

HeAL3D: Heuristical-enhanced Active Learning for 3D Object Detection

Esteban Rivera* Surya Prabhakaran Markus Lienkamp

Technical University of Munich and Munich Institute of Robotics and Machine Intelligence
 Munich, Germany

name.lastname@tum.de

Abstract

Active Learning has proved to be a relevant approach to perform sample selection for training models for Autonomous Driving. Particularly, previous works on active learning for 3D object detection have shown that selection of samples in uncontrolled scenarios is challenging. Furthermore, current approaches focus exclusively on the theoretical aspects of the sample selection problem but neglect the practical insights that can be obtained from the extensive literature and application of 3D detection models. In this paper, we introduce HeAL (Heuristical-enhanced Active Learning for 3D Object Detection) which integrates those heuristical features together with Localization and Classification to deliver the most contributing samples to the model's training. In contrast to previous works, our approach integrates heuristical features such as object distance and point-quantity to estimate the uncertainty, which enhance the usefulness of selected samples to train detection models. Our quantitative evaluation on KITTI shows that HeAL presents competitive mAP with respect to the State-of-the-Art, and achieves the same mAP as the full-supervised baseline with only 24% of the samples.

1. Introduction

3D object detection performs well in controlled scenarios and specific data distributions [28, 31]. However, once the data distribution changes, it is necessary to record, collect and label new data, followed by retraining of the object detector. Additionally, issues like class novelty and class imbalance significantly restrict the overall performance of 3D object detection in real-world settings [3, 11, 14]. A straightforward approach to solve those issues involves a supervised training procedure with extensive labelled data across all scenarios. However, labelling data for every possible distribution is impractical, if not impossible. Therefore, a scalable method is essential for robust generalization of 3d object detectors across diverse environments.

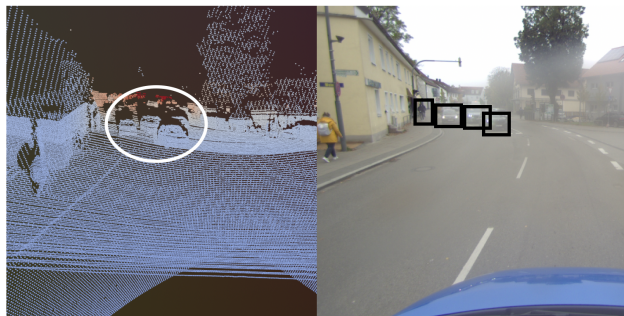


Figure 1. Typical driving scenario for an autonomous vehicle. Even for LiDAR detectors, which can measure depth directly, farther away objects are more difficult to localize, estimate and classify because of the distance and the lower number of points they reflect. Such information can be leveraged by Active Learning approaches to find the better samples to be labeled. Even human labelers require the support of camera image to label such examples, which reinforces the idea of difficulty and uncertainty.

Scalability can be pursued through several approaches. Among these, self-supervised and semi-supervised learning offer strategies that reduce the reliance on labelled data. However, initial human input remains necessary to provide the model with a foundational knowledge of objects, scenes, and defining characteristics. Despite this, minimizing human intervention is a key goal, making Active Learning (AL) an appealing strategy. AL focuses on identifying samples that are particularly challenging to classify or detect, enabling selective human analysis and labelling rather than annotating the entire dataset. In this way, human expertise is only required for specific, high-impact cases, delivering substantial improvements to the models in training while keeping the labelling effort manageable.

These strategies are broadly applicable across various deep learning domains; however, in the specific context of autonomous driving, they play a crucial role in 3D object detection. This application poses unique challenges, as it demands reliable detection in outdoor environments with constantly changing weather, lighting conditions, and a multitude of diverse scenarios. Additionally, while cameras are

* Corresponding author

often the standard for perception tasks, LiDAR has become the preferred sensor for autonomous driving due to its advantageous 3D information capture capabilities. Although algorithms originally developed for camera-based detection, including active learning, have been adapted to LiDAR data, new challenges arise with the distinct data structure associated with this sensor. These challenges necessitate specialized approaches tailored to the unique characteristics of LiDAR-based detection [25].

Furthermore, uncertainty estimation is a widely utilized approach by AL to identify the most optimal samples for labeling, as indicated by previous studies [8, 9, 12, 18]. It is therefore evident that objects for which the model is less confident should provide the most value to improve its generalization capability and consequently its performance. Moreover, the results of several 3D detection architectures [11, 33] suggests that objects situated at a distance from the sensor or enclosing a smaller number of points are consistently more challenging to detect. Such practical knowledge is given in the literature already and can be heuristically implemented into the uncertainty estimation to enhance and complement it. Therefore, in this paper, we present a new active learning strategy for 3D LiDAR object detection based on both localization and classification uncertainty of augmented samples. We introduce a Gaussian mixture model (GMM) approach to calculate an inconsistency score while incorporating detection distance and class features to improve the information quality of the selected samples.

Our contributions are summarized as below:

- We propose HeAL, an explainable uncertainty estimation method derived from the real-world application of 3D detection. It uses heuristic information like object distance to enhance an augmentation inconsistency AL approach.
- Our sample selection strategy represents the output of the 3D detection as a Gaussian Mixture Model and therefore is agnostic to arbitrary model architectures, facilitating its implementation.
- We conducted several experiments with the KITTI dataset which show an improvement of 3% mAP with respect to the current AL state-of-the-art methods (To the best of our knowledge) and the same accuracy as a fully supervised model training with 24% of the data.
- We show that previous state-of-the-art approaches perform satisfactorily in a restricted data-quantity domain and propose a piece-wise approach for the AL problem, where different methods are used for different data-domains to keep a constant improvement above random sampling

2. Related Work

2.1. Active Learning

Active learning is a strategy to minimize labeling costs by selecting the most informative samples for annotation, par-

ticularly useful in data-intensive tasks like object detection [30]. In recent years, several AL approaches have been developed, targeting the optimization of acquisition functions, uncertainty measures, and efficiency improvements in the labeling process. Classical AL methods rely on uncertainty-based sampling [1, 5, 8, 12, 17], where samples with the highest uncertainty scores are prioritized for labeling. Techniques such as BALD (Bayesian Active Learning by Disagreement) [4] and entropy-based sampling [36] have been used to capture the most ambiguous samples for model improvement. Ensemble models, like those proposed by Beluch et al. [2], increase robustness by training multiple networks and aggregating predictions.

Bayesian approaches utilize Monte Carlo dropout [13, 19] or variational autoencoders (VAEs) [34] to estimate uncertainty, which is then used as an acquisition function to select the most valuable samples. For low-dimensional problems, the uncertainty calculation is tractable, but for high-dimensional domains like image classification the uncertainty can only be estimated, giving as a result possible sub-optimal solutions for the samples to be labeled.

Core-set approaches [26, 29, 37] focus on sample diversity, selecting a representative subset of data points that minimizes model error on the entire dataset. Diversity-based methods aim to prevent overfitting to specific data distributions.

2.2. Active Learning for 2D Object detection

In image classification, uncertainty or sample-specific information is derived from the score vector. In object detection, however, additional localization and dimension data are available, helping to identify samples that are valuable for labeling. Kao et al. [22] calculate localization uncertainty by measuring the difference between the region proposal and the final object detection, which reflects the tightness of localization. Choi et al. [7] introduce a mixture density network to generate a probabilistic distribution for both the classification and localization heads, allowing for the estimation of both aleatoric and epistemic uncertainties in the data.

2.3. Active Learning for 3D Object detection

The jump from 2D to 3D object detection implies in theory just depth as an extra spatial dimension and new rotation parameters. However, the weakness of monocular cameras to measure and estimate depth increases the localization uncertainty, decreasing the usefulness of the samples to be selected. To deal with this, Hekimoglu et al. [18] propose a heatmap-based uncertainty estimation, predicting the position of the objects as a sum of gaussians, where the variance of the gaussians work as an uncertainty estimation. The problem with this implementation is that they use a specific architecture which outputs position as a probability distribu-

tion and therefore can not be implemented in general with any network.

In recent years, there has been a notable surge in interest in LiDAR architectures, with the aim of enhancing the localization capabilities in 3D object detection. This shift has simultaneously created a need for novel active learning approaches that can optimize data sampling and annotation efforts in this new domain. One prevalent approach is to apply active learning methodologies from classical machine learning or image classification to the domain of 3D detection. In their study, Greer et al. [16] employed the entropy uncertainty measurement to identify the most optimal point-cloud samples. However, given the distinct structure of point clouds compared to images or other forms of inputs, it is necessary to develop methods that are specifically tailored to this domain. Luo et al. [25] prioritize label diversity and the representativeness of sampled data, particularly by leveraging point cloud density inside detected bounding boxes. The objective of this strategy is to enhance the generalization capabilities of the selected training subsets. In a subsequent study, Luo et al. [24], revisited the fundamentals of information theory, selecting point clouds that maximize the minimal coding length required by a neural network to represent their latent features. Compared to them, we leverage the class, distance and point-density information directly on the scene representation. Additionally, our implementation does not require an extra proxy-network to calculate the coding rate or several forward passes to estimate the feature vector as a representativeness measure. Lastly, to facilitate the implementation and evaluation of these active learning methods, Ghita et al. [15] and Luo et al. [25] introduce ActiveAnno3D and Active-3D-Det frameworks respectively, which are designed to enable the testing and comparison of LiDAR 3D active learning strategies.

2.4. Consistency-based active learning

There are two ways to exploit inconsistency inside the AL domain: the query by committee approach, and the augmentation one. In query by committee, several models are trained with the same data and then the prediction of one input is queried. The inputs for which the models do not agree are the ones selected for labeling, as they present the maximal information gain. On the other hand, in the augmentation strategy there is only one model which is queried, but the inputs are preprocessed with light augmentations which should not change the result of the output. In the ideal case, both the original and augmented sample should output the same result, but when not, it means that the evaluated sample has a possible greater information gain and therefore should be labeled. For example, Yu et al. [38] demonstrate how classification-based active learning approaches underperform when used for 2D detection tasks. They propose to augment the images and calculate a consistency metric to estimate the uncertainty

of the detector. This consistency is used as a measurement of how interesting the sample is. Jeong et al. [21] applies consistency not for finding the best samples to label, but to improve the quality of pseudolabels in a semi-supervised learning approach. For multimodal setups, with simultaneous LiDAR and Camera data, the inconsistency between the two modalities can also give hints about the most informative samples [27, 39]. More recent advancements, as presented by Elezi et al. [9], emphasize a framework that combines uncertainty with robustness scores for a more comprehensive acquisition function. This method considers both the uncertainty and consistency of predictions across augmentations, addressing class imbalances and distribution drift by focusing on hard-to-detect classes. This unified approach not only enhances class-agnostic performance but also incorporates auto-labeling to balance labeling costs. Nevertheless, similar to several AL approaches for classification or 2D detection, they focus on the class vector of the output to calculate the inconsistency, whereas for 3D Object detection the localization is the greatest challenge looking forward to better performance.

3. Method

3.1. Problem formulation

Given a point cloud $P = (x, y, z)$, where (x, y, z) represents the 3D coordinates of each point, the objective of 3D object detection is to identify and localize the objects of interest as a set of 3D bounding boxes $B = \{b_k\}_{k \in [0, N_B]}$, where N_B indicates the total number of detected bounding boxes. Each bounding box comprises the center position (p_x, p_y, p_z) , the dimensions (l, w, h) , and the orientation angle θ with respect to the ego vehicle. The associated labels are denoted as $Y = \{y_k\}_{k \in [0, N_B]} \in \{1, \dots, C\}$, where C is the number of classes.

Our active learning pipeline begins with a small labeled dataset $D_L = \{(P, B, Y)_i\}_{i \in [m]}$ and a larger pool of unlabeled data $D_U = \{P_j\}_{j \in [n]}$, where $m \ll n$. Additionally, we augment the original point cloud P by rotating it 180 degrees along the z-axis to generate an augmented point cloud P' , resulting in the corresponding augmented bounding boxes B' and labels Y' [18, 27]. We choose this specific augmentation, because in the augmented version of the point cloud the represented scene is also plausible in a real scenario, compared to scaling or filtering approaches.

During the training phase, the 3D detection model is initially trained with D_L . For each active learning round $r \in [R]$, we aim to select a subset of point clouds $\{P_j\}_{j \in [0, N_r]}$ from D_U based on our active learning policy. We then query the labels of the 3D bounding boxes from an oracle $\Omega : P \rightarrow B \times Y$ and combine these with the labeled dataset to form $D_S = \{(P, B, Y)_j\}_{j \in [0, N_r]} \cup D_L$. The model is subsequently retrained with this updated dataset until the

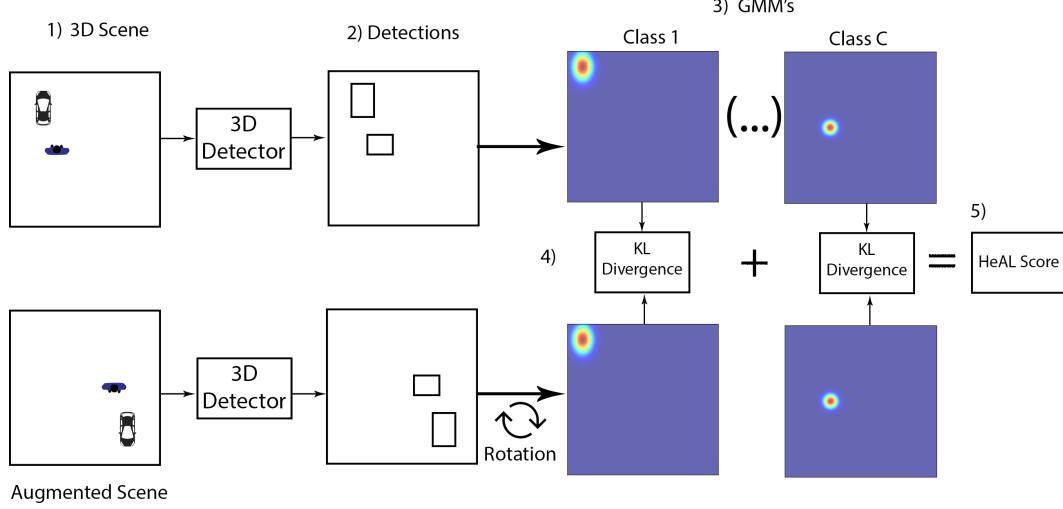


Figure 2. HeAL Score calculation pipeline, presented as BEV for clarity whereas the actual procedure is done on 3D. 1) Our input is a 3D scene represented by a point cloud, which is augmented through a 180 degree rotation around the Z axis perpendicular to the ground. 2) The 3D model from the previous AL iteration is used to detect the bounding boxes on the scene, for both the original and the augmented scene. The augmented detections are rotated again so they can be compared directly with the original detections 3) A GMM is calculated for each one of the classes in the detection, where the variance is determined by the dimensions of the object and the distance to the sensor or the point quantity of the box. 4) The inconsistency between original and augmented GMM is calculated with the KL Divergence. 5) The final HeAL score is the mean KL Divergence across all the classes. For the next AL cycle, the 100 samples with the greatest HeAL score are selected for labeling

selected samples reach the final budget B . We evaluate with respect to the number of samples and not to the number of boxes[24]. In Fig. 2 the pipeline to calculate the score is shown.

3.2. Localization Probability Map Calculation

While certain approaches calculate uncertainty by directly applying measures like Smooth-L1 Loss between predictions of original and augmented point clouds [20], we find it more effective to represent detected objects within a gaussian mixture model (GMM)[18, 28]. To create a robust probability map for both the original and augmented point clouds, we apply a 3D Gaussian probability density function (PDF) to the bounding box predictions generated by the detector. For each detected bounding box $b_k \in B$ in the original point cloud P and $b'_k \in B'$ in the augmented point cloud P' , we define a probability density function that models the likelihood of object's presence within the predicted boundaries. With this approach, we do not need the network to output directly the heatmap, but we created it from the usual output of object detection models, making our approach compatible with most of 3D detection architectures.

Each bounding box b_k is represented by its center coordinates $\mu = (p_x, p_y, p_z)$, dimensions (l, w, h) , and orientation angle θ . We define a 3D Gaussian distribution centered at (p_x, p_y, p_z) , with standard deviations proportional to the dimensions (l, w, h) , to capture the spatial uncertainty around the detected object center.

The probability at the location $\mathbf{x} = \{x, y, z\}$ for a given bounding box b_k is expressed as:

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{3/2} |\Sigma_k|^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{x} - \mu)^\top \Sigma_k^{-1} (\mathbf{x} - \mu) \right) \quad (1)$$

$$\Sigma_k = \begin{pmatrix} \sigma_x^2 & 0 & 0 \\ 0 & \sigma_y^2 & 0 \\ 0 & 0 & \sigma_z^2 \end{pmatrix}$$

where the covariance matrix Σ is diagonal, and $\sigma_x^2 = l$, $\sigma_y^2 = w$, $\sigma_z^2 = h$ are the standard deviations derived from the bounding box dimensions.

For each bounding box prediction b_k in P and its augmented prediction b'_k in P' , we compute the 3D Gaussian PDF over a spatial grid around the bounding box center. This yields probability densities representing the likelihood distribution of an object's presence within each predicted box in both original and augmented frames.

Then, the probability map for the sample is defined as the equally-weighted-GMM resulting of the sum of all gaussians:

$$f(\mathbf{x}) = \sum_k f_k(\mathbf{x}) \quad (2)$$

3.3. KL Divergence for Uncertainty Estimation

To estimate uncertainty within our active learning framework, we utilize the Kullback-Leibler (KL) Divergence, a

measure of how one probability distribution diverges from a reference distribution. In our work, KL Divergence calculates the difference between the probability maps generated from the detections on original and augmented point clouds, capturing the spatial uncertainty around detected objects.

Given the probability map $f(\mathbf{x})$ for a LiDAR point cloud in the original point cloud P and the corresponding probability map $g(\mathbf{x})$ in the augmented point cloud P' , we first rotate $g(\mathbf{x})$ by 180 degrees along the z-axis to align it with the original orientation in P . This alignment ensures that the probability distributions $f(\mathbf{x})$ and $g(\mathbf{x})$ are spatially comparable.

Once aligned, we compute the KL Divergence $D_{KL}(f||g)$ to measure the divergence between the two probability distributions:

$$D_{KL}(f||g) = \sum_{\mathbf{x}} f(\mathbf{x}) \log \frac{f(\mathbf{x})}{g(\mathbf{x})} \quad (3)$$

In this equation, $f(x)$ represents the probability map for the predictions in the original point cloud, while $g(x)$ reflects the same in the augmented (and now aligned) point cloud. A high KL Divergence indicates a significant difference in object positioning or shape between the original and augmented frames, suggesting greater uncertainty in the model's predictions for that particular bounding box.

The KL Divergence calculated between the original and augmented probability maps serves as an uncertainty metric that informs our active learning policy to prioritize samples with higher uncertainty for labeling in subsequent training rounds. By focusing on samples with higher KL Divergence, we aim to improve the model's robustness and enhance overall detection accuracy through iterative learning.

3.4. Classwise Probability Maps for Enhanced Uncertainty Estimation

The creation of probability maps, as described in Section 3.2, focuses on the 3D localization of the object. However, 3D object detection also requires accurately classifying the detected objects into predefined categories. To take the classification component into account for our uncertainty estimation, we introduce a novel method to incorporate classwise probability maps that capture the likelihood of each object class within the detected bounding boxes.

For each class label predicted by the 3D detection model, we generate a separate GMM for the corresponding class in both the original point cloud P and the augmented point cloud P' representing all the objects in the scene which correspond to that class. This classwise approach allows us to account for both the spatial distribution and the predicted category, enhancing the accuracy of uncertainty estimation across different object classes.

To calculate uncertainty, we compute the KL Divergence between the probability maps of corresponding classes from

the original and augmented point clouds. Specifically, let $f_c(x)$ represent the probability map for class c in the original point cloud and $g_c(x)$ the probability map for the same class in the augmented point cloud. After aligning $g_c(x)$ by rotating the bounding box predictions in P' by 180 degrees along the z-axis to match P , we calculate the classwise KL Divergence $D_{KL}(f_c||g_c)$ as follows:

$$D_{KL}(f_c||g_c) = \sum_{\mathbf{x}} f_c(\mathbf{x}) \log \frac{f_c(\mathbf{x})}{g_c(\mathbf{x})} \quad (4)$$

In this equation, $f_c(x)$ represents the probability map for the object of class c within the original point cloud. In contrast, $g_c(x)$ represents the same in the augmented point cloud after alignment. This process is repeated across all classes in each active learning round, yielding class-specific uncertainty scores.

To compute the overall KL Divergence D_{KL} for uncertainty estimation, we calculate the classwise KL Divergence values $D_{KL}(f_c||g_c)$ for each class c , and then take the average over all classes C .

$$D_{KL} = \frac{1}{C} \sum_{c=1}^C D_{KL}(f_c||g_c) \quad (5)$$

With this approach, the objects with wrong predicted classes will have a greater contribution to the general uncertainty estimation. This provides a comprehensive uncertainty score that integrates both spatial and class information across all detected classes.

By implementing this approach of calculating classwise KL Divergence, our active learning strategy prioritizes data samples with higher uncertainty on the basis of spatial and class dimensions. This approach aims to improve the detection model's performance on challenging classes and enhance overall accuracy in class-specific detection and localization through iterative active learning.

3.5. Distance and Point-quantity Uncertainty Correction

In LiDAR point clouds, distant objects are often represented by fewer points, which can affect the detector's accuracy. To address this issue, we propose two additional corrections to the initial GMMs, namely one for the distance of the bounding boxes and one for the number of points enclosed to the bounding box. Both farther-away and low-density boxes are more difficult to recognise, localise and estimate, and therefore, their uncertainty should be higher. On the one side, we propose the following correction parameter for the distance uncertainty for each box b_k :

$$U_{kd} = \frac{\sqrt{(x_k)^2 + (y_k)^2}}{\max_k(\sqrt{(x_k)^2 + (y_k)^2})} \quad (6)$$

Where x_k and y_k are the cartesian coordinates of b_k .

On the other side, we utilize the number of points within each bounding box to calculate point uncertainty. The uncertainty is determined using the following logarithmic formula:

$$U_{kp} = \log \left(\frac{S + \epsilon}{N_k + \epsilon} \right) + 1 \quad (7)$$

where U_k represents the uncertainty for bounding box k , S is the total number of points across all bounding boxes, N_k is the number of points within bounding box k , and ϵ is a small constant (e.g., $\epsilon = 1 \times 10^{-9}$) to prevent division by zero.

This formulation ensures that bounding boxes with fewer points or larger distances are assigned greater uncertainty, effectively emphasizing distant objects that may be challenging for the detector to identify accurately. The calculated point uncertainties are integrated into the creation of Gaussian probability maps, enhancing the spatial representation of detected objects.

$$\begin{aligned} \Sigma_{kd} &= U_d \cdot \Sigma_k \\ \Sigma_{kp} &= U_p \cdot \Sigma_k \end{aligned} \quad (8)$$

where Σ_{kp} and Σ_{kd} are the adjusted standard deviation for bounding box k , Σ_k is the original standard deviation used in Equation 1, and U_{kd} and U_{kp} are both distance and point uncertainties.

This strategic approach ensures that even challenging classes, particularly those represented by distant objects, are adequately considered during the training process, ultimately leading to improved detection accuracy in complex environments, whose individual components are explainable.

3.6. Experiment

3.6.1. Datasets

The KITTI dataset [14] is a widely recognized benchmark for point cloud-based 3D object detection. The dataset consists of 3,712 training samples and 3,769 validation samples. It contains labels of cars, pedestrians, and cyclists. The nuScenes dataset [3] offers a comprehensive suite of sensory data, with precise annotations for 40,000 frames. To ensure consistency with the KITTI dataset representation, we selected a subset of 7,481 samples divided into 3,712 samples for training and 3,769 samples for validation.

3.6.2. Evaluation Metrics

For a fair comparison of the baselines and our proposed method, we utilize the evaluation protocol established for the KITTI dataset, following the approach of Shi et al. [31]. Specifically, we employ Average Precision (AP) for both 3D object detection, and the evaluation difficulty is categorized as EASY, MODERATE, and HARD. The related Intersection over Union (IoU) thresholds are set at 0.7 for the car class and 0.5 for both pedestrian and cyclist classes. The results are computed using 40 recall positions on the validation split.

3.6.3. Implementation Details

To ensure the reproducibility of the baselines and the proposed approach, we implemented our method using the publicly available ACTIVE-3D-DET toolbox [25]. We utilize the PV-RCNN [31] backbone as the standard 3D object detection model for a fair comparison across baselines and our proposed approach. All models were trained using an NVIDIA A40 GPU. The batch sizes are set to 1 for training and 4 for evaluation on the KITTI [14] dataset. We adopt the Adam [23] optimizer with an initial learning rate of 0.01, following a one-cycle learning rate scheduler for efficient convergence. The code for the strategy will be available upon acceptance.

For each active learning cycle, the number of unlabeled samples to be queried was defined as $N_r = 100$, training the models 30 epochs for each cycle. To evaluate 3D object detection, the 3D mAP is presented. In each plot, the mean results for three different seeds were presented with a shaded region representing the standard deviation.

3.7. 3D Object Detector

For our object detection pipeline, we employ the PV-RCNN model, a state-of-the-art architecture in point cloud processing for autonomous driving applications. PV-RCNN [31] combines voxel-based and point-based feature extraction methods to achieve high accuracy in 3D object detection. The model’s architecture utilizes both 3D voxel CNNs and PointNet [6] to effectively capture and encode fine-grained spatial features, which enables robust and accurate predictions of bounding boxes around objects within point cloud data.

Our pipeline begins by feeding raw point cloud data into PV-RCNN to obtain initial predictions. These predictions consist of bounding boxes, object classes, and associated confidence scores. In the following steps, these predictions serve as the basis for uncertainty estimation and sample selection in an active learning framework.

4. Results

For the comparison with the literature, we compared several methods: First, Random sampling as baseline; for domain agnostic AL, Entropy[35] and Montecarlo[10]; Finally, for LiDAR 3D-Detection methods we compared with the current State-of-the-art, namely CRB[25] and KECOR [24].

Figure 3 shows the results across eight active learning cycles. In the low-data regime, only KECOR has the same performance as the Random baseline. However, as more samples are incorporated, HeAL demonstrates increasingly robust results, outperforming both KECOR, CRB and the Random baseline by 3%, 4.9% and 4.2% after the eighth active learning cycle, respectively. Notably, in the final iteration, even the simple entropy-based selection strategy surpasses KECOR in performance. Discriminating according to

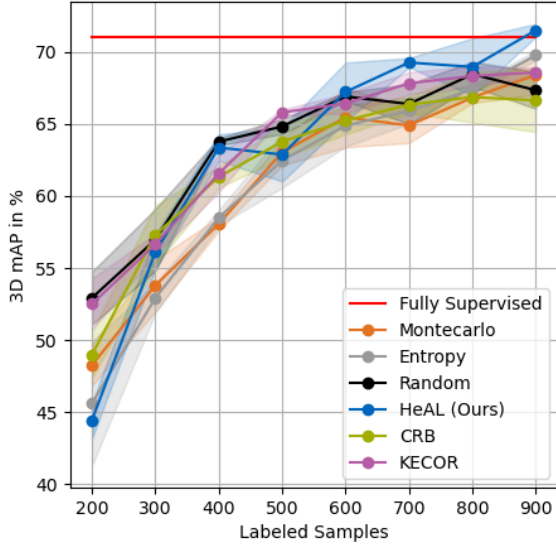


Figure 3. mAP in % with respect to the number of labeled samples for selected AL strategies on KITTI.

difficulty level confirms the trend, where for the 3 categories a similar trend as in the combined mAP is observed.

We provide an evaluation of both Heal and the state-of-the-art KECOR method using the NuScenes dataset in Fig. 4, where our approach consistently outperforms KECOR by 2% across three different seeds. Furthermore, its performance across iterations demonstrates a clear upward trend, in contrast to KECOR, which exhibits more unpredictable behavior. This pattern aligns with the evaluation of KECOR and CRB on the Waymo dataset [32], where their trends are similarly harder to discern. This variability can be attributed to the overall complexity of datasets like NuScenes and Waymo, which pose significant challenges for detectors.

In Fig. 5, HeAL outperforms the concurrent method for all difficulty levels. Similarly, separating the results by class (Fig. 6) shows an interesting behavior for HeAL: For the class Car, which is the most representative class in KITTI, there is almost no improvement with respect to the random baseline; however, for both pedestrian and cyclist there is a clear improvement on the AP. The sample selection method improve the performance of unrepresented classes, mostly because for them is easier to find extra information to improve the detection, rather than from the Car class, where there are already too many samples to start with.

More generally, in the higher data regime, both CRB and KECOR demonstrate minimal to no improvement, whereas HeAL and even entropy (in the final iteration) exhibits superior performance. It is evident that KECOR exhibit its greatest strength during the initial iterations, with a reduced number of boxes, and present the most significant positive differ-

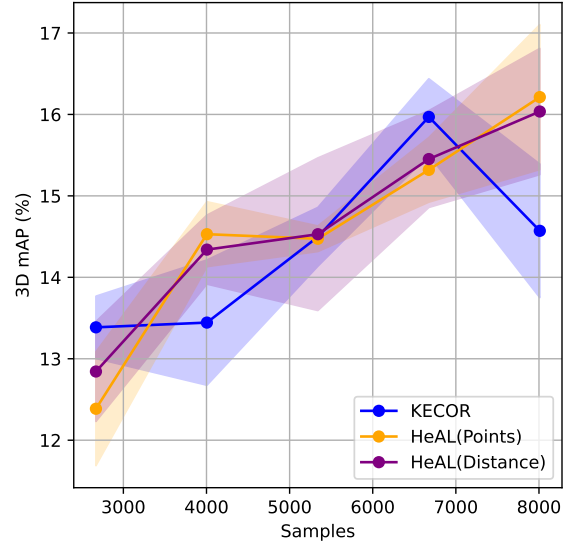


Figure 4. mAP in % with respect to the number of labeled samples for selected AL strategies on Nuscenes.

ence in performance. However, the trend indicates that for the higher-data regime, their performance reaches a plateau, indicating that no further improvement can be extracted from the information-theoretic insights they propose. We argue that, given the current state of the art in this field, there is no optimal AL strategy for the entire data range. Rather, it may be more beneficial to employ a combination of strategies for different data domains, particularly when working with a fixed budget. Specifically, KECOR can be employed as a preliminary approach, followed by a transition to an alternative strategy if the budget permits and their performance does not improve. This alternative could be HeAL which can enhance the model’s performance through the whole range. This complements the goal of AL, which is to obtain the highest possible metrics with the smallest number of samples. Even if the labeling budget is not extremely small, the goal is still to find the best samples to label in the whole data space: HeAL is then a good complementary strategy.

4.1. Ablation studies

Table 1. Ablation mAP results for 700 samples on KITTI

GMM	Class	Distance	Point	Easy	Moderate	Hard	Total
✓				75.9	63.3	59.5	66.3
✓	✓			77.2	63.8	59.9	67.0
✓		✓		74.5	61.1	57.1	64.3
✓			✓	75.4	61.87	57.9	65.0
✓	✓		✓	76.5	62.8	58.9	66.1
✓	✓	✓		79.3	66.2	62.3	69.3

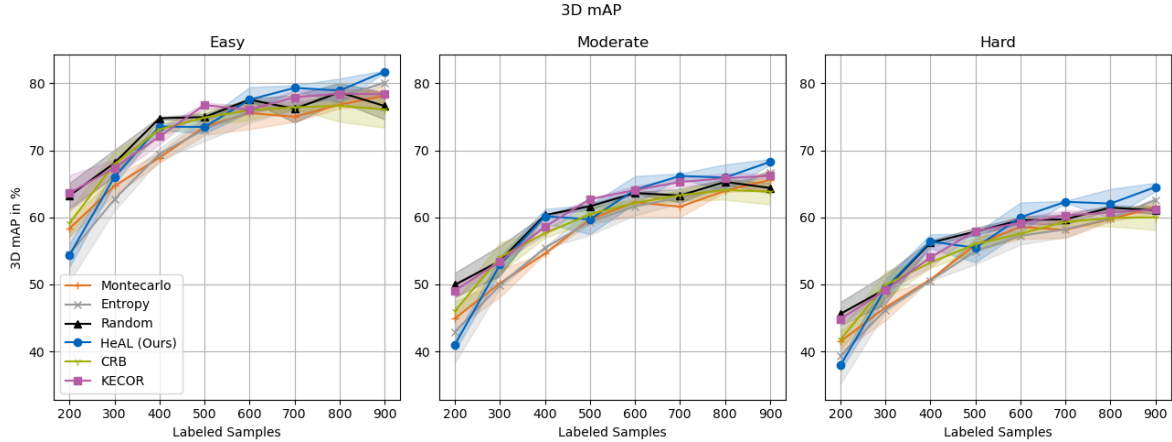


Figure 5. mAP in KITTI for each difficulty level

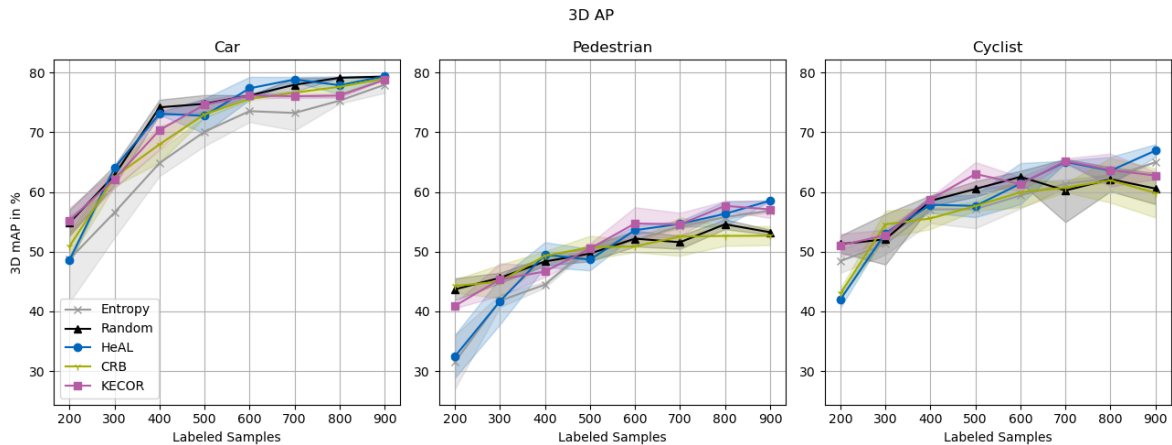


Figure 6. AP in KITTI for each class

Finally, we investigate separately the different modules which compose our method, namely localization uncertainty, classification correction and difficulty correction. In Tab. 1 we show which separated modules contribute the most to the final performance. It can be seen, that the point uncertainty presents the lowest performance coupled directly with the localization uncertainty, being even lower than the vanilla localization uncertainty. In contrast, the class information improves the pure heatmap in 1.3%. The best performance is shown by the Localization+Class+Distance uncertainty approach, which achieves an improvement of 2.3% with respect to the second best, the Localization+Class approach. The separation of Class into multiple GMM appears to be the defining factor to improve the performance. This could be because this separation penalizes more strictly a missclassification of the object than differences in the classification score vector of a regular object detector.

Lastly, even though both the distance and number of points are correlated, it is clear that the number of points damages the overall efficiency of the sampling strategy. A possible explanation, is that occlusions cause that some ob-

jects, even when they enclose few points, are still close enough to the sensor, and are therefore easier to localize, classify and detect than farther away boxes, which would make them less suitable to be labeled, given their low amount of valuable information for the training.

5. Conclusion

In this paper we present HeAL, a novel and explainable AL sampling method for 3D Object detection. Our method is derived from heuristic observations regarding the difficulty of detecting objects located at a distance. This information is then used to adjust the uncertainty of a GMM representing the detection heatmap from a scene. Subsequently, this heatmap is compared with an augmented version of it, with the aim of selecting the samples exhibiting the greatest inconsistencies. Our approach achieves results competitive to the state-of-the-art for the data domain where previous works have reached a plateau. We therefore propose a multi-approach strategy, in which several AL strategies are employed sequentially to achieve the largest mAP improvement across the entire data range.

References

- [1] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *ICLR*. ACM, 2020. 2
- [2] William H. Beluch, Tim Genewein, Andreas Nurnberger, and Jan M. Kohler. The power of ensembles for active learning in image classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9368–9377, 2018. 2
- [3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1, 6
- [4] Xiaofeng Cao and Ivor Wai-Hung Tsang. Bayesian active learning by disagreements: A geometric perspective. *ArXiv*, abs/2105.02543, 2021. 2
- [5] Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Sequential graph convolutional network for active learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9578–9587, 2021. 2
- [6] R. Qi Charles, Hao Su, Mo Kaichun, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, 2017. 6
- [7] Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clement Farabet, and Jose M. Alvarez. Active learning for deep object detection via probabilistic modeling. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10244–10253, 2021. 2
- [8] Pan Du, Suyun Zhao, Hui Chen, Shuwen Chai, Hong Chen, and Cuiping Li. Contrastive coding for active learning under class distribution mismatch. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8907–8916, 2021. 2
- [9] Ismail Elezi, Zhiding Yu, Anima Anandkumar, Laura Leal-Taixé, and Jose M. Alvarez. Not all labels are equal: Rationalizing the labeling costs for training object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14472–14481, 2022. 2, 3
- [10] Di Feng, Xiao Wei, Lars Rosenbaum, Atsuto Maki, and Klaus Dietmayer. Deep active learning for efficient training of a lidar 3d object detector. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 667–674, 2019. 6
- [11] Felix Fent, Fabian Kutenreich, Florian Ruch, Farija Rizwin, Stefan Juergens, Lorenz Lechermann, Christian Nissler, Andrea Perl, Ulrich Voll, Min Yan, and Markus Lienkamp. Man truckscenes: A multimodal dataset for autonomous trucking in diverse conditions. In *Advances in Neural Information Processing Systems*, pages 62062–62082. Curran Associates, Inc., 2024. 1, 2
- [12] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1050–1059, New York, New York, USA, 2016. PMLR. 2
- [13] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, page 1183–1192. JMLR.org, 2017. 2
- [14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 1, 6
- [15] Ahmed Ghita, Bjørk Antoniussen, Walter Zimmer, Ross Greer, Christian Creß, Andreas Møgelmoose, Mohan M. Trivedi, and Alois C. Knoll. Activeanno3d - an active learning framework for multi-modal 3d object detection. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 1699–1706, 2024. 3
- [16] Ross Greer, Bjork Antoniussen, Mathias V. Andersen, Andreas Mogelmoose, and Mohan M. Trivedi. The why, when, and how to use active learning in large-data-driven 3d object detection for safe autonomous driving: An empirical exploration. *arXiv.org*, abs/2401.16634, 2024. 3
- [17] Denis Gudovskiy, Alec Hodgkinson, Takuya Yamaguchi, and Sotaro Tsukizawa. Deep active learning for biased datasets via fisher kernel self-supervision. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9038–9046, 2020. 2
- [18] Aral Hekimoglu, Michael Schmidt, Alvaro Marcos-Ramiro, and Gerhard Rigoll. Efficient active learning strategies for monocular 3d object detection. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 295–302, 2022. 2, 3, 4
- [19] Neil Houlsby, Ferenc Huszar, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *CoRR*, abs/1112.5745, 2011. 2
- [20] Sihwan Hwang, Sanmin Kim, Youngseok Kim, and Dong-suk Kum. Joint semi-supervised and active learning via 3d consistency for 3d object detection. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4819–4825. IEEE, 2023. 4
- [21] Jisoo Jeong, Seungeui Lee, Jeessoo Kim, and Nojun Kwak. *Consistency-based semi-supervised learning for object detection*. Curran Associates Inc., Red Hook, NY, USA, 2019. 3
- [22] Chieh-Chi Kao, Teng-Yok Lee, Pradeep Sen, and Ming-Yu Liu. Localization-aware active learning for object detection. In *Computer Vision – ACCV 2018*, pages 506–522, Cham, 2019. Springer International Publishing. 2
- [23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6
- [24] Yadan Luo, Zhuoxiao Chen, Zhen Fang, Zheng Zhang, Mahsa Baktashmotlagh, and Zi Huang. Kecor: Kernel Coding Rate Maximization for Active 3D Object Detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18233–18244, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 3, 4, 6

- [25] Yadan Luo, Zhuoxiao Chen, Zijian Wang, Xin Yu, Zi Huang, and Mahsa Baktashmotlagh. Exploring active 3d object detection from a generalization perspective. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. [2](#), [3](#), [6](#)
- [26] Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Reza Haf-fari, Anton Van Den Hengel, and Javen Qinfeng Shi. Active Learning by Feature Mixing . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12227–12236, Los Alamitos, CA, USA, 2022. IEEE Computer Society. [2](#)
- [27] Esteban Rivera, Ana Clara Serra Do Nascimento, and Markus Lienkamp. Camera-lidar inconsistency analysis for active learning in object detection. In *35th IEEE Intelligent Vehicles Symposium, IV 2024*, pages 97–103. Institute of Electrical and Electronics Engineers Inc., 2024. Publisher Copyright: © 2024 IEEE.; 35th IEEE Intelligent Vehicles Symposium, IV 2024 ; Conference date: 02-06-2024 Through 05-06-2024. [3](#)
- [28] Thomas Roddick, Alex Kendall, and Roberto Cipolla. Orthographic feature transform for monocular 3d object detection. *British Machine Vision Conference*, 2019. [1](#), [4](#)
- [29] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. [2](#)
- [30] H. S. Seung, M. Oppor, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, page 287–294, New York, NY, USA, 1992. Association for Computing Machinery. [2](#)
- [31] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10529–10538, 2020. [1](#), [6](#)
- [32] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. [7](#)
- [33] Pei Sun, Weiyue Wang, Yuning Chai, Gamaleldin Elsayed, Alex Bewley, Xiao Zhang, Cristian Sminchisescu, and Dragomir Anguelov. RSN: Range Sparse Net for Efficient, Accurate LiDAR 3D Object Detection . In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5721–5730, Los Alamitos, CA, USA, 2021. IEEE Computer Society. [2](#)
- [34] Toan Tran, Thanh-Toan Do, Ian D. Reid, and G. Carneiro. Bayesian generative active deep learning. In *International Conference on Machine Learning*, 2019. [2](#)
- [35] Dan Wang and Yi Shang. A new active labeling method for deep learning. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 112–119, 2014. [6](#)
- [36] Jiaxi Wu, Jiaxin Chen, and Di Huang. Entropy-based Active Learning for Object Detection with Progressive Diversity Constraint . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9387–9396, Los Alamitos, CA, USA, 2022. IEEE Computer Society. [2](#)
- [37] Donggeun Yoo and In So Kweon. Learning Loss for Active Learning . In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 93–102, Los Alamitos, CA, USA, 2019. IEEE Computer Society. [2](#)
- [38] Weiping Yu, Sijie Zhu, Taojiannan Yang, and Chen Chen. Consistency-based active learning for object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3950–3959, 2022. [3](#)
- [39] Jiakang Yuan, Bo Zhang, Xiangchao Yan, Tao Chen, Bortian Shi, Yikang Li, and Yu Qiao. Bi3D: Bi-Domain Active Learning for Cross-Domain 3D Object Detection . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15599–15608, Los Alamitos, CA, USA, 2023. IEEE Computer Society. [3](#)