reg} = 0$ based on the zero training error assumption. Considering that test samples and their associated labels are not observable during training, we make an assumption on the prior distributions of test data. WLOG, we assume that the prior distribution of bounding box labels and geometric features are uniform. Note that we can adopt the KL-divergence for the implementation of $d_{\mathcal{A}}$ assuming that latent representations follow the univariate Gaussian distribution.

Connections with existing AL approaches. The proposed criteria jointly optimize the discrepancy distance for both tasks with three objectives, which shows the connections with existing AL strategies. The uncertainty-based methods focus strongly on the first term, based on the assumption that learning more difficult samples will help to improve the suprema of the loss. This rigorous assumption can result in a bias towards hard samples, which will be accumulated and amplified across iterations. Diversity-based methods put more effort into minimizing the second term, aiming to align the distributions in the latent subspace. However, the diversity-based approaches are unable to discover the latent features specified for regression, which can be critical when dealing with a detection problem. We introduce the third term for the 3D detection task, motivated by the fact that aligning the geometric characteristics of point clouds helps to preserve the fine-grained details of objects, leading to more accurate regression. Our empirical study provided in Sec. 3.3 suggests jointly optimizing three terms can lead to the best performance.

2.3 Our Approach

To optimize the three criteria outlined in Eq. 1, we derive an AL scheme consisting of three components. In particular, to reduce the computational overhead, we hierarchically filter the samples that meet the selection criteria (illustrated in Fig. 1): we first pick $\mathcal{K}_1$ candidates by concise label sampling (**Stage 1**), from which we select $\mathcal{K}_2$ representative prototypes (**Stage 2**), with $\mathcal{K}_1, \mathcal{K}_2 << n$. Finally, we leverage greedy search (**Stage 3**) to find the N_r prototypes that match with the prior marginal distribution of test data. The hierarchical sampling scheme can save $\mathcal{O}((n - \mathcal{K}_1)T_2 + (n - \mathcal{K}_2)T_3)$ cost, with T_2 and T_3 indicating the runtime of criterion evaluation. The algorithm is summarized in the supplemental material. In the following, we describe the details of the three stages.

Stage 1: Concise Label Sampling (CLS). By using *label conciseness* as a sampling criterion, we aim to alleviate label redundancy and align the source label distribution with the target prior label distribution. Particularly, we find a subset $\mathcal{D}_{S_1}^*$ of size $\mathcal{K}_1$ that minimizes Kullback-Leibler (KL) divergence between the probability distribution P_{Y_S} and the uniform distribution P_{Y_T}. To this end, we formulate the KL-divergence with Shannon entropy $H(\cdot)$ and define an optimization problem of

maximizing the entropy of the label distributions:

$$D_{KL}(P_{\widehat{Y}_{S_1}} \parallel P_{Y_T}) = -H(\widehat{Y}_{S_1}) + \log|\widehat{Y}_{S_1}|, \tag{2}$$

$$\mathcal{D}^*_{S_1} = \underset{\mathcal{D}_{S_1}\subset\mathcal{D}_U}{\arg\min}\, D_{KL}(P_{\widehat{Y}_{S_1}} \parallel P_{Y_T}) = \underset{\mathcal{D}_{S_1}\subset\mathcal{D}_U}{\arg\max}\, H(\widehat{Y}_{S_1}), \tag{3}$$

where $\log|\widehat{Y}_{S_1}| = log\mathcal{K}_1$ indicates the number of values Y_{S_1} can take on, which is a constant. Note that P_{Y_T} is a uniform distribution, and we removed the constant values from the formulations. We pass all point clouds $\{(\mathcal{P})_j\}_{i\in[n]}$ from the unlabeled pool to the detector and extract the predictive labels $\{\hat{y}_i\}_{i=1}^{N_B}$ for N_B bounding boxes, with $\hat{y}_i = \arg\max_{y\in[C]} f(x_i; \boldsymbol{w}_f)$. The label entropy of the j-th point cloud $H(\widehat{Y}_{j,S})$ can be calculated as,

$$H(\widehat{Y}_{j,S}) = -\sum_{c=1}^{C} \boldsymbol{p}_{i,c}\log\boldsymbol{p}_{i,c}, \quad \boldsymbol{p}_{i,c} = \frac{e^{|\hat{y}_i=c|/N_B}}{\sum_{c=1}^{C} e^{|\hat{y}_i=c|/N_B}}. \tag{4}$$

Based on the calculated entropy scores, we filter out the top-$\mathcal{K}_1$ candidates and validate them through the **Stage 2** representative prototype selection.

Stage 2: Representative Prototype Selection (RPS). In this stage, we aim to to identify whether the subsets cover the *unique* knowledge encoded only in $\mathcal{D}_U$ and not in $\mathcal{D}_L$ by measuring the *feature representativeness* with gradient vectors of point clouds. Motivated by this, we find the representative prototypes on the gradient space $\mathcal{G}$ to form the subset $\mathcal{D}_{S_2}$, where magnitude and orientation represent the uncertainty and diversity of the new knowledge. For a classification problem, gradients can be retrieved by feeding the hypothetical label $\hat{y} = \arg\max_{y\in[C]} \boldsymbol{p}(y|x)$ to the networks. However, the gradient extraction for regression problem is not explored yet in the literature, due to the fact that the hypothetical labels for regression heads cannot be directly obtained. To mitigate this, we propose to enable Monte Carlo dropout (MC-DROPOUT) at the **Stage 1**, and get the averaging predictions $\bar{B}$ of M stochastic forward passes through the model as the hypothetical labels for regression loss:

$$\bar{B} \approx \frac{1}{M}\sum_{i=1}^{M} g(\boldsymbol{x}; \boldsymbol{w}_d, \boldsymbol{w}_g), \boldsymbol{w}_d \sim \texttt{Bernoulli}(1-p), \tag{5}$$

$$G_{S_2} = \{\nabla_\Theta \ell^{reg}(g(\boldsymbol{x}), \bar{B}; \boldsymbol{w}_g), \boldsymbol{x} \sim \mathcal{D}_{S_2}\}, \tag{6}$$

with p indicating the dropout rate, $\boldsymbol{w}_d$ the random variable of the dropout layer, and Θ the parameters of the convolutional layer of the shared block. The gradient maps $G_{S_2} \in \mathcal{G}$ can be extracted from shared layers and calculated by the chain rule. Since the gradients for test samples are not observable, we make an assumption that its prior distribution follows a Gaussian distribution, which allows us to rewrite the optimization function as,

$$\begin{aligned}\mathcal{D}^*_{S_2} &= \underset{\mathcal{D}_{S_2}\subset\mathcal{D}_{S_1}}{\arg\min}\, D_{KL}(P_{X_{S_2}} \parallel P_{X_T}) \approx \underset{\mathcal{D}_{S_2}\subset\mathcal{D}_{S_1}}{\arg\min}\, D_{KL}(P_{G_{S_2}} \parallel P_{G_T}) \\ &= \underset{\mathcal{D}_{S_2}\subset\mathcal{D}_{S_1}}{\arg\min}\, \log\frac{\sigma_T}{\sigma_{S_2}} + \frac{\sigma^2_{S_2} + (\mu_{S_2} - \mu_T)}{2\delta^2_T} - \frac{1}{2} \approx \mathcal{K}_2\text{-}\texttt{medoids}(G_{S_1}),\end{aligned} \tag{7}$$

with μ_{S_2}, σ_{S_2} (μ_T, and σ_T) being the mean and a standard deviation of the univariate Gaussian distribution of the selected set (test set), respectively. Based on Eq. 7, the task of finding a representative set can be viewed as picking $\mathcal{K}_2$ prototypes (*i.e.,* $\mathcal{K}_2$-medoids) from the clustered data, so that the centroids (mean value) of the selected subset and the test set can be naturally matched. The variance σ_{S_2} and σ_T, basically, the distance of each point to its prototypes will be minimized simultaneously. We test different approaches for selecting prototypes in Sec. 3.3.

Stage 3: Greedy Point Density Balancing (GPDB). The third criterion adopted is *geometric balance*, which targets at aligning the distribution of selected prototypes with the marginal distribution of testing point clouds. As point clouds typically consist of thousands (if not millions) of points, it is computationally expensive to directly align the meta features (*e.g.,* coordinates) of points. Furthermore, in representation learning for point clouds, the common practice of using voxel-based architecture typically relies on quantized representations of point clouds and loses the object details due to the limited perception range of voxels. Therefore, we utilize the point density $\phi(\cdot,\cdot)$ within each bounding box to preserve the geometric characteristics of an object in 3D point clouds. By

aligning the geometric characteristic of the selected set and unlabeled pool, the fine-tuned detector is expected to predict more accurate localization and size of bounding boxes and recognize both close (*i.e.,* dense) and distant (*i.e.,* sparse) objects at the test time. The probability density function (PDF) of the point density is not given and has to be estimated from the bounding box predictions. To this end, we adopt Kernel Density Estimation (KDE) using a finite set of samples from each class which can be computed as:

$$\boldsymbol{p}(\phi(\mathcal{P}, \widehat{\mathcal{B}})) = \frac{1}{N_B h} \sum_{j=1}^{N_B} \mathcal{K}er(\frac{\phi(\mathcal{P}, \widehat{\mathcal{B}}) - \phi(\mathcal{P}, \widehat{\mathcal{B}}_j)}{h}), \tag{8}$$

with $h > 0$ being the pre-defined bandwidth that can determine the smoothing of the resulting density function. We use Gaussian kernel for the kernel function $\mathcal{K}er(\cdot)$. With the PDF defined, the optimization problem of selecting the final candidate sets $\mathcal{D}_S$ of size N_r for the label query is:

$$\mathcal{D}_S^* = \underset{\mathcal{D}_S \subset \mathcal{D}_{S_2}}{\arg\min} D_{KL}(\phi(\mathcal{P}_S, \widehat{\mathcal{B}}_S) \parallel \phi(\mathcal{P}_T, \mathcal{B}_T)), \tag{9}$$

where $\phi(\cdot, \cdot)$ measures the point density for each bounding box. We use greedy search to find the optimal combinations from the subset $\mathcal{D}_{S_2}$ that can minimize the KL distance to the uniform distribution $\boldsymbol{p}(\phi(\mathcal{P}_T, \mathcal{B}_T)) \sim \texttt{uniform}(\alpha_{lo}, \alpha_{hi})$. The upper bound α_{hi} and lower bound α_{lo} of the uniform distribution are set to the 95% density interval, *i.e.,* $\boldsymbol{p}(\alpha_{lo} < \phi(\mathcal{P}, \widehat{\mathcal{B}}_j) < \alpha_{hi}) = 95\%$ for every predicted bounding box j. Notably, the density of each bounding box is recorded during the **Stage 1**, which will not cause any computation overhead. The analysis of time complexity against other active learning methods is presented in Sec. 3.4.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Datasets. KITTI (Geiger et al., 2012) is one of the most representative datasets for point cloud based object detection. The dataset consists of 3,712 training samples (*i.e.,* point clouds) and 3,769 *val* samples. The dataset includes a total of 80,256 labeled objects with three commonly used classes for autonomous driving: cars, pedestrians, and cyclists. The Waymo Open dataset (Sun et al., 2020) is a challenging testbed for autonomous driving, containing 158,361 training samples and 40,077 testing samples. The sampling intervals for KITTI and Waymo are set to 1 and 10, respectively.

Generic AL Baselines. We implemented the following five generic AL baselines of which the implementation details can be found in the supplementary material. (1) **RAND**: is a basic sampling method that selects N_r samples at random for each selection round; (2) **ENTROPY** (Wang & Shang, 2014): is an *uncertainty*-based active learning approach that targets the *classification* head of the detector, and selects the top N_r ranked samples based on the entropy of the sample's predicted label; (3) **LLAL** (Yoo & Kweon, 2019): is an *uncertainty*-based method that adopts an auxiliary network to predict an indicative loss and enables to select samples for which the model is likely to produce wrong predictions; (4) **CORESET** (Sener & Savarese, 2018): is a *diversity*-based method performing the core-set selection that uses the greedy furthest-first search on both labeled and unlabeled embeddings at each round; and (5) **BADGE** (Ash et al., 2020): is a *hybrid* approach that samples instances that are disparate and of high magnitude when presented in a hallucinated gradient space.

Applied AL Baselines for 2D and 3D Detection. For a fair comparison, we also compared three variants of the deep active learning method for 3D detection and adapted one 2D active detection method to our 3D detector. (6) **MC-MI** (Feng et al., 2019) utilized Monte Carlo dropout associated with mutual information to determine the uncertainty of point clouds. (7) **MC-REG**: Additionally, to verify the importance of the uncertainty in regression, we design an *uncertainty*-based baseline that determines the *regression* uncertainty via conducting M-round MC-DROPOUT stochastic passes at the test time. The variances of predictive results are then calculated, and the samples with the top-N_r greatest variance will be selected for label acquisition. We further adapted two applied AL methods for 2D detection to a 3D detection setting, where (8) **LT/C** (Kao et al., 2018) measures the class-specific localization tightness, *i.e.*, the changes from the intermediate proposal to the final bounding box and (9) **CONSENSUS** (Schmidt et al., 2020) calculates the variation ratio of minimum IoU value for each RoI-match of 3D boxes.

Table 1: Performance comparisons (3D AP scores) with generic AL and applied AL for detection on KITTI *val* set with 1% queried bounding boxes.

	Method	Car			Pedestrian			Cyclist			Average		
		Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
Generic	Coreset	87.77	77.73	72.95	47.27	41.97	38.19	81.73	59.72	55.64	72.26	59.81	55.59
	Badge	89.96	75.78	70.54	51.94	46.24	40.98	84.11	62.29	58.12	75.34	61.44	56.55
	LLAL	89.95	78.65	**75.32**	56.34	49.87	45.97	75.55	60.35	55.36	73.94	62.95	58.88
AL-Det	MC-Reg	88.85	76.21	73.47	35.82	31.81	29.79	73.98	55.23	51.85	66.21	54.41	51.70
	MC-MI	86.28	75.58	71.56	41.05	37.50	33.83	86.26	60.22	56.04	71.19	57.77	53.81
	Consensus	90.14	78.01	74.28	56.43	49.50	44.80	78.46	55.77	53.73	75.01	61.09	57.60
	Lt/c	88.73	78.12	73.87	55.17	48.37	43.63	83.72	63.21	59.16	75.88	63.23	58.89
	Crb	**90.98**	**79.02**	74.04	**64.17**	**54.80**	**50.82**	**86.96**	**67.45**	**63.56**	**80.70**	**67.81**	**62.81**

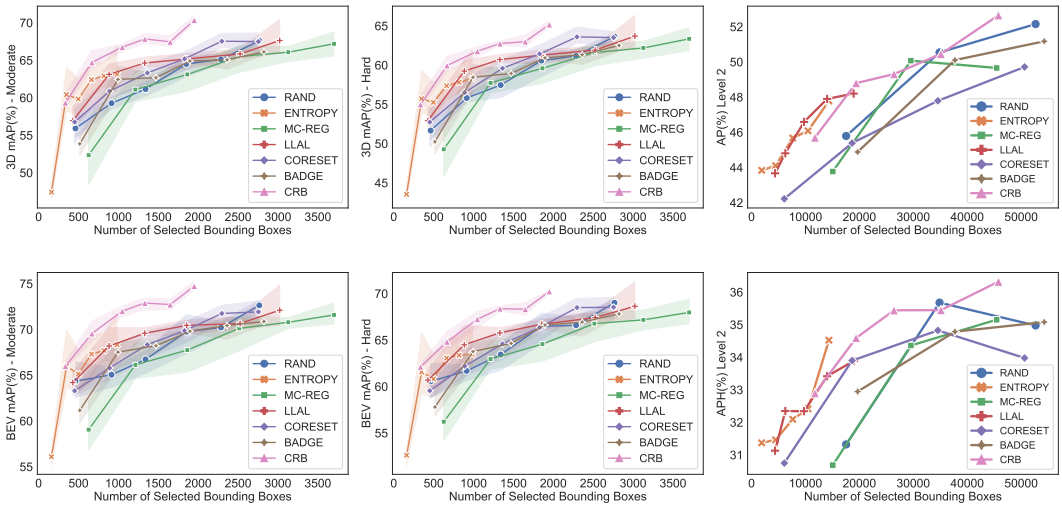

Figure 2: 3D and BEV mAP (%) of Crb and AL baselines on the KITTI and Waymo *val* split.

3.2 Comparisons against Active Learning Methods

Quantitative Analysis. We conducted comprehensive experiments on the KITTI and Waymo datasets to demonstrate the effectiveness of the proposed approach. The $\mathcal{K}_1$ and $\mathcal{K}_2$ are empirically set to 300 and 200 for KITTI and $2,000$ and $1,200$ for Waymo. Under a fixed budget of point clouds, the performance of 3D and BEV detection achieved by different AL policies are reported in Figure 2, with standard deviation of three trials shown in shaded regions. We can clearly observe that Crb consistently outperforms all state-of-the-art AL methods by a noticeable margin, irrespective of the number of annotated bounding boxes and difficulty settings. It is worth noting that, on the KITTI dataset, the annotation time for the proposed Crb is 3 times faster than Rand, while achieving a comparable performance. Moreover, AL baselines for regression and classification tasks (*e.g.*, LLAL) or for regression only tasks (*e.g.*, MC-Reg) generally obtain higher scores yet leading to higher labeling costs than the classification-oriented methods (*e.g.*, Entropy).

Table 1 reports the major experimental results of the state-of-the-art generic AL methods and applied AL approaches for 2D and 3D detection on the KITTI dataset. It is observed that LLAL and Lt/c achieve competitive results, as the acquisition criteria adopted jointly consider the classification and regression task. Our proposed Crb improves the 3D mAP scores by 6.7% which validates the effectiveness of minimizing the generalization risk.

Qualitative Analysis. To intuitively understand the merits of our proposed active 3D detection strategy, Figure 10 demonstrates that the 3D detection results yielded by **Rand** (bottom left) and **Crb** selection (bottom right) from the corresponding image (upper row). Both 3D detectors are trained under the budget of 1K annotated bounding boxes. False positives and corrected predictions are indicated with red and green boxes. It is observed that, under the same condition, Crb produces more accurate and more confident predictions than Rand. Besides, looking at the cyclist highlighted in the orange box in Figure 10, the detector trained with Rand produces a significantly lower

Table 2: Ablative study of different active learning criteria on the KITTI *val* split. 3D and BEV AP scores (%) are reported when 1,000 bounding boxes are annotated.

			3D Detection mAP			BEV Detection mAP		
CLS	RPS	GPDB	EASY	MODERATE	HARD	EASY	MODERATE	HARD
-	-	-	$70.70_{\pm 1.60}$	$58.27_{\pm 1.04}$	$54.69_{\pm 1.30}$	$75.37_{\pm 1.65}$	$64.54_{\pm 1.69}$	$61.36_{\pm 1.61}$
√	-	-	$77.76_{\pm 1.70}$	$64.56_{\pm 1.39}$	$59.54_{\pm 1.13}$	$81.07_{\pm 1.67}$	$69.76_{\pm 1.45}$	$65.01_{\pm 1.31}$
-	√	-	$74.93_{\pm 3.11}$	$61.65_{\pm 1.95}$	$57.70_{\pm 1.52}$	$78.85_{\pm 2.31}$	$67.07_{\pm 1.36}$	$63.47_{\pm 1.21}$
-	-	√	$69.11_{\pm 13.22}$	$56.12_{\pm 12.74}$	$52.85_{\pm 11.49}$	$73.57_{\pm 10.45}$	$62.49_{\pm 10.62}$	$59.45_{\pm 9.78}$
√	√	-	$76.19_{\pm 2.13}$	$62.81_{\pm 1.31}$	$58.03_{\pm 1.18}$	$80.73_{\pm 0.92}$	$68.67_{\pm 0.21}$	$64.42_{\pm 0.22}$
√	-	√	$76.72_{\pm 0.78}$	$64.70_{\pm 1.07}$	$59.68_{\pm 0.93}$	$80.71_{\pm 0.26}$	$70.01_{\pm 0.40}$	$65.47_{\pm 0.56}$
√	√	√	$\mathbf{79.03}_{\pm 1.39}$	$\mathbf{65.86}_{\pm 1.21}$	$\mathbf{61.06}_{\pm 1.43}$	$\mathbf{82.60}_{\pm 1.34}$	$\mathbf{70.74}_{\pm 0.57}$	$\mathbf{66.41}_{\pm 1.22}$

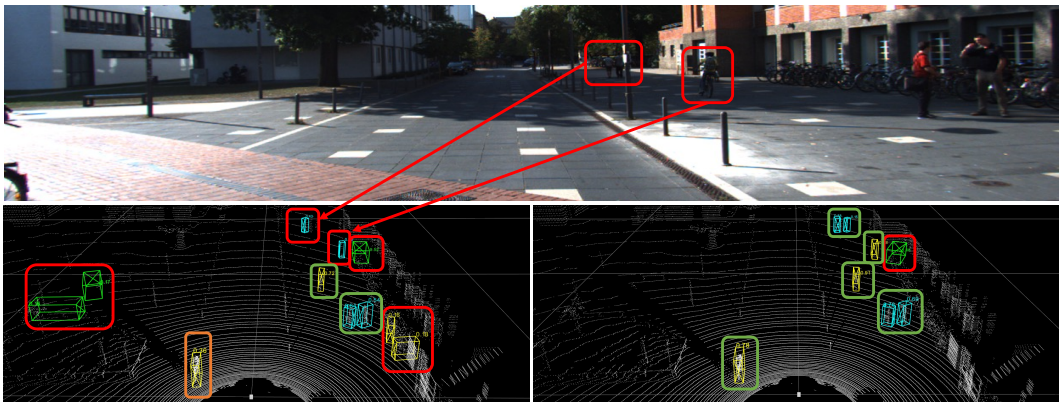

Figure 3: A case study of active 3D detection performance of **RAND** (bottom left) and **CRB** (bottom right) under the budge of 1,000 annotated bounding boxes. False positive (corrected predictions) are highlighted in red (green) boxes. The orange box denotes the detection with low confidence.

confidence score compared to our approach. This confirms that the samples selected by CRB are aligned better with the test cases. More visualizations can be found in the supplemental material.

3.3 ABLATION STUDY

Study of Active Selection Criteria. Table 2 reports the performance comparisons of six variants of the proposed CRB method and the basic random selection baseline (1-st row) on the KITTI dataset. We report the 3D and BEV mAP metrics at all difficulty levels with 1,000 bounding boxes annotated. We observe that only applying GPDB (4-th row) produces 12.5% lower scores and greater variance than the full model (the last row). However, with CLS (6-th row), the performance increases by approximately 10% with the minimum variance. This phenomenon evidences the importance of optimizing the discrepancy for both classification and regression tasks. It's further shown that removing any selection criteria from the proposed CRB triggers a drop on mAP scores, confirming the importance of each in a sample-efficient AL strategy.

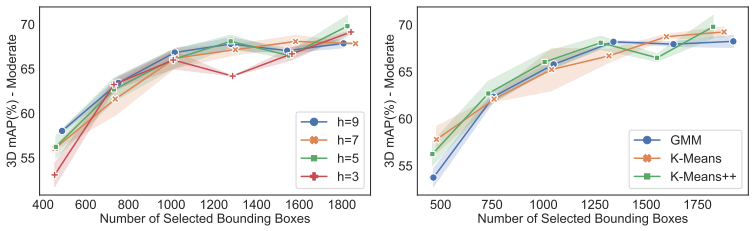

Figure 4: Results on KITTI *val* set with varying KDE bandwidth h (left) and prototype selection approaches (right) with increasing queried bounding boxes.

Sensitivity to Prototype Selection. We examine the sensitivity of performance to different prototype selection methods used in the RPS module on the KITTI dataset (moderate difficulty level). Particularly, In Figure 4 (right), we show the performance of our approach using different prototype selection methods of the Gaussian mixture model (GMM), K-MEANS, and K-MEANS++. To fairly reflect the trend in the performance curves, we run two trials for each prototype selection approach and plot the mean and the variance bars. K-MEANS is slightly more stable than the other two, with higher time complexity and better representation learning. It is observed that there is very little variation ($\sim 1.5\%$) in the performance of our approach when using different prototype selection methods. This confirms that the CRB's superiority over existing baselines is not coming from the prototype selection method.

Table 3: Performance comparisons on KITTI *val* set *w.r.t.* varying thresholds $\mathcal{K}_1$ and $\mathcal{K}_2$ after two rounds of active selection (8% point clouds, 1% bounding boxes). Results are reported with 3D AP with 40 recall positions. † indicates the reported performance of the backbone trained with the full labeled set (100%).

		CAR			PEDESTRIAN			CYCLIST			AVERAGE		
$\mathcal{K}_1$	$\mathcal{K}_2$	EASY	MOD.	HARD	EASY	MOD.	HARD	EASY	MOD.	HARD	EASY	MOD.	HARD
500	400	90.04	79.08	**74.66**	57.11	51.10	**51.12**	81.97	63.40	59.62	76.50	64.53	60.10
500	300	90.98	79.02	74.04	64.17	54.80	50.82	**86.96**	**67.45**	**63.56**	**80.70**	**67.81**	**62.81**
400	300	**91.30**	**79.21**	74.00	62.93	55.67	49.27	79.02	60.50	56.74	77.75	65.13	60.00
300	200	90.45	78.81	73.44	**65.00**	**55.91**	**51.12**	84.82	65.77	61.53	80.09	67.32	62.05
PV-RCNN†		92.56	84.36	82.48	64.26	56.67	51.91	88.88	71.95	66.78	81.75	70.99	67.06

Sensitivity to Bandwidth h. Figure 4 depicts the results of CRB with the bandwidth h varying in $\{3, 5, 7, 9\}$. Choosing the optimal bandwidth value h^* can avoid under-smoothing ($h < h^*$) and over-smoothing ($h > h^*$) in KDE. Except $h = 3$ which yields a large variation, CRB with the bandwidth of all other values reach similar detection results within the 2% absolute difference on 3D mAP. This evidences that the CRB is robust to different values of bandwidth.

Sensitivity to Detector Architecture. We validate the sensitivity of performance to choices of one-stage and two-stage detectors. Table 4 reports the results with the SECOND detection backbone on the KITTI dataset. With only 3% queried 3D bounding boxes, it is observed that the proposed CRB approach consistently outperforms the SOTA generic active learning approaches across a range of detection difficulties, improving 4.7% and 2.8% on 3D mAP and BEV mAP scores.

Sensitivity Analysis of Thresholds $\mathcal{K}_1$ and $\mathcal{K}_2$. We examine the sensitivity of our approach to different values of threshold parameters $\mathcal{K}_1$ and $\mathcal{K}_2$. We report the mean average precision (mAP) on the KITTI dataset, including both 3D and BEV views at all difficulty levels. We check four possible combinations of $\mathcal{K}_1$ and $\mathcal{K}_2$ and show the results in Table 3. We can observe that at MODERATE and HARD levels, there is only 3.28% and 2.81% fluctuation on average mAP. In the last row, we further report the accuracy achieved by the backbone detector trained with all labeled training data and a larger batch size. With only 8% point clouds and 1% annotated bounding boxes, CRB achieves a comparable performance to the full model.

3.4 COMPLEXITY ANALYSIS

Table 5 shows the time complexity of training and active selection for different active learning approaches. n indicates the total number of unlabeled point clouds, N_r is the quantity selected, and E is the training epochs, with $N_r \ll n$. We can clearly observe that, at training stage, the complexity of all AL strategies is $\mathcal{O}(En)$, except LLAL that needs extra epochs E_l to train the loss prediction module. At the active selection stage, RAND randomly generates N_r indices to retrieve samples from the pool. CORESET computes pairwise distances between the embedding of selected samples and unlabeled samples that yields the time complexity of $\mathcal{O}(N_r n)$. BADGE iterates through the gradients of all unlabeled samples passing gradients into K-MEANS++ algorithm, with the complexity of $\mathcal{O}(N_r n)$ bounded by K-MEANS++. Given $\mathcal{K}_1, \mathcal{K}_2 \approx N_r$, the time complexity of our method is $\mathcal{O}(n \log n + 2N_r^2)$, with $\mathcal{O}(n \log(n))$ being the complexity of sorting the entropy scores in CLS, and $\mathcal{O}(N_r^2)$ coming from $\mathcal{K}_2$-medoids and greedy search in RPS and GPDB. Note that, in our case, $\mathcal{O}(n \log n + 2N_r^2) < \mathcal{O}(N_r n)$. The complexity of simple ranking-based baselines is $\mathcal{O}(n \log(n))$ due to sorting the sample acquisition scores. Comparing our method with recent state-of-the-arts, LLAL has the highest training complexity, and BADGE and CORESET have the highest selection complexity. Unlike the existing baseline, training and selection complexities of the proposed CRB are upper bounded by the reasonable asymptotic growth rates.

Table 4: AL Results with one-stage 3D detector SECOND.

	3D Detection mAP			BEV Detection mAP		
	EASY	MODERATE	HARD	EASY	MODERATE	HARD
RAND	75.23	60.83	56.55	80.20	67.56	63.30
LLAL	72.02	58.96	54.21	79.50	66.82	62.48
CORESET	74.74	58.86	54.61	79.71	65.53	61.39
BADGE	75.38	61.65	56.72	80.81	68.83	64.17
CRB	**78.96**	**64.27**	**59.60**	**83.28**	**70.49**	**66.09**

Table 5: Complexity Analysis.

AL Strategy	Training	Selection
RAND	$\mathcal{O}(En)$	$\mathcal{O}(N_r)$
ENTROPY	$\mathcal{O}(En)$	$\mathcal{O}(n \log n)$
MC-REG	$\mathcal{O}(En)$	$\mathcal{O}(n \log n)$
LLAL	$\mathcal{O}((E + E_l)n)$	$\mathcal{O}(n \log n)$
CORESET	$\mathcal{O}(En)$	$\mathcal{O}(N_r n)$
BADGE	$\mathcal{O}(En)$	$\mathcal{O}(N_r n)$
CRB	$\mathcal{O}(En)$	$\mathcal{O}(n \log n + 2N_r^2)$

4 DISCUSSION

This paper studies three novel criteria for sample-efficient active 3D object detection, that can effectively achieve high performance with minimum costs of 3D box annotations and runtime complexity. We theoretically analyze the relationship between finding the optimal acquired subset and mitigating the sets discrepancy. The framework is versatile and can accommodate existing AL strategies to provide in-depth insights into heuristic design. The limitation of this work lies in a set of assumptions made on the prior distribution of the test data, which could be violated in practice. For more discussions, please refer to Sec. A.1 in Appendix. In contrast, it opens an opportunity of adopting our framework for active domain adaptation, where the target distribution is accessible for alignment. Addressing these two avenues is left for future work.

REPRODUCIBILITY STATEMENT

The source code of the developed active 3D detection toolbox is available in the supplementary material, which accommodates various AL approaches and one-stage and two-stage 3D detectors. We specify the settings of hyper-parameters, the training scheme and the implementation details of our model and AL baselines in Sec. B of the supplementary material. We show the proofs of Theorem C.1 in Sec. C followed by the overview of the algorithm in Sec. D in the supplementary material. We repeat the experiments on the KITTI dataset 3 times with different initial labeled sets and show the standard deviation in plots and tables.

ETHICS STATEMENT

Our work may have a positive impact on communities to reduce the costs of annotation, computation, and carbon footprint. The high-performing AL strategy greatly enhances the feasibility and practicability of 3D detection in critical yet data-scarce fields such as medical imaging. We did not use crowdsourcing and did not conduct research with human subjects in our experiments. We cited the creators when using existing assets (*e.g.*, code, data, models).

In this appendix, we discuss the prior distribution selection, motivation of the Stage 2, and evaluation division of difficulty in Sec A.1 and Sec A.2, respectively. In the rest of the supplementary material, we provide the implementation details of all baselines and the proposed approach in Sec B followed by the proof of Theorem C.1. In Sec D, the overall algorithm is summarized. Additional experimental results on KITTI (Sec E) and Waymo (Sec F) datasets are reported and analyzed. We further conducted supplemental experiments on parameter sensitivity (Sec H) and visualizations (Sec G). In the end, we leave the related work and the associated discussion in Sec I.

A APPENDIX

A.1 MORE DISCUSSIONS ON PRIOR DISTRIBUTION

In mainstream 3D detection datasets, the curated test set is commonly long-tailed distributed, with a few head classes (*e.g.*, car) possessing a large number of samples and all the rest of the tail classes possessing only a few samples. As such, the trained detector can be easily biased towards head classes with massive training data, resulting in high accuracy on head classes and low accuracy on tail classes. This suggests that for 3D detection tasks, **mean average precision (mAP)** can be a **fairer** metric of evaluation, by taking an average of all AP values per class. When the test label is uniformly distributed, mAP scores will be equal to the AP scores for all samples. This motivates us to choose the uniform distribution as the prior distribution, rather than estimating the test label distribution from the initial labeled set $\mathcal{D}_L$. In this case, the trained model tends to be more robust and resilient to the imbalanced training data, achieving higher mAP scores.

To justify the effectiveness of choosing the uniform distribution, we provide more comparisons with the SOTA active learning methods in Table 7 and Table 6, which do not take the uniform distribution as an assumption. We clearly observe that such AL methods perform poorly on **tail classes** (*e.g.*, pedestrian and cyclist), confirming that the yielded models are biased towards learning car samples.

Table 6: Performance gap (%) between different AL methods and fully supervised backbone when acquiring approximately 1% queried bounding boxes on KITTI. Gaps are calculated by subtracting the performance of a fully supervised backbone from the performance of AL methods.

	Car (↓)			Pedestrian (↓)			Cyclist (↓)			Average (↓)		
Method	EASY	MOD.	HARD	EASY	MOD.	HARD	EASY	MOD.	HARD	EASY	MOD.	HARD
LLAL	2.61	5.71	**7.16**	7.92	6.80	5.94	13.33	11.60	11.42	7.81	8.04	8.18
CORESET	4.79	6.63	9.53	16.99	14.70	13.72	7.15	12.23	11.14	9.49	11.18	11.47
BADGE	2.60	8.58	11.94	12.32	10.43	10.93	4.77	9.66	8.66	6.41	9.55	10.51
CRB	**1.58**	**5.34**	8.44	**0.09**	**1.87**	**1.09**	**1.92**	**4.50**	**3.22**	**1.05**	**3.18**	**4.25**

Table 7: Performance comparisons on KITTI *val* set with different SOTA AL methods when acquiring approximately 1% queried bounding boxes. Results are reported with 3D AP with 40 recall positions. † indicates the reported performance of the backbone trained with the full labeled set (100%).

	Car			Pedestrian			Cyclist			Average		
Method	EASY	MOD.	HARD	EASY	MOD.	HARD	EASY	MOD.	HARD	EASY	MOD.	HARD
LLAL	89.95	78.65	**75.32**	56.34	49.87	45.97	75.55	60.35	55.36	73.94	62.95	58.88
CORESET	87.77	77.73	72.95	47.27	41.97	38.19	81.73	59.72	55.64	72.26	59.81	55.59
BADGE	89.96	75.78	70.54	51.94	46.24	40.98	84.11	62.29	58.12	75.34	61.44	56.55
CRB	**90.98**	**79.02**	74.04	**64.17**	**54.80**	**50.82**	**86.96**	**67.45**	**63.56**	**80.70**	**67.81**	**62.81**
PV-RCNN†	92.56	84.36	82.48	64.26	56.67	51.91	88.88	71.95	66.78	81.75	70.99	67.06

A.2 MORE DISCUSSIONS ON EVALUATION DIVISION OF DIFFICULTY

On the KITTI dataset, the evaluation difficulty is set based on the visual look[1] of the images, which is supposed to be unavailable for our LiDAR-based detection task. On the other hand, the Waymo dataset leverages a more reasonable and general setting of difficulty evaluation, with LEVEL 1 and LEVEL 2 difficulties indicating "more than five points" and "at least one point" inside labeled bounding boxes, respectively. This aligns with the design of the balance criterion (Stage 3), as the sparse point clouds or dense point clouds can be equally learned. In Table 8, we report the performance of the proposed approach with a small portion of point clouds and the fully supervised baseline reported in (Zhang et al., 2022), on the Waymo dataset. From Table 8, we can observe that the performance gap between the detectors trained with active learning (approx. 50K bounding box annotations) and fully supervised learning (approx. 8 million bounding box annotations) is smaller in LEVEL 2 (7.18% in LEVEL 2 vs 8.08% in LEVEL 1), which aligns with the balance criteria in the proposed CRB framework.

A.3 MORE DISCUSSIONS ON THE MOTIVATION OF THE STAGE 2

Our main objective of Stage 2, *i.e.*, Representative Prototype Selection is to determine a subset $\mathcal{D}^*_{S_2}$ from the pre-selected set S_1 in the last stage, by minimizing the set discrepancy in the latent feature space. However, the test features are not observable during the training phase, and it is hard to guarantee that the feature distribution can be comprehensively captured. As stated in Remark section, we focus on the features that are not learned well from the training set due to the zero training error assumption and reconsider the feature matching problem from a gradient perspective. In particular, we split the test set into two group In the gradient space: (1) seen test samples that can be easily recognized will cluster near the origin, (2) while the novel test samples will diversely distribute in the subspace. As the first group of samples have been sufficiently covered by the initiated, in this stage, we focus on finding matching with the latter group. By assuming the prior distribution of gradients follows a Gaussian distribution, finding the K-metroids is naturally a choice to mitigate the gap between mean and variance. K-metroids algorithm breaks the dataset up into groups and attempts to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster (*i.e.*, prototype). By selecting the prototypes in the second stage, we implicitly bridge the gap between the selected set and the test set at a latent feature level.

[1] http://www.cvlibs.net/datasets/kitti/eval_object.php

B IMPLEMENTATION DETAILS

B.1 EVALUATION METRICS.

To fairly evaluate baselines and the proposed method on KITTI dataset (Geiger et al., 2012), we follow the work of (Shi et al., 2020): we utilize Average Precision (AP) for 3D and bird eye view (BEV) detection, and the task difficulty is categorized to EASY, MODERATE, and HARD, with a rotated IoU threshold of 0.7 for cars and 0.5 for pedestrian and cyclists. The results evaluated on the validation split are calculated with 40 recall positions. To evaluate on Waymo dataset (Sun et al., 2020), we adopt the officially published evaluation tool for performance comparisons, which utilizes AP and the average precision weighted by heading (APH). The respective IoU thresholds for vehicles, pedestrians, and cyclists are set to 0.7, 0.5, and 0.5. Regarding detection difficulty, the Waymo test set is further divided into two levels. LEVEL 1 (and LEVEL 2) indicates there are more than five inside points (at least one point) in the ground-truth objects.

Table 8: Comparing the performance of detectors with active learning (AL) by CRB and fully supervised learning (FSL) on Waymo *val* set. Results (mAP $\%$) are calculated by Waymo official evaluation metric.

Method	mAP Level 1	mAP Level 2
CRB	58.60	52.65
FSL	66.68	59.83
Gap (↓)	-8.08	-7.18

B.2 IMPLEMENTATION DETAILS OF TRAINING

To ensure the reproducibility of the baselines and the proposed approach, we develop a PyTorch-based active 3D detection toolbox (attached in the supplemental material) that implements mainstream AL approaches and can accommodate most of the public benchmark datasets. For fair comparison, all active learning methods are constructed from the PV-RCNN (Shi et al., 2020) backbone. All experiments are conducted on a GPU cluster with three V100 GPUs. The runtime for an experiment on KITTI and Waymo is around 11 hours and 100 hours, respectively. Note that, training PV-RCNNon the full set typically requires 40 GPU hours for KITTI and 800 GPU hours for Waymo.

Parameter Settings. The batch sizes for training and evaluation are fixed to 6 and 16 on both datasets. The Adam optimizer is adopted with a learning rate initiated as 0.01, and scheduled by one cycle scheduler. The number of MC-DROPOUT stochastic passes is set to 5 for all methods.

Active Learning Protocols. As our work is the first comprehensive study on active 3D detection task, the active training protocol for all AL baselines and the proposed method is empirically defined. For all experiments, we first randomly select m fully labeled point clouds from the training set as the initial $\mathcal{D}_L$. With the annotated data, the 3D detector is trained with E epochs, which is then freezed to select N_r candidates from $\mathcal{D}_U$ for label acquisition. We set the m and N_r to 2.5 3% point clouds (*i.e.*, $N_r = m =$100 for KITTI, $N_r = m =$400 for Waymo) to trade-off between reliable model training and high computational costs.The aforementioned training and selection steps will alternate for R rounds. Empirically, we set $E = 30$, $R = 6$ for KITTI, and fix $E = 40$, $R = 5$ for Waymo.

B.3 IMPLEMENTATION DETAILS OF BASELINES AND CRB

In this section, we introduce more implementation details of both baselines and the proposed CRB.

CRB. In comparison with baselines as reported in Figure 2, the $\mathcal{K}_1$ and $\mathcal{K}_2$ are empirically set to $300, 200$ for KITTI and $2,000$ and $1,200$ for Waymo. The gradient maps used for RPS are extracted from the second convolutional layer in the shared block of PV-RCNN. Three dropout layers in PV-RCNN are enabled during the MC-DROPOUT and the dropout rate is fixed to 0.3 for both datasets. The number of MC-DROPOUT stochastic passes are set to 5 for all methods. In the GPCB stage, we measure the KL-divergence between the KDE PDF of the selected set and the uniform prior distribution of the point cloud density for each class. The goal of conducting a greedy search is to find the optimal subset that can achieve the minimum sum of KL divergence for all classes. Considering the high variance of KL divergence across different classes, we unify the scale of KL-divergence to $\bar{d}_c$ by applying the following function,

$$\bar{d}_c = \frac{2}{\pi} \arctan \frac{\pi}{2} d_c,$$

where d_c denotes the KL-divergence for the c-th class. To this end, the ultimate objective for greedy search is $\arg\min_{\mathcal{D}_S \subset \mathcal{D}_{S_2}} \sum_{c\in[C]} \bar{d}_c$. The normalized measurement can avoid dominance by any single class.

CORESET (Sener & Savarese, 2018). The embeddings extracted for both labeled and unlabeled data are the output from the shared block, with the dimension of 128 by 256. The CORESET adopts the furthest-first traversal for k-Center clustering strategy, which computes the Euclidean distance between each embedding pair.

LLAL (Yoo & Kweon, 2019). For implementing the loss prediction module in LLAL, we construct a two-block module that connects to two layers of the PV-RCNN, which takes multi-level knowledge into consideration for loss prediction. Particularly, each block consists of a convolutional layer with a channel size of 265 and a kernel size of 1, a batchnorm layer, and a relu activation layer. The outputs are then concatenated and fed to a fully connected layer and map to a loss score. All real loss for each training data point is saved and serves as the ground-truth to train the loss prediction module.

BADGE. According to (Ash et al., 2020), hypothetical labels for the classifier are determined by the classes with the highest predicted probabilities. The gradient matrix with the dimension 256 by 256 for each unlabeled point cloud is extracted from the last convolutional layer of the PV-RCNN's classification head and then fed into the BADGE algorithm.

C PROOF OF THEOREM 2.1

Theorem C.1. *Let $\mathcal{H}$ be a hypothesis space of Vapnik-Chervonenkis (VC) dimension d, with f and g being the classification and regression branches, respectively. The $\widehat{\mathcal{D}}_S$ and $\widehat{\mathcal{D}}_T$ represent the empirical distribution induced by samples drawn from the acquired subset $\mathcal{D}_S$ and the test set $\mathcal{D}_T$, and ℓ the loss function bounded by $\mathcal{J}$. It is proven that $\forall\, \delta \in (0,1)$, and $\forall f, g \in \mathcal{H}$, with probability at least $1-\delta$ the following inequality holds,*

$$\mathfrak{R}_T[\ell(f,g;\boldsymbol{w})] \leq \mathfrak{R}_S[\ell(f,g;\boldsymbol{w})] + \frac{1}{2} disc(\widehat{\mathcal{D}}_S, \widehat{\mathcal{D}}_T) + \lambda^* + const,$$

where $const = 3\mathcal{J}(\sqrt{\frac{\log\frac{4}{\delta}}{2N_r}} + \sqrt{\frac{\log\frac{4}{\delta}}{2N_t}}) + \sqrt{\frac{2d\log(eN_r/d)}{N_r}} + \sqrt{\frac{2d\log(eN_t/d)}{N_t}}$.

Notably, $\lambda^* = \mathfrak{R}_T[\ell(f^*, g^*; \boldsymbol{w}^*)] + \mathfrak{R}_S[\ell(f^*, g^*; \boldsymbol{w}^*)]$ denotes the joint risk of the optimal hypothesis f^* and g^*, with $\boldsymbol{w}^*$ being the model weights. N_r and N_t indicate the number of samples in the $\mathcal{D}_S$ and $\mathcal{D}_T$. The proof can be found in the supplementary material.

Proof. For brevity, we omit the model weights $\boldsymbol{w}$ in the following proof. Based on the triangle inequality of ℓ and the definition of the discrepancy distance $disc(\cdot,\cdot)$, the following inequality holds,

$$\begin{aligned}
\mathfrak{R}_T[\ell(f,g)] &\leq \mathfrak{R}_T[\ell(f^*,g^*)] + \frac{1}{2}\mathfrak{R}_T[\ell(f,f^*)] + \frac{1}{2}\mathfrak{R}_T[\ell(g,g^*)] \\
&\leq \mathfrak{R}_T[\ell(f^*,g^*)] + \mathfrak{R}_S[\ell(f^*,g^*)] + \frac{1}{2}|\mathfrak{R}_T[\ell(f,f^*)] - \mathfrak{R}_S[\ell(f,f^*)]| \\
&+ \frac{1}{2}|\mathfrak{R}_T[\ell(g,g^*)] - \mathfrak{R}_S[\ell(g,g^*)]| \\
&\leq \mathfrak{R}_T[\ell(f^*,g^*)] + \mathfrak{R}_S[\ell(f^*,g^*)] + \frac{1}{2} disc(\mathcal{D}_S, \mathcal{D}_T) \\
&\leq \mathfrak{R}_T[\ell(f^*,g^*)] + \mathfrak{R}_S[\ell(f,g)] + \mathfrak{R}_S[\ell(f^*,g^*)] + \frac{1}{2} disc(\mathcal{D}_S, \mathcal{D}_T).
\end{aligned}$$

Algorithm 1 The algorithm of CRB for active 3D object detection

Inputs:
$\mathcal{D}_L$: initially labeled point clouds
$\mathcal{D}_U$: unlabeled pool of point clouds
$\mathbf{\Omega}$: oracle
B: total budget of active selection
$e(\cdot) : \mathcal{P} \rightarrow \boldsymbol{x}$: point cloud encoder of 3D detector
$f(\cdot)$: classifier of 3D detector
$g(\cdot)$: regression head of 3D detector
R: total active learning rounds

$\mathcal{D}_S \leftarrow \emptyset$
Pre-train 3D detector $\{e(\cdot), f(\cdot), g(\cdot)\}$ with $\mathcal{D}_L$ until converge
$\mathcal{D}_S \leftarrow \mathcal{D}_S \cup \mathcal{D}_L$
for $r \in [R]$ **do** ▷ For each round of active selection
  $\widehat{Y} \leftarrow f \circ e(\mathcal{D}_U)$ ▷ Get the predicted labels
  $\widehat{\mathcal{B}}, \phi \leftarrow g \circ e(\mathcal{D}_U)$ ▷ Get the predicted boxes $\widehat{\mathcal{B}}$ and box point densities ϕ
  $\bar{B} \leftarrow$ Hypothetical labels computed by Equation (5)
  $\mathcal{D}^*_{S_1} \leftarrow \textsc{Cls}(\mathcal{D}_U, \widehat{Y})$ selects $\mathcal{K}_1$ samples via Equation (2),(3),(4)
  ▷ Stage 1: Concise Label Sampling
  $\mathcal{D}^*_{S_2} \leftarrow \textsc{Rps}(\mathcal{D}^*_{S_1}, e(\mathcal{D}^*_{S_1}), \bar{B})$ selects $\mathcal{K}_2$ samples via Equation (6),(7)
  ▷ Stage 2: Representative Prototype Selection
  $\mathcal{D}^*_S \leftarrow \textsc{Gpdb}(\mathcal{D}^*_{S_2}, \widehat{\mathcal{B}}, \phi)$ selects N_r samples via Equation (8),(9)
  ▷ Stage 3: Greedy Point Cloud Density Balancing
  $\mathcal{D}^*_S \leftarrow \mathbf{\Omega}(\mathcal{D}^*_S)$ ▷ Query labels from oracles
  $\mathcal{D}_S \leftarrow \mathcal{D}_S \cup \mathcal{D}^*_S$
  $\mathcal{D}_U \leftarrow \mathcal{D}_U \backslash \mathcal{D}^*_S$
  Train 3D detector $\{e(\cdot), f(\cdot), g(\cdot)\}$ with $\mathcal{D}_S$ until converge
end for

By defining the joint risk of the optimal hypothesis $\lambda^* = \mathfrak{R}_T[\ell(f^*, g^*)] + \mathfrak{R}_S[\ell(f^*, g^*)]$ and the Corollary 6 in (Mansour et al., 2009), we have,

$$\begin{aligned}\mathfrak{R}_T[\ell(f,g)] &\leq \mathfrak{R}_S[\ell(f,g)] + \frac{1}{2} disc(\mathcal{D}_S, \mathcal{D}_T) + \lambda^* \\ &\leq \mathfrak{R}_S[\ell(f,g)] + \frac{1}{2} disc(\widehat{\mathcal{D}}_S, \widehat{\mathcal{D}}_T) + \lambda^* + 4q(\text{Rad}_S(\mathcal{H}) + \text{Rad}_T(\mathcal{H})) \\ &+ 3\mathcal{J}(\sqrt{\frac{\log\frac{4}{\delta}}{2N_r}} + \sqrt{\frac{\log\frac{4}{\delta}}{2N_t}}),\end{aligned}$$

where N_r and N_t indicate the sample size of the selected set and the test set, respectively. q stands for the function is q-Lipschitz. As our regression loss, ℓ^{reg} is the smooth-L1 loss function and bounded by $\mathcal{J}$, q equals 1 in our case. $\text{Rad}_S(\mathcal{H})$ and $\text{Rad}_T(\mathcal{H})$ indicates the empirical Rademacher complexity of a hypothesis set $\mathcal{H}$ whose VC dimension is d over the selected set and the test set.

Considering the Rademacher complexity is bounded by:

$$\text{Rad}_S(\mathcal{H}) \leq \sqrt{\frac{2d\log(eN_r/d)}{N_r}}, \quad \text{Rad}_T(\mathcal{H}) \leq \sqrt{\frac{2d\log(eN_t/d)}{N_t}},$$

then we can rewrite the inequality as,

$$\mathfrak{R}_T[\ell(f,g)] \leq \mathfrak{R}_S[\ell(f,g)] + \frac{1}{2} disc(\widehat{\mathcal{D}}_S, \widehat{\mathcal{D}}_T) + \lambda^* + const,$$

where $const = 3\mathcal{J}(\sqrt{\frac{\log\frac{4}{\delta}}{2N_r}} + \sqrt{\frac{\log\frac{4}{\delta}}{2N_t}}) + \sqrt{\frac{2d\log(eN_r/d)}{N_r}} + \sqrt{\frac{2d\log(eN_t/d)}{N_t}}$. □

D ALGORITHM DESCRIPTION

To thoroughly describe the procedure of active 3D object detection by the proposed CRB, we present the Algorithm 1 in detail. Firstly, the 3D detector consisting of an encoder $\{e(\cdot)$, a classifier $f(\cdot)$, and regression heads $g(\cdot)\}$ is pre-trained with a small set $\mathcal{D}_L$ of labeled point clouds. During the stage 1: CLS, the pre-trained 3D detector infers all samples from the unlabeled pool $\mathcal{D}_U$ and obtains the predicted bounding boxes $\widehat{\mathcal{B}}$, predicted box labels $\widehat{Y}$, and calculated box point densities ϕ for each point cloud. Also, the hypothetical labels $\bar{B}$ through stochastic monte-carlo sampling are computed by Equation (5) during inference. Based on the criterion of maximizing the label entropy, the set $\mathcal{D}^*_{S_1}$ containing $\mathcal{K}_1$ candidates are formed via Equations (2), (3), (4). In stage 2, we set the model to the training mode and allow the gradient back-propagation to retrieve gradients for each point cloud. Yet, the model weights will be fixed and not updated. RPS selects the set $\mathcal{D}^*_{S_2}$ of size $\mathcal{K}_2$ from the previous candidate set $\mathcal{D}^*_{S_1}$ based on the Equation (6) and (7). In stage 3, GBPS selects the set $\mathcal{D}^*_S$ of size N_r from set $\mathcal{D}^*_{S_2}$, predicted boxes $\widehat{\mathcal{B}}$ and box point densities ϕ via Equation (8), (9). The final set $\mathcal{D}^*_S$ at this round is then annotated by an oracle $\mathbf{\Omega}$ and merged with the selected set in the previous round as the training data. Notably, the selected set at the 0-th round is $\mathcal{D}_L$. When the training data is determined, we re-train the 3D detector with the merged selected set until the model is converged. We iterate the above process starting with model inference for R rounds and add N_r queried samples to the selected set $\mathcal{D}_S$ for each round.

E MORE EXPERIMENTAL RESULTS ON KITTI

E.1 AL PERFORMANCE COMPARISONS ON EASY DIFFICULTY LEVEL

In addition to the MODERATE and HARD difficulties reported in the body text, we provide the additional quantitative analysis *w.r.t.* the EASY mode. Figure 5 depicts the mAP(%) variation of the baselines against the proposed CRB with an increasing number of selected bounding boxes. The solid lines indicate the mean value from three running trials and the standard deviation are shown in the shaded area. The results indicate that with increasing annotation cost, CRB consistently achieves the highest mAP and outperforms the state-of-the-art active learning approaches on both 3D and BEV views. Note that CRB with only 1k boxes selected for annotation reaches the comparable performance of RAND that selects around 3k boxes. Other AL baselines share the same trend as the ones under the difficulties of MODERATE and HARD (reported in Figure 2 of the main paper).

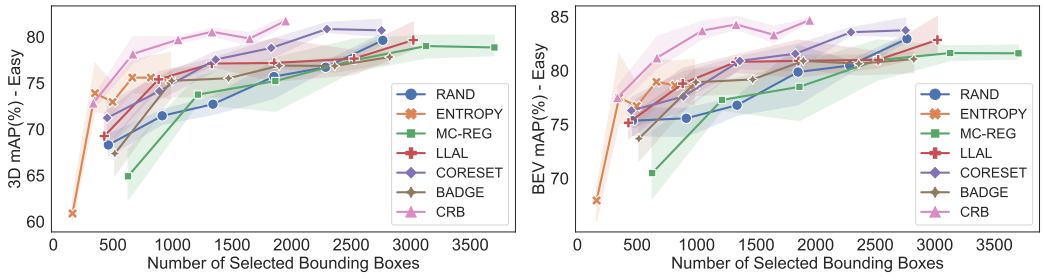

Figure 5: 3D and BEV mAP (%) of CRB and AL baselines on the KITTI *val* split at the EASY level.

E.2 AL PERFORMANCE COMPARISONS FOR EACH CLASS

To investigate the effectiveness of AL strategies on detecting specific classes, we plot the results of Cyclist and Pedestrian at all difficulty levels in Figure 6 (3D AP) and Figure 7 (BEV AP). We mainly compare three aspects: performance, annotation cost and error variance. 1)Performance: the plots in Figure 6 and Figure 7 show that the proposed CRB outperforms all state-of-the-art AL methods by a noticeable margin, for all settings of difficulty, classes and views, except at easy cyclist. This evidences that our proposed AL approach explores samples with more conceptual semantics covering test sets so that the detector tends to perform better on more challenging samples. 2) Annotation cost: all the plots consistently demonstrate that the proposed CRB reaches comparable performance

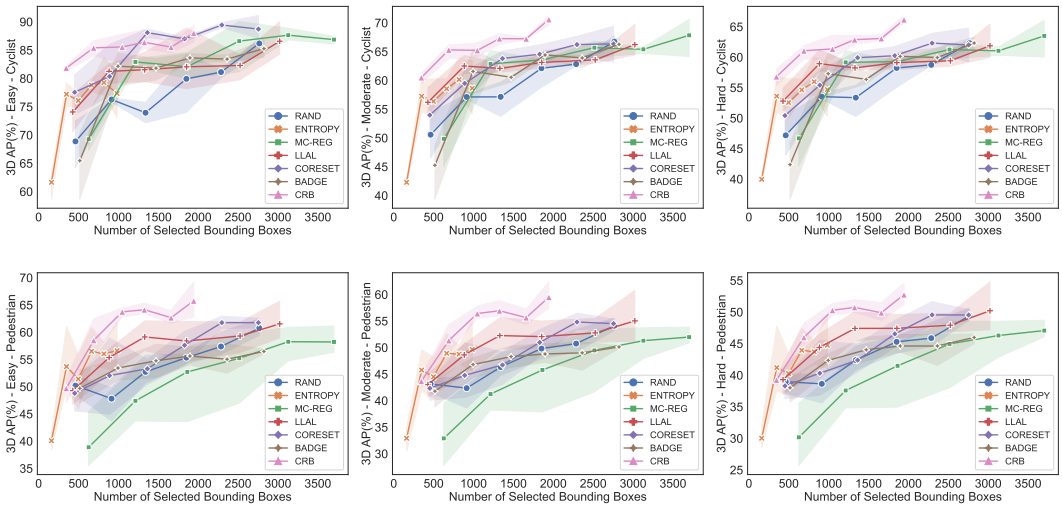

Figure 6: Detection results of different classes on the KITTI *val* set (3D view) with an increasing number of queried bounding boxes.

while requiring very few ($\sim$1/3) annotation costs as baselines, except ENTROPY. ENTROPY takes the minimal annotation cost, yet its result is inferior, especially for difficult classes like Cyclist. 3) Variance: we observe that AP variance of CRB is lower than all baselines, which shows that our method is less sensitive to randomness and more stable to produce expected results.

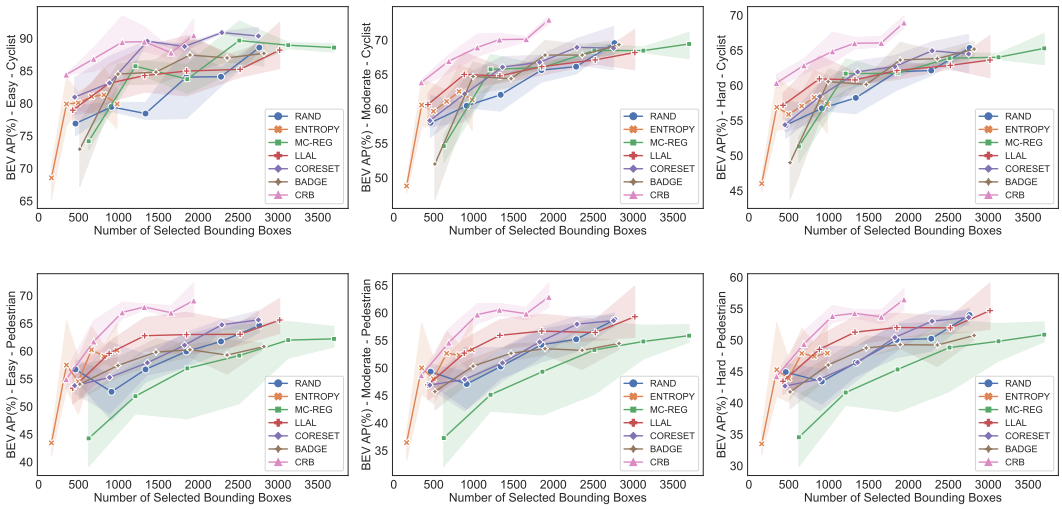

Figure 7: Detection results of different classes on the KITTI *val* set (BEV view) with increasing number of queried bounding boxes.

E.3 AL PERFORMANCE COMPARISONS FOR EACH ACTIVE SELECTION ROUND

Figure 8 compares the performance variation of the AL baselines against the proposed CRB with the increasing percentage of queried point clouds (from 2.7% to 16.2%). The reported performance is mAP scores (%) $\pm$ the standard deviation of three trials for both 3D view (top row) and BEV view (bottom row) and all difficulty levels. We clearly observe that our method CRB consistently outperforms the state-of-the-art results, irrespective of percentage of annotated point clouds and difficulty settings. Surprisingly, when the annotation costs reaches 16.2%, RAND strategy outperforms all the baselines at the MODERATE and HARD level. This implicitly evidences that existing uncertainty and diversity-based AL strategies fail to select samples that are aligned with test cases.

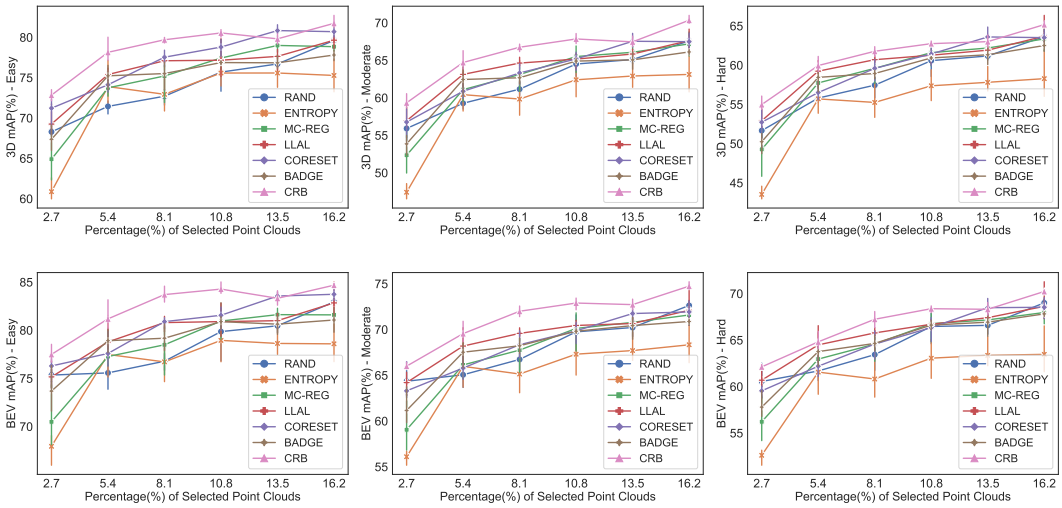

Figure 8: Results on KITTI datasets with an increasing percentage of queried point clouds.

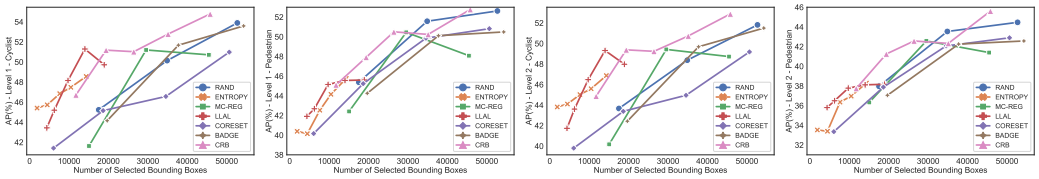

Figure 9: Results of CRB and baselines on the Waymo *val* splt for different classes at Level 2.

F MORE EXPERIMENTAL RESULTS ON WAYMO

To explore the performance for different classes on the Waymo dataset, we plot the AP(%) variation of Cyclist and Pedestrian yielded by the baselines and CRB with increasing annotated bounding boxes in Figure 9. We present the results at two levels of difficulty officially defined by Waymo. LEVEL 1 (and LEVEL 2) indicates there are more than five inside points (at least one point) of the ground-truth objects. As can be observed by the AP curves in the plots, CRB achieves the superiror recognition accuracy when the annotation cost comes to $\sim$ 45k bounding boxes. Specifically, the AP values of CRB are boosted by the largest margin (3.1% on LEVEL 2 Cyclist and 1.6% on LEVEL 2 Pedestrian) over the best performing baseline (RAND) that takes extra cost of 5k bounding boxes than ours. Surprisingly, note the results on the class of Pedestrian, the AP curves of most baselines except ENTROPY and LLAL are bounded by RAND. The AP curves of ENTROPY and LLAL are bounded by CRB with the increasing cost to 15k $\sim$ 20k bounding boxes. This confirms the CRB's superiority over compared AL baselines. Besides, the boosted margin achieved by CRB set for LEVEL 2 Pedestrian is larger than set for LEVEL 1 Pedestrian. This indicates that the samples selected by CRB matches well with the data at the time, covering more diverse samples that span different difficulties.

G ADDITIONAL QUALITATIVE ANALYSIS

To intuitively demonstrate the benefits of our proposed active 3D detection strategy, Figure 10 visualizes that the 3D detection results produced by **RAND** (bottom left) and **CRB** selection (bottom right) from the corresponding image (upper row). Both 3D detectors are trained under the budget of 1K annotated bounding boxes. False positives and corrected predictions are indicated with red and green boxes. It is observed that, under the same condition, CRB produces more accurate and more confident predictions than RAND. Specifically, our CRB yields accurate predictions for multiple

pedestrians on the right sidewalk, while RAND fails. Besides, note the car parked on the left that is highlighted in the orange box in Figure 10, the detector trained with RAND produces a significantly lower confidence score (0.62) compared to our approach (0.95). This validates that the point clouds selected by CRB are aligned more tightly with the test samples.

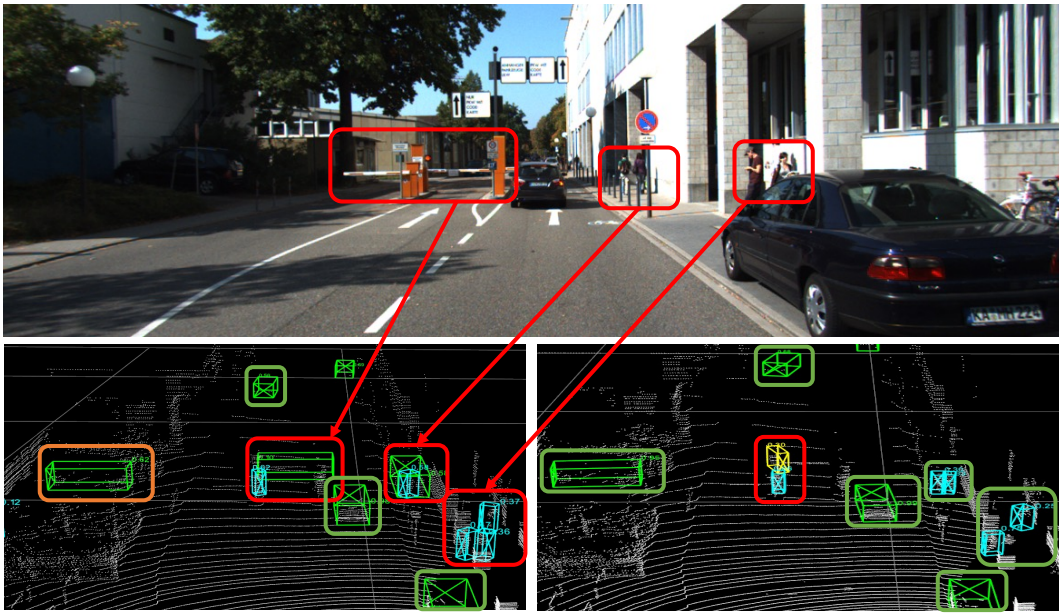

Figure 10: Another case study of active 3D detection performance of **RAND** (bottom left) and **CRB** (bottom right) under the budge of 1,000 annotated bounding boxes. False positive (corrected predictions) are highlighted in red (green) boxes. The orange box denotes the detection with low confidence.

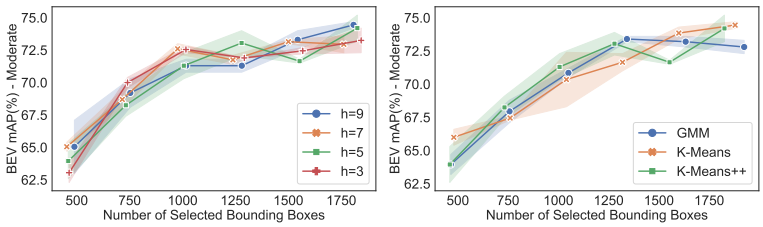

Figure 11: Performance comparison on KITTI *val* set with varying KDE bandwidth h (left) and prototype selection approaches (right) with increasing queried bounding boxes.

H ADDITIONAL RESULTS FOR PARAMETER SENSITIVITY ANALYSIS

Sensitivity to Prototype Selection. To further analyze the sensitivity of performance to different prototype selection approaches, *i.e.*, GMM, K-MEANS, and K-MEANS++, we show more results on BEV views in Figure 11 (right). We again run two trials for each prototype selection method and plot the mean and the variance bars. Note that there is very little difference (1.65% in the last round) in the mAP(%) of our approach when using different prototype selection methods. This evidences that the more performance gains achieved by CRB than existing baselines do not depend on choosing the prototype selection method.

Sensitivity to Bandwidth h. Figure 11 shows additional results w.r.t the BEV views of CRB with the bandwidth h varying in $\{3, 5, 7, 9\}$. Observing the trends of four curves, CRB with the bandwidth of all values yields consistent results within the 1.7% variation. This demonstrates that the CRB is insensitive to different values set for bandwidth and can produce similar mAP(%) on BEV views.

I RELATED WORK

Generic Active Learning. For a comprehensive review of classic active learning methods and their applications, we refer readers to (Ren et al., 2021). Most active learning approaches were tailored for image classification task, where the *uncertainty* (Wang & Shang, 2014;?; Lewis & Catlett, 1994; Joshi et al., 2009; Roth & Small, 2006; Parvaneh et al., 2022; Du et al., 2021; Kim et al., 2021b; Bhatnagar et al., 2021) and *diversity* (Sener & Savarese, 2018; Elhamifar et al., 2013; Guo, 2010; Yang et al., 2015; Nguyen & Smeulders, 2004; Hasan & Roy-Chowdhury, 2015; Aodha et al., 2014) of samples are measured as the acquisition criteria. The hybrid works (Kim et al., 2021a; Citovsky et al., 2021; Ash et al., 2020; MacKay, 1992; Liu et al., 2021; Kirsch et al., 2019; Houlsby et al., 2011) combine both paradigms such as by measuring uncertainty as to the gradient magnitude (Ash et al., 2020) at the final layer of neural networks and selecting gradients that span a diverse set of directions. In addition to the above two mainstream methods, (Settles et al., 2007; Roy & McCallum, 2001b; Freytag et al., 2014; Yoo & Kweon, 2019) estimate the expected model changes or predicted losses as the sample importance.

Active Learning for 2D Detection. Lately, the attention of AL has shifted from image classification to the task of object detection (Siddiqui et al., 2020; Li & Yin, 2020). Early work (Roy et al., 2018) exploits the detection inconsistency of outputs among different convolution layers and leverages the query by committee approach to select informative samples. Concurrent work (Kao et al., 2018) introduces the notion of localization tightness as the regression uncertainty, which is calculated by the overlapping area between region proposals and the final predictions of bounding boxes. Other uncertainty-based methods attempt to aggregate pixel-level scores for each image (Aghdam et al., 2019), reformulate detectors by adding Bayesian inference to estimate the uncertainty (Harakeh et al., 2020) or replace conventional detection head with the Gaussian mixture model to compute aleatoric and epistemic uncertainty (Choi et al., 2021). A hybrid method (Wu et al., 2022) considers image-level uncertainty calculated by entropy and instance-level diversity measured by the similarity to the prototypes. Lately, AL technique is leveraged for transfer learning by selecting a few uncertain labeled source bounding boxes with high transferability to the target domain, where the transferability is defined by domain discriminators (Tang et al., 2021b; Al-Saffar et al., 2021). Inspired by neural architecture searching, Tang et al. (2021a) adopted the 'swap-expand' strategy to seek a suitable neural architecture including depth, resolution, and receptive fields at each active selection round. Recently, some works augment the weakly-supervised object detection (WSOD) with an active learning scheme. In WSOD, only image-level category labels are available during training. Some conventional AL methods such as predicted probability, probability margin are explored in (Wang et al., 2022), while in (Vo et al., 2022), "box-in-box" is introduced to select images where two predicted boxes belong to the same category and the small one is "contained" in the larger one. Nevertheless, it is not trivial to adapt all existing AL approaches for 2D detection as the ensemble learning and network modification leads to more model parameters to learn, which could be hardly affordable for 3D tasks.

Active Learning for 3D Detection. Active learning for 3D object detection has been relatively under-explored than other tasks, potentially due to its large-scale nature. Most existing works (Feng et al., 2019; Schmidt et al., 2020) simply apply the off-the-shelf generic AL strategies and use hand-crafted heuristics including Shannon entropy (Wang & Shang, 2014), ensemble (Beluch et al., 2018), localization tightness (Kao et al., 2018) and MC-DROPOUT (Gal & Ghahramani, 2016) for 3D detection learning. However, the abovementioned solutions base on the cost of labelling point clouds rather than the number of 3D bounding boxes, which inherently being biased to the point clouds containing more objects. However, in our work, the proposed CRB greedily search for the unique point clouds while maintaining the same marginal distribution for generalization, which implicitly quires objects to annotate without repetition and save labeling costs.

Active Learning for 3D Semantic Segmentation. The adoption of active learning techniques has successfully reduced the significant burden of point-by-point human labeling in large-scale point cloud datasets. Super-point (Shi et al., 2021) is introduced to represent a spectral clustering containing points which are most likely belonging to the same category, then only super-points with high score are labeled at each round. An improved work Shao et al. (2022) further encoded the super-points with a graph neural network, where the edges denote distance between super-points, and then projects the super-point features into the diversity space to select the most representative super-points. Another streaming of work (Wu et al., 2021) is to obtain point labels for uncertain

and diverse regions to prevent the high cost of labeling the entire point cloud. Although semantic segmentation and object detection are different vision tasks, both can benefit from active learning to substantially alleviate the manual labelling cost.

Connections to Semi-supervised Active Learning. Aiming at unifying unlabeled sample selection and model training, the concept of semi-supervised active learning (Drugman et al., 2016; Rhee et al., 2017; Sinha et al., 2019; Gao et al., 2020; Liu et al., 2021; Kim et al., 2021a;b; Zhang & Plank, 2021; Guo et al., 2021; Caramalau et al., 2021; Citovsky et al., 2021; Elezi et al., 2022; Gudovskiy et al., 2020) has been raised. (Drugman et al., 2016) combines the semi-supervised learning (SSL) and active learning (AL) for speech understanding that leverages the confidence score obtained from the posterior probabilities of decoded texts. (Sener & Savarese, 2018) incorporated a Ladder network for SSL during AL cycles, while the performance gains are marginal compared to the supervised counterpart. (Sinha et al., 2019) trained a variational adversarial active learning (VAAL) model with both labeled and unlabeled data points, where the discriminator is able to estimate how representative each sample is from the pool. (Elezi et al., 2022) proposed a combined strategy for training 2D object detection, which queries samples of high uncertainty and low robustness for supervised learning and takes full advantage of easy samples via auto-labeling. As our work is under the umbrella of the pool-based active learning, accessible unlabeled data are not used for model training in our setting, hereby the semi-supervised active learning algorithms were not considered in experimental comparisons.

REFERENCES

Hamed H. Aghdam, Abel Gonzalez-Garcia, Joost van de Weijer, and Antonio M. Lopez. Active learning for deep detection neural networks. In *Proc. International Conference on Computer Vision (ICCV)*, pp. 3672–3680, 2019.

Syeda Mariam Ahmed, Yan Zhi Tan, Chee-Meng Chew, Abdullah Al Mamun, and Fook Seng Wong. Edge and corner detection for unorganized 3d point clouds with application to robotic welding. In *Proc. International Conference on Intelligent Robots and Systems (IROS)*, pp. 7350–7355, 2018.

Ahmed Al-Saffar, Alina Bialkowski, Mahsa Baktashmotlagh, Adnan Trakic, Lei Guo, and Amin M. Abbosh. Closing the gap of simulation to reality in electromagnetic imaging of brain strokes via deep neural networks. *IEEE Transactions on Computational Imaging*, 7:13–21, 2021.

Oisin Mac Aodha, Neill D. F. Campbell, Jan Kautz, and Gabriel J. Brostow. Hierarchical subquery evaluation for active learning on a graph. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 564–571, 2014.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *Proc. International Conference on Learning Representations (ICLR)*, 2020.

William H. Beluch, Tim Genewein, Andreas Nürnberger, and Jan M. Köhler. The power of ensembles for active learning in image classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9368–9377, 2018.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Journal of Machine Learning*, 79 (1-2):151–175, 2010.

Shubhang Bhatnagar, Sachin Goyal, Darshan Tank, and Amit Sethi. PAL : Pretext-based active learning. In *Proc. British Machine Vision Conference (BMVC)*, pp. 195. BMVA Press, 2021.

Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Sequential graph convolutional network for active learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9583–9592, 2021.

Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clément Farabet, and Jose M. Alvarez. Active learning for deep object detection via probabilistic modeling. In *Proc. International Conference on Computer Vision (ICCV)*, pp. 10244–10253, 2021.

Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. Batch active learning at scale. In *Proc. Annual Conference on Neural Information Processing (NeurIPS)*, pp. 11933–11944, 2021.

Boyang Deng, Charles R. Qi, Mahyar Najibi, Thomas A. Funkhouser, Yin Zhou, and Dragomir Anguelov. Revisiting 3d object detection from an egocentric perspective. In *Proc. Annual Conference on Neural Information Processing (NeurIPS)*, pp. 26066–26079, 2021.

Thomas Drugman, Janne Pylkkönen, and Reinhard Kneser. Active and semi-supervised learning in ASR: benefits on the acoustic and language models. In Nelson Morgan (ed.), *Interspeech Annual Conference of the International Speech Communication Association*, pp. 2318–2322, 2016.

Pan Du, Suyun Zhao, Hui Chen, Shuwen Chai, Hong Chen, and Cuiping Li. Contrastive coding for active learning under class distribution mismatch. In *Proc. International Conference on Computer Vision (ICCV)*, pp. 8907–8916, 2021.

Ismail Elezi, Zhiding Yu, Anima Anandkumar, Laura Leal-Taixe, and Jose M Alvarez. Not all labels are equal: Rationalizing the labeling costs for training object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14492–14501, 2022.

Ehsan Elhamifar, Guillermo Sapiro, Allen Y. Yang, and S. Shankar Sastry. A convex optimization framework for active learning. In *Proc. International Conference on Computer Vision (ICCV)*, pp. 209–216, 2013.

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

Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Information, prediction, and query by committee. In *Proc. Annual Conference on Neural Information Processing (NeurIPS)*, pp. 483–490, 1992.

Alexander Freytag, Erik Rodner, and Joachim Denzler. Selecting influential examples: Active learning with expected model output changes. In *Proc. European Conference on Computer Vision (ECCV)*, pp. 562–577, 2014.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proc. International Conference on Machine Learning (ICML)*, volume 48, pp. 1050–1059, 2016.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proc. International Conference on Machine Learning (ICML)*, volume 70, pp. 1183–1192, 2017.

Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ömer Arik, Larry S. Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *Proc. European Conference on Computer Vision (ECCV)*, volume 12355, pp. 510–526, 2020.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3354–3361, 2012.

Denis A. Gudovskiy, Alec Hodgkinson, Takuya Yamaguchi, and Sotaro Tsukizawa. Deep active learning for biased datasets via fisher kernel self-supervision. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9038–9046, 2020.

Jiannan Guo, Haochen Shi, Yangyang Kang, Kun Kuang, Siliang Tang, Zhuoren Jiang, Changlong Sun, Fei Wu, and Yueting Zhuang. Semi-supervised active learning for semi-supervised models: Exploit adversarial examples with graph-based virtual labels. In *Proc. International Conference on Computer Vision (ICCV)*, pp. 2876–2885. IEEE, 2021.

Yuhong Guo. Active instance sampling via matrix partition. In *Proc. Annual Conference on Neural Information Processing (NeurIPS)*, pp. 802–810, 2010.

Ali Harakeh, Michael Smart, and Steven L. Waslander. Bayesod: A bayesian approach for uncertainty estimation in deep object detectors. In *Proc. International Conference on Robotics and Automation (ICRA)*, pp. 87–93, 2020.

Mahmudul Hasan and Amit K. Roy-Chowdhury. Context aware active learning of activity recognition models. In *Proc. International Conference on Computer Vision (ICCV)*, pp. 4543–4551, 2015.

Neil Houlsby, Ferenc Huszar, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *CoRR*, abs/1112.5745, 2011.

Ajay J. Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2372–2379, 2009.

Chieh-Chi Kao, Teng-Yok Lee, Pradeep Sen, and Ming-Yu Liu. Localization-aware active learning for object detection. In *Proc. Asian Conference on Computer (ACCV)*, pp. 506–522, 2018.

Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *(e)Proceedings of International Conference on Very Large Data Bases ((VLDB)*, pp. 180–191. Morgan Kaufmann, 2004.

Kwanyoung Kim, Dongwon Park, Kwang In Kim, and Se Young Chun. Task-aware variational adversarial active learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8166–8175, 2021a.

Yoon-Yeong Kim, Kyungwoo Song, JoonHo Jang, and Il-Chul Moon. LADA: look-ahead data acquisition via augmentation for deep active learning. In *Proc. Annual Conference on Neural Information Processing (NeurIPS)*, pp. 22919–22930, 2021b.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *Proc. Annual Conference on Neural Information Processing (NeurIPS)*, pp. 7024–7035, 2019.

David D. Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Proc. International Conference on Machine Learning (ICML)*, pp. 148–156, 1994.

Haohan Li and Zhaozheng Yin. Attention, suggestion and annotation: A deep active learning framework for biomedical image segmentation. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz (eds.), *Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 12261, pp. 3–13, 2020.

Zhuoming Liu, Hao Ding, Huaping Zhong, Weijia Li, Jifeng Dai, and Conghui He. Influence selection for active learning. In *Proc. International Conference on Computer Vision (ICCV)*, pp. 9254–9263, 2021.

Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. In *Proc. International Conference on Learning Representations (ICLR)*, 2021.

David J. C. MacKay. Information-based objective functions for active data selection. *Journal of Neural Computation*, 4(4):590–604, 1992.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Proc. Conference on Learning Theory (COLT)*, 2009.

Hector A. Montes, Justin Le Louedec, Grzegorz Cielniak, and Tom Duckett. Real-time detection of broccoli crops in 3d point clouds for autonomous robotic harvesting. In *Proc. International Conference on Intelligent Robots and Systems (IROS)*, pp. 10483–10488, 2020.

Hieu Tat Nguyen and Arnold W. M. Smeulders. Active learning using pre-clustering. In Carla E. Brodley (ed.), *Proc. International Conference on Machine Learning (ICML)*, 2004.

Adam M. Oberman and Jeff Calder. Lipschitz regularized deep neural networks converge and generalize. *CoRR*, abs/1808.09540, 2018.

Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Gholamreza (Reza) Haffari, Anton van den Hengel, and Javen Qinfeng Shi. Active learning by feature mixing. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12237–12246, 2022.

Robert Pinsler, Jonathan Gordon, Eric T. Nalisnick, and José Miguel Hernández-Lobato. Bayesian batch active learning as sparse subset approximation. In *Proc. Annual Conference on Neural Information Processing (NeurIPS)*, pp. 6356–6367, 2019.

Jun Qi, Jun Du, Sabato Marco Siniscalchi, Xiaoli Ma, and Chin-Hui Lee. On mean absolute error for deep neural network based vector-to-vector regression. *IEEE Signal Processing Letters*, 27: 1485–1489, 2020.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Computing Survey*, 54(9):40, 2021.

Phill-Kyu Rhee, Enkhbayar Erdenee, Shin Dong Kyun, Minhaz Uddin Ahmed, and SongGuo Jin. Active and semi-supervised learning for object detection with imperfect data. *Cognition System Research*, 45:109–123, 2017.

Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *Proc. European Conference on Machine Learning (ECML)*, pp. 413–424, 2006.

Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proc. International Conference on Machine Learning (ICML)*, pp. 441–448, 2001a.

Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. In *Proc. International Conference on Machine Learning (ICML)*, pp. 441–448, 2001b.

Soumya Roy, Asim Unmesh, and Vinay P. Namboodiri. Deep active learning for object detection. In *Proc. British Machine Vision Conference (BMVC)*, pp. 91, 2018.

Sebastian Schmidt, Qing Rao, Julian Tatsch, and Alois C. Knoll. Advanced active learning strategies for object detection. In *Proc. Intelligent Vehicles Symposium, (IV)*, pp. 871–876, 2020.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *Proc. International Conference on Learning Representations (ICLR)*, 2018.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In *Proc. Annual Conference on Neural Information Processing (NeurIPS)*, pp. 1289–1296, 2007.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Feifei Shao, Yawei Luo, Ping Liu, Jie Chen, Yi Yang, Yulei Lu, and Jun Xiao. Active learning for point cloud semantic segmentation via spatial-structural diversity reasoning. In *Proc. International Conference on Multimedia (MM)*, pp. 2575–2585, 2022.

Feng Shi and Yu-Feng Li. Rapid performance gain through active model reuse. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3404–3410, 2019.

Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–779. Computer Vision Foundation / IEEE, 2019.

Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN: point-voxel feature set abstraction for 3d object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10526–10535, 2020.

Weishi Shi and Qi Yu. Integrating bayesian and discriminative sparse kernel machines for multi-class active learning. In *Proc. Annual Conference on Neural Information Processing (NeurIPS)*, pp. 2282–2291, 2019.

Xian Shi, Xun Xu, Ke Chen, Lile Cai, Chuan Sheng Foo, and Kui Jia. Label-efficient point cloud semantic segmentation: An active learning approach. *CoRR*, abs/2101.06931, 2021.

Yawar Siddiqui, Julien Valentin, and Matthias Nießner. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9430–9440, 2020.

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proc. International Conference on Computer Vision (ICCV)*, pp. 5971–5980, 2019.

Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 567–576, 2015.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2443–2451, 2020.

Fuhui Tang, Chenhan Jiang, Dafeng Wei, Hang Xu, Andi Zhang, Wei Zhang, Hongtao Lu, and Chunjing Xu. Towards dynamic and scalable active learning with neural architecture adaption for object detection. In *Proc. British Machine Vision Conference (BMVC)*, 2021a.

Ying-Peng Tang, Xiu-Shen Wei, Borui Zhao, and Sheng-Jun Huang. Qbox: Partial transfer learning with active querying for object detection. *Journal of IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2021b. doi: 10.1109/TNNLS.2021.3111621.

Toan Tran, Thanh-Toan Do, Ian D. Reid, and Gustavo Carneiro. Bayesian generative active deep learning. In *Proc. International Conference on Machine Learning (ICML)*, volume 97, pp. 6295–6304, 2019.

Huy V Vo, Oriane Siméoni, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, and Jean Ponce. Active learning strategies for weakly-supervised object detection. In *Proc. European Conference on Computer Vision (ECCV)*, pp. 211–230, 2022.

Dan Wang and Yi Shang. A new active labeling method for deep learning. In *Proc. International Joint Conference on Neural Networks (IJCNN)*, pp. 112–119, 2014.

Jun Wang, Shiyi Lan, Mingfei Gao, and Larry S. Davis. Infofocus: 3d object detection for autonomous driving with dynamic information modeling. In *Proc. European Conference on Computer Vision (ECCV)*, volume 12355, pp. 405–420, 2020.

Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *Journal of IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2017.

Li Wang, Ruifeng Li, Jingwen Sun, Xingxing Liu, Lijun Zhao, Hock Soon Seah, Chee Kwang Quah, and Budianto Tandianus. Multi-view fusion-based 3d object detection for robot indoor scene perception. *Sensors*, 19(19):4092, 2019.

Xiao Wang, Xiang Xiang, Baochang Zhang, Xuhui Liu, Jianying Zheng, and QingLei Hu. Weakly supervised object detection based on active learning. *Journal of Neural Processing Letters*, pp. 1–15, 2022.

Jiaxi Wu, Jiaxin Chen, and Di Huang. Entropy-based active learning for object detection with progressive diversity constraint. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9387–9396. IEEE, 2022.

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 *Proc. International Conference on Computer Vision (ICCV)*, pp. 15490–15499, 2021.

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.

Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. STD: sparse-to-dense 3d object detector for point cloud. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1951–1960. IEEE, 2019.

Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11037–11045. Computer Vision Foundation / IEEE, 2020.

Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 93–102, 2019.

Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5330–5339, 2021.

Beichen Zhang, Liang Li, Shijie Yang, Shuhui Wang, Zheng-Jun Zha, and Qingming Huang. State-relabeling adversarial active learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8753–8762, 2020.

Mike Zhang and Barbara Plank. Cartography active learning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proc. Findings of the Association for Computational Linguistics (EMNLP)*, pp. 395–406. Association for Computational Linguistics, 2021.

Yifan Zhang, Qingyong Hu, Guoquan Xu, Yanxin Ma, Jianwei Wan, and Yulan Guo. Not all points are equal: Learning highly efficient point-based detectors for 3d lidar point clouds. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18953–18962, 2022.