Monocular 3D Object Detection with LiDAR Guided Semi Supervised Active Learning

Aral Hekimoglu
Technical University Munich
Munich, Germany
aral.hekimoglu@tum.de

Michael Schmidt
BMW Group
Munich, Germany
michael.se.schmidt@bmw.de

Alvaro Marcos-Ramiro
BMW Group
Munich, Germany
alvaro.marcos-ramiro@bmw.de

Abstract

We propose a novel semi-supervised active learning framework for monocular 3D object detection with LiDAR guidance (MonoLiG), which leverages all modalities of collected data during model development. We utilize LiDAR to guide the data selection and training of monocular 3D detectors without introducing any overhead in the inference phase. During training, we leverage the LiDAR teacher, monocular student cross-modal framework from semi-supervised learning to distill information from unlabeled data as pseudo-labels. To handle the differences in sensor characteristics, we propose a data noise-based weighting mechanism to reduce the effect of propagating noise from LiDAR modality to monocular. For selecting which samples to label to improve the model performance, we propose a sensor consistency-based selection score that is also coherent with the training objective. Extensive experimental results on KITTI and Waymo datasets verify the effectiveness of our proposed framework. In particular, our selection strategy consistently outperforms state-of-the-art active learning baselines, yielding up to 17% better saving rate in labeling costs. Our training strategy attains the top place in KITTI 3D and bird's-eye-view (BEV) monocular object detection official benchmarks by improving the BEV Average Precision (AP) by 2.02. Code is shared at https://github.com/aralhekimoglu/monolig.

1. Introduction

3D object detection is fundamental in scene understanding for autonomous driving vehicles. Detectors operating on point cloud scans from the LiDAR sensor achieve impressive performance on benchmarks like KITTI [16]; however, they are costly for consumer vehicles. Monocular RGB cameras offer a cheaper alternative. Therefore, there has been a surge of interest in research on monocular 3D object detectors. Convolutional Neural Network (CNN) based

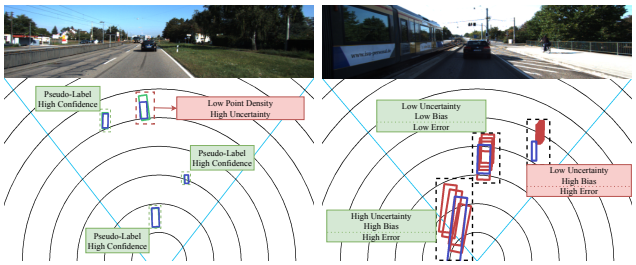


Figure 1. RGB image and predictions in BEV space of 2 frames from KITTI. (a) In regions with low point cloud return, the LiDAR detector's predictions (green) are not safe to use as pseudo-labels since they do not perfectly overlap with ground truth (blue). (b) In predictions with low uncertainty but high variation from the ground-truth (bias), uncertainty from an ensemble of monocular detectors (red) is not enough to capture erroneous samples.

monocular detectors achieve state-of-the-art (SOTA) performance with the help of massive annotated datasets. However, annotating a large amount of 3D detection data is time and labor-consuming. Specifically for monocular 3D object detectors, manually annotating 3D boxes from monocular imagery is infeasible due to a lack of depth information. Therefore, LiDAR point clouds are recorded during data collection, and annotators label 3D box locations on the collected point clouds. To save annotation costs, only the most informative frames in the collected samples are labeled to train models. Consequently, large amounts of LiDAR data with beneficial 3D information remain unlabeled.

Semi-supervised learning (SSL) and active learning (AL) are two related techniques that aim to improve model performance while minimizing labeling effort by utilizing unlabeled data. AL focuses on selecting the most informative samples for labeling, while SSL focuses on training the model using unlabeled data.

In a recent cross-modal SSL method [30], predictions from the LiDAR detector (teacher) are treated as the ground truth of unlabeled data. They are combined with annota-

tions of labeled data to train the monocular detector (student). However, as shown in Fig. 1a, some predictions from the LiDAR detector are not accurate, and thus learning from them is not optimal for the monocular detector. We observe that these inaccurate predictions are typically from areas with low point cloud densities, i.e., distant or occluded objects. These correspond to regions in LiDAR where aleatoric (data) uncertainty is high [12].

The key idea of AL for object detection is to leverage the current detector to select the most informative samples for labeling under a fixed-labeling budget. The selection is based on an acquisition function that estimates the detector's uncertainty on samples. Then, samples with high uncertainty are selected for labeling. However, as shown in Fig. 1b, some detections with low uncertainty are still far from the ground truth. For these samples, having an acquisition function that measures this discrepancy is essential.

To solve the problems mentioned above, we present the MonoLiG framework, illustrated in Fig. 2, that consists of a coherent selection and training phase. During our training phase (Sec. 3.3), we utilize the cross-modal teacher-student framework with a LiDAR teacher and a monocular student detector. To reduce the effect of incorrect LiDAR predictions, we propose to scale the loss of the monocular detector based on the confidence of generated labels. To this end, we extend the LiDAR detector with an aleatoric uncertainty estimation head and define the confidence of predicted labels with the aleatoric uncertainty of LiDAR. During our selection phase (Sec. 3.4), inspired by the cross-modal teacher-student frameworks from SSL, we extend the uncertainty-based selection score and use LiDAR predictions as pseudo-labels to measure the distance of monocular predictions to the ground truth. To our knowledge, our work is the first to leverage the teacher-student paradigm for AL selection and integrate it with a coherent SSL training strategy. By combining AL and SSL, MonoLiG is able to select challenging samples that are difficult to learn with semi-supervised training and thus achieve higher model performance with minimal labeling costs.

Our main contributions are summarized as follows:

- We propose MonoLiG, a novel framework that consists of a coherent selection and training phase. The proposed strategies outperform AL and SSL baselines separately, and achieves the best performance when utilized coherently.
- We extend current uncertainty strategies for AL selection by adapting the teacher-student paradigm and adding an inconsistency term, resulting in a better data savings rate than the SOTA AL baselines.
- We identify the potential error propagation from the teacher to the student model in cross-modal teacher-student SSL methods and propose a pseudo-label weighting mechanism based on the aleatoric uncertainty of the teacher. Our proposed training strategy define a new SOTA for monocular 3D object detection in the KITTI test benchmark.

2. Related work

2.1. Active learning for object detection

Pool-based AL selection methods can be grouped into two categories: uncertainty-based [4, 9, 14] and diversity-based [1, 36, 41]. One approach to estimate the uncertainty is through ensembles [4]. Different models are trained with different random initializations to construct a committee with slightly different predictions for uncertain samples. Then, the informativeness score is obtained by an acquisition function like entropy [37], or BALD [18] for classification tasks or total variance (TV) [46] for regression tasks. In contrast, diversity-based methods target maintaining the distribution of the unlabeled pool by selecting a set of samples that covers the remaining points within a distance. Core-set [36] uses Euclidean distance in the feature space learned by CNNs, and CDAL [1] utilizes KL-divergence between context features, which they define as a mixture of predicted softmax probabilities in a detection network. One recent task-agnostic approach, LL4AL [51], trains a loss-learning module during training and uses the predicted loss as the score to select samples.

AL is extended for 2D object detection [2, 11, 17, 52], and 3D object detection from LiDAR [9, 13, 35]. Elezi *et al*. [11] select samples using uncertainty and robustness of the detector, defined by the consistency between a sample and its augmented version. Yu *et al*. [52] propose a 2-stage selection strategy. First, they select samples using a consistency-based metric and continue selection with a score that promotes the class distribution of selected samples to be different from the labeled pool.

Most works for AL for object detection focus on classification [2, 17, 53] and ignore localization of the bounding boxes. Choi *et al*. [9] estimate the aleatoric and epistemic uncertainty for both classification and localization and combine the uncertainties in a single selection score. In [35], Schmidt *et al*. train an ensemble of models and define the localization uncertainty as the intersection over union (IoU)-based matching score. Since localization is more challenging for monocular 3D detectors, our work also utilizes the localization uncertainty-based selection score.

To our knowledge, our work is to first to exploit the bias, defined by the teacher-student inconsistency, as a selection criterion.

2.2. Semi-supervised learning in object detection

SSL aims to improve the performance of a model by training with a limited labeled dataset and exploiting infor-

mation from a large amount of unlabeled data. SSL approaches can be categorized into two groups: consistency regularization [7, 21, 22] and pseudo-labeling [6, 24, 27]. Consistency regularization trains the model's parameters on unlabeled data by penalizing the inconsistency between predictions for the same input under different perturbations. In pseudo-labeling methods, [24], a trained model predicts labels for unlabeled samples. Then the model is trained on the unlabeled data using the *pseudo-labels* as the target.

An issue with pseudo-labeling is overfitting to incorrect predictions due to the confirmation bias [3]. One solution is filtering pseudo-labels based on a confidence score (hard-thresholding). FixMatch [42] enhances the quality of pseudo-labels by filtering predictions from the teacher with low classification confidence. Wang *et al*. [47] extend this approach to 3D object detection by using an additional IoU-based localization confidence score. Another solution is using soft-pseudo-labels [33, 39] and scaling the effect of each prediction based on their confidence.

Recently, the teacher-student paradigm has been used for monocular 3D object detection in a cross-modal setting to transfer information from one modality (LiDAR) to another (monocular) [10, 30]. Chong *et al*. [10] distill information from the LiDAR detector with feature and label guidance. Similarly, Peng *et al*. [30] generate pseudo-labels using a LiDAR teacher model to train a student monocular detector.

In this cross-modal setting, our work is the first to identify the theoretical error propagation from teacher to student in the form of modality-specific aleatoric uncertainty and propose a confidence-based method to mitigate the effect of highly uncertain predictions.

2.3. Semi-supervised active learning

Recent works [15, 20, 40, 43, 48] combine SSL and AL using semi-supervised techniques like pseudo-labeling in the training phases of AL cycles to distill information from the unlabeled data. Huang *et al*. [20] constructs a Mean Teacher [45] by applying an exponential moving average (EMA) to weights obtained at the end of each AL cycle. Gao et al. [15] proposed an AL framework that utilizes a selection score based on augmentation consistency with a SSL training strategy penalizing augmentation inconsistency.

Our work is the first to jointly formulate the SSL and AL optimization problems and propose a coherent semi-supervised active learning (SSAL) framework.

3. Methodology

3.1. Optimization problem

Let (x, y) be a sample pair drawn from the dataset space D. For an input scene x - consisting of a synchronized point cloud and an image - label y contains bounding box parameters b_o and a semantic class label c_o for all objects o within the scene. We use the cross-modal teacher-student paradigm [30], where the student model f_s is a monocular detector and the teacher model f_t is a LiDAR detector. During training, we use both models, for inference, we only deploy the student model. Therefore, the optimization objective is to reduce the expected loss of the student model $f_s(x;\theta)$ given by,

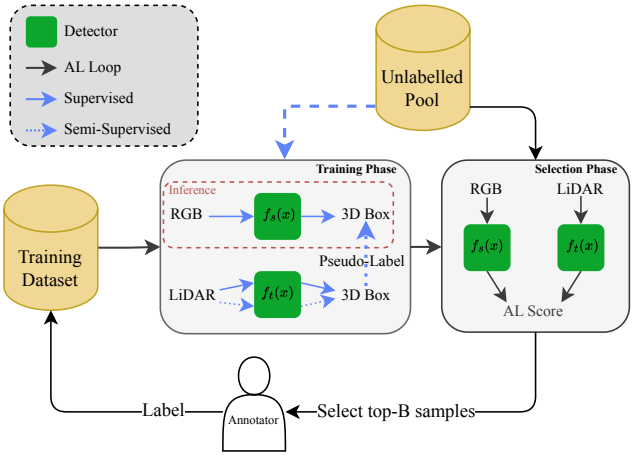


Figure 2. Overview of the proposed MonoLiG framework. At each cycle, we first train our LiDAR detector with the training dataset and the monocular detector in a semi-supervised manner with LiDAR predictions as pseudo-labels. During selection, we use predictions from both detectors to compute a selection score for each unlabeled sample. Then, a human annotator labels samples with the highest score under a labeling budget B.

$$E[L;\theta] = \iint_{x,y\sim D} L(f_s(x_s;\theta), y)dxdy \quad (1)$$

where $L(f_s(x;\theta), y)$ represents a loss function. We formulate our theory using mean-square error (MSE) as the regression loss for a single bounding box and compute the sample loss as the sum of losses of all boxes. We decompose Eq. (1) into two components (derivation in the supplement):

$$L(f_s(x;\theta), y) = (f_s(x;\theta) - h_s(x))^2 + (h_s(x) - y)^2 \quad (2)$$

where $h_s(x)$ is the global optimum for $f_s(x;\theta)$. The first component of this equation can be optimized during training, whereas the second component represents u_s^{al}, the aleatoric uncertainty of the student (data noise), that cannot be reduced by optimization.

We follow pool-based AL, where we assume access to a labeled sample set (x_T, y_T) belonging to the training dataset T, and to unlabeled samples (x_U) from a large unlabeled data pool U, randomly (i.i.d.) sampled from the dataset space D. Then, our optimization objective is ap-

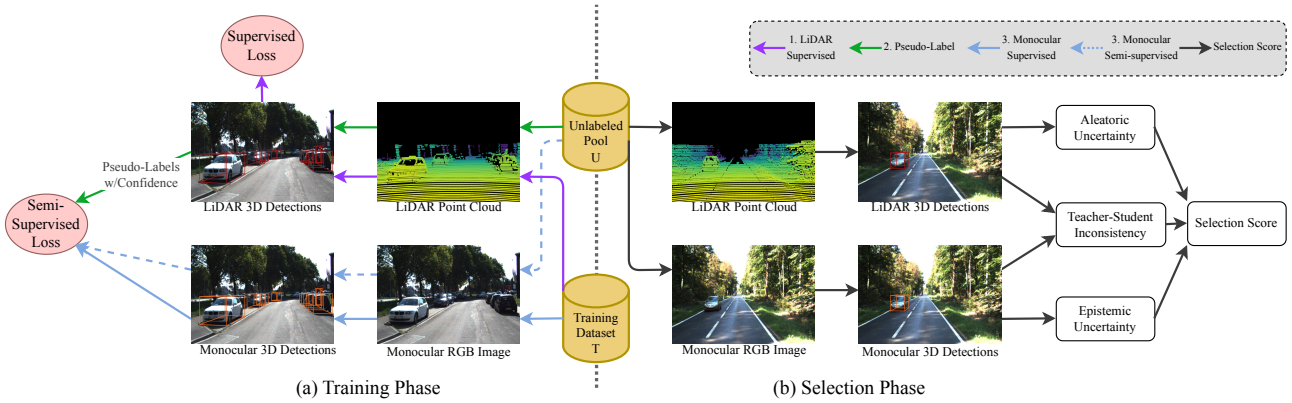


Figure 3. Illustration of the proposed MonoLiG framework. **(a) In the training phase**, 1) We train a LiDAR detector as our teacher model. 2) Using its predictions on unlabeled data, we generate pseudo-labels and assign confidence based on the aleatoric uncertainty. 3) We train a monocular detector as our student model with SSL. **(b) In the selection phase**, we select samples based on the epistemic uncertainty of the student model, the inconsistency between predictions from the teacher and the student, and the aleatoric uncertainty of the teacher.

proximated as follows:

$$E[L;\theta] \approx \sum_{(x_T,y_T)\epsilon T} (f_s(x_T;\theta) - y_T)^2 + \sum_{(x_U,y_U)\epsilon U} (f_s(x_U;\theta) - h_s(x_U))^2 + (h_s(x_U) - y_U)^2 \tag{3}$$

3.2. MonoLiG overview

The MonoLiG framework, illustrated in Fig. 2, optimizes Eq. (3) during each AL cycle through two phases: a semi-supervised training phase and an active learning selection phase. During the training phase (Fig. 3a, Sec. 3.3), we train with supervised learning using the labels y_T as our target. We employ the teacher-student paradigm on unlabeled data and use the predictions from the teacher model $f_t(x)$ as a proxy target for $f_s(x)$ in the form of pseudo-labels. We extend this paradigm with a pseudo-labeling weighting mechanism based on aleatoric uncertainty. During the selection phase (Fig. 3b, Sec. 3.4), we minimize the expected loss of the student model by selecting a subset S from U, under a fixed budget $|S| = B$, to be labeled by an oracle and moved to T for retraining the student model in the next cycle. We propose a scoring function consisting of three components: epistemic uncertainty of the student, inconsistency between the predictions of the teacher and the student, and aleatoric uncertainty of the teacher to be coherent with our training phase. The training and selection cycle repeats until a stop condition is satisfied, i.e., the detector's performance converges for several iterations or reaches the desired performance.

3.3. Training phase with semi-supervised learning

We follow the recent cross-modal pseudo-labeling approaches [10, 30] to optimize the objective in Eq. (3). For the training dataset, we optimize the student model with supervised learning using labels y_T. For unlabeled samples, predictions from the teacher model $f_t(x)$ are given as a proxy for $h_s(x)$, and the student model is optimized towards the proxy target. If we replace $h_s(x)$ with an optimal teacher model $h_t(x)$ in Eq. (3), we can rewrite the unlabeled part as:

$$\sum_{(x_U,y_U)\epsilon U} (f_s(x_U;\theta) - h_t(x_U))^2 + u_t^{al}(x_U) \tag{4}$$

Using teacher model predictions as labels, the first term can theoretically be fully reduced after optimization. However, our initial objective of minimizing the expected loss of the student model does not converge toward its global minima due to the additional second term, aleatoric uncertainty of the teacher $u_t^{al}(x)$. With this formulation, we identify the potential error propagation from the teacher to the student model in our baseline framework [30]. Therefore, in MonoLiG, we reduce the effect of pseudo-labels with high teacher aleatoric uncertainty while training the student model.

Fig. 3a presents the training phase in MonoLiG with the following steps:

1. Teacher training with aleatoric uncertainty using T
2. Pseudo-label generation on samples of U using the predictions of the teacher model
3. Student training with T using y_T and with U using pseudo-labels $f_t(x)$

Training of the teacher model with aleatoric uncertainty calculation. We design MonoLiG to utilize any

object detector as its teacher model. A typical 3D object detector outputs seven bounding box regression parameters defined by the center coordinates (x, y, z), dimensions (w, h, l), and rotation angle α. We use gaussian modeling to model the aleatoric uncertainty of a bounding box. We assume a Gaussian distribution for each regression variable and modify the teacher detector to output the mean $\mu(x;\theta)$ and the uncertainty $\sigma^2(x;\theta)$. To optimize the teacher model with the uncertainty head, we use negative log-likelihood (NLL). For a Gaussian distribution, the NLL can be written as:

$$L(x;\theta) = \frac{y - \mu(x;\theta)^2}{2\sigma^2(x;\theta)} + \frac{log\sigma^2(x;\theta)}{2} \tag{5}$$

To define a single confidence score per bounding box, we sum the uncertainty of the regression parameters:

$$u_t^{al}(x) = \sigma_x(x;\theta) + \sigma_y(x;\theta) + \sigma_z(x;\theta) \tag{6}$$

We give the statistics of u_t^{al} in the supplementary.

Pseudo-label and confidence generation. To generate pseudo-labels for all samples in the unlabeled pool, we perform inference using our teacher model to detect objects and apply post-processing, such as non-maximum suppression (NMS). To scale the effect of teacher's predictions based on their uncertainty during the training of the student model, we assign a confidence score to each prediction. We propose to use $1 - u_t^{al}(x)$ from Eq. (6) as the localization confidence and the probability of the predicted class $p(x)$ as the classification confidence and combine as follows:

$$c(x) = max\{0, p(x) * (1 - u_t^{al}(x))\} \tag{7}$$

Student training using semi-supervised learning. The MonoLiG framework is compatible with any object detector as its student model. The loss function is updated to incorporate the ability to scale the effect of each bounding box label based on its confidence. We scale the original loss of the student model with the confidence as follows:

$$L_c(x, y, c) = c(x) * L(f_s(x), y) \tag{8}$$

The student model is trained with labels y_T for the training set T and the pseudo-labels $\tilde{y}_U$ for the unlabeled set U. Joint loss is given as:

$$L = L_c(x_T, y_T, c_T) + \lambda_U L_c(x_U, \tilde{y}_U, c_U) \tag{9}$$

where λ_U is the weight of the loss for unlabeled samples. We set the confidence of labeled samples c_T to 1 and λ_U to 0.5 following our ablation study in the supplementary.

3.4. Selection phase with active learning

The goal of the selection phase is to select the best subset S^*, such that after training with it results in a student model with a lower error than any other S [34].

$$\forall S, E[L; \theta_{T+S^*}] < E[L; \theta_{T+S}] \tag{10}$$

However, this requires re-training the model for every possible subset and evaluating the expectation, which is practically infeasible. LL4AL [51] proposes a greedy solution by selecting samples with the highest loss from the unlabeled set U and optimizing it with supervised learning after labeling. This way, the remaining set $U - S$ has a smaller expected loss. The greedy selection score $s(x)$ can be defined as,

$$s(x) = L(f_s(x;\theta_T), y) \tag{11}$$

Note that this criterion depends on the current model parameters θ_T, trained with dataset T. Following Bayesian AL [14], we argue that the optimal selection score should not depend on a specific parameter value but the expectation over the parameters E_θ for a weight distribution $p(\theta|D)$, due to the stochastic nature of training with random initialization and data shuffles. We decompose the loss-based criteria as follows (derivation in the supplement):

$$\begin{aligned} s(x) = E_\theta[(f_s(x;\theta) - E_\theta[f_s(x;\theta)])^2] \\ + (E_\theta[f_s(x;\theta)] - h_s(x))^2 \\ + u_s^{al}(x) \end{aligned} \tag{12}$$

Selecting based on the total loss, like LL4AL, leads to selecting samples with high aleatoric uncertainty u_s^{al} that potentially harms the optimization. Therefore, we propose a selection criterion focusing on the first two components. The first component corresponds to the epistemic uncertainty of the student model. Using epistemic uncertainty as an AL scoring function is well-researched [9, 46]. However, previous AL methods cannot capture the second term without the ground-truth or $h_s(x)$. We propose using the teacher model's predictions as an estimate to $h_s(x)$ and define a new selection score, the inconsistency between the teacher and the student.

To have a coherent selection score with our semi-supervised training objective, we propose to select samples with high teacher aleatoric uncertainty u_t^{al} for annotation instead of generating pseudo-labels. Recall that these samples harm the optimization of the student, and with this selection score, the remaining samples in U contain pseudo-labels with high confidence.

Fig. 3b presents the selection phase in MonoLiG consisting of epistemic uncertainty of the student model, teacher-student inconsistency, and aleatoric uncertainty of the teacher.

Epistemic uncertainty of student model. Following ensembling techniques to capture epistemic uncertainty [4, 17], we estimate E_θ using an ensemble of five models trained with different random initialization. Following [26],

we match and cluster ensemble predictions using intersection over union (IoU). Within each cluster, we sum the variances of each regression parameter to obtain the total variance $u_s^{tv}(x)$.

$$u_s^{tv}(x) = E_\theta[(f_s(x;\theta) - E_\theta[f_s(x;\theta)])^2] \quad (13)$$

Teacher-student inconsistency. Following the teacher-student paradigm in SSL, we propose to use $f_t(x)$ as an estimate to $h_s(x)$. Using the matching algorithm with IoU, we match predictions from the teacher with predictions from different ensemble members. Then we define teacher-student inconsistency i_{ts} as the difference between the teacher model's regression parameters and the mean of the regression parameters from ensembles of student models:

$$i_{ts}(x) = (E_\theta[f_s(x;\theta)] - f_t(x))^2 \quad (14)$$

Selection strategy. We propose the selection score of MonoLiG as a combination of the three aforementioned scores as follows:

$$s(x) = (u_s^{tv}(x) + i_{ts}(x)) * (u_t^{al}(x)) \quad (15)$$

We sum u_s^{tv} and $i_{ts}(x)$ based on formulation in Eq. (12) and multiply with u_t^{al} due to difference in scales. Then, we aggregate the object-based score by taking the maximum score of objects to obtain a sample selection score.

4. Experiments

4.1. Experimental setup

Datasets and evaluation metric. We present our evaluation results on two autonomous driving datasets with synchronized LiDAR and camera frames and 3D bounding box labels: KITTI [16], and the Waymo Open Dataset [44].

KITTI contains 7481 images for training and 7518 samples for testing. Since labels of the test set are unavailable, we further split the training set following [23], which results in 3712 training and 3769 validation samples. For the AL scoring comparison and the ablation study, we report on the validation set and present the performance of our semi-supervised training strategy on the test set. We report BEV AP and 3D AP with 40 recall points on the car class for moderate difficulty with a 0.7 IoU threshold. We also present results on a larger scale Waymo Open Dataset which contains 798 training and 202 validation sequences. Following CaDDN [32], we downsample the original training set by selecting every third frame, resulting in a training set of approximately 51K samples labeled with 3D bounding boxes. For the Waymo dataset, we present our results using the official Level 2 mAP metric with 0.5 IoU.

Active learning details. For KITTI, we randomly split the training set for each experiment into a 30% labeled pool as an initial training dataset and a 70% unlabeled pool. The initial 30% of training data is used to pre-train the model. At each AL cycle, we compute scores on all samples in the unlabeled pool and select the 10% with the highest score to add to the training set. To imitate labeling, we use the already available annotations. For Waymo, since it contains more samples, we start with a training set with 5% samples and, at each cycle, add 5%. We present the mean of the corresponding metrics for three experiments with three different random initial training datasets.

Model architectures. For our AL experiments, we use the SOTA DD3D [28] as our student model and the well-established PV-RCNN [38] as our teacher model. We train for the same number of epochs and use the same hyperparameters and optimization scheme described in their respective papers. All experiments are conducted on an NVIDIA Tesla V100 GPU with PyTorch [29].

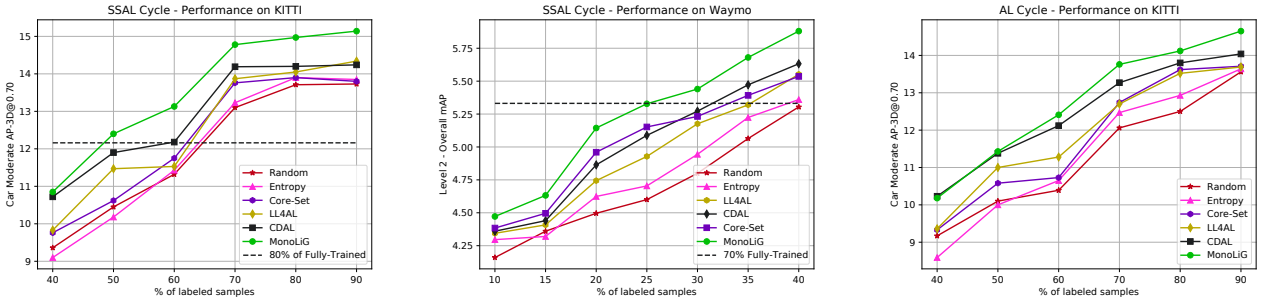
4.2. Comparison with AL selection baselines

With semi-supervised training. To demonstrate the effectiveness of our MonoLiG framework, we perform an evaluation to other AL selection methods using the same semi-supervised training phase. Specifically, we compare our approach with six baseline methods: **Random** sampling, **Entropy** sampling, a diversity-based **Core-Set** method [36], a task-agnostic **LL4AL** method [51], and the state-of-the-art **CDAL** [1] method. We also include the AP of a "fully-trained" detector, which is trained on the entire training set, to demonstrate the detector's performance capability.

In Fig. 4a, we present the comparison with AL methods for KITTI. Our method outperforms all the uncertainty-based baselines by at least 1.02 3D AP in the first AL cycle. As the number of actively selected labels increases, our method outperforms Random by 1.61 and the second-best method, CDAL, by 0.75. In the final cycle, where we use 90% of all the available data, with 60% of it actively labeled, our method outperforms all methods by at least 6.32%. Our approach reaches 80% of the fully-trained performance using only 48% of the data, compared to 60% of CDAL and 65% of random selection. This corresponds to a 17% improvement in data savings. We consistently outperform LL4AL in all cycles, validating that our approach of decomposing the loss function and ignoring aleatoric uncertainty leads to a better selection strategy.

In Fig. 4b, we present the results for the Waymo dataset, which is larger and thus more intuitively benefits from AL data selection. Towards the end, we reach 4.4% higher than the second-highest performing CDAL and 10.9% higher than random selection. Our approach reaches 70% of the fully-trained performance using only 25% of available data, compared to CDAL at 32% and random selection at 40%, corresponding to an 15% better data saving rate.

With supervised training. To evaluate the effectiveness



(a) With semi-supervised training on KITTI (b) With semi-supervised training on Waymo (c) With supervised training on KITTI

Figure 4. Comparison with SOTA AL methods. Lines indicate the averaged results over three trials. Note that all methods start from the same network trained with the initial labeled data, corresponding to 30% for KITTI and 5% for Waymo.

of our selection criteria in the absence of a coherent training strategy, we conduct an evaluation by comparing it with the same baselines, but using only supervised learning. In Fig. 4c, our selection criterion achieves better results than other uncertainty-based selection scores by at least 0.77 3D AP. As we actively selected more samples, we reach 4.34% higher than CDAL and 8.03% higher than Random.

4.3. Comparison with semi-supervised learning

We compare the performance of the monocular 3D detector on the KITTI *test* to our semi-supervised training strategy. Following [30], we train with the entire KITTI training dataset and, as the unlabeled pool, use KITTI raw scenes, excluding the samples from the validation set. This dataset is called the KITTI-depth set and contains approximately 26K samples. For a fair comparison against LPCG-MonoFlex [30], we use MonoFlex as the student model.

Table 1 shows the results of our method compared to other SOTA monocular 3D detectors. Among the methods that use semi-supervised LiDAR guidance, our approach reaches +2.02 and +4.24 BEV AP than the SOTA LPCG and MonoDistill, respectively. Considering the performance weighted by the number of samples in each case, MonoLiG has a higher overall AP of 28.62 compared to 26.94 of LPCG. Using our semi-supervised training strategy, MonoFlex, proposed in 2019, lagging behind the current SOTA detector MonoDDE by 3.71 AP, reaches 3.37 higher performance.

We further investigate how our pseudo-label weighting strategy compares to other pseudo-label filtering strategies from the literature. We compare with **FixMatch** [42], which uses the confidence score to filter out uncertain pseudo-label, and **3DIoUMatch** [47], which, in addition to the confidence score, uses an estimated IoU for filtering. We follow our approach of scaling the pseudo-labels with the corresponding uncertainty and present the results in Table 2. Our pseudo-label uncertainty approach reaches the highest performance, reaching 7.82 AP higher than the base model and

Approaches	Extra	Mod.	Easy	Hard
M3D-RPN [5]	-	13.67	21.02	10.23
MonoRUn [8]	-	17.34	27.94	15.24
DDMP-3D [49]	KD	17.89	28.08	13.44
PCT [50]	KD	19.03	29.65	15.92
MonoFlex [54]	-	19.75	28.23	16.89
MonoDTR [19]	-	20.38	28.59	17.14
DID-M3D [31]	-	22.76	32.95	19.83
DD3D [28]	DDAD	23.41	32.35	20.42
MonoDDE [25]	-	23.46	33.58	20.37
MonoDistill [10]	-	22.59	31.87	19.72
LPCG [30]	KD	<u>24.81</u>	**35.96**	<u>21.86</u>
MonoLiG-SSL	KD	**26.83**	<u>35.73</u>	**24.24**

Table 1. Comparison of BEV detection results on KITTI *test* for monocular detectors. Note that both LPCG and MonoLiG use MonoFlex as the base detector. MonoDistill, LPCG, and MonoLiG are semi-supervised methods using additional information from LiDAR during training. KD and DDAD represent the extra datasets, KITTI-depth and DDAD15M [28], respectively. Bold and underlined values represent best and second best respectively.

1.59 AP higher than 3DIoUMatch. Furthermore, compared to 3DIoUMatch, our approach reduces additional time per iteration by 16% and additional memory consumption by 3-fold. We observe that methods that consider localization uncertainty (3DIoUMatch, MonoLiG) perform better than methods that only use classification uncertainty (FixMatch).

In Fig. 5, we compare qualitatively how different pseudo-labeling uncertainty strategies work. In the first row, our approach and 3DIoUMatch filter out different vehicles, both difficult to localize for a monocular 3D object detector. We observe that our predictions correspond to more distant and occluded objects with low LiDAR point returns and, therefore, higher aleatoric uncertainty. Also, in the second row, the middle car with high bounding box localization error has low classification uncertainty but high localization

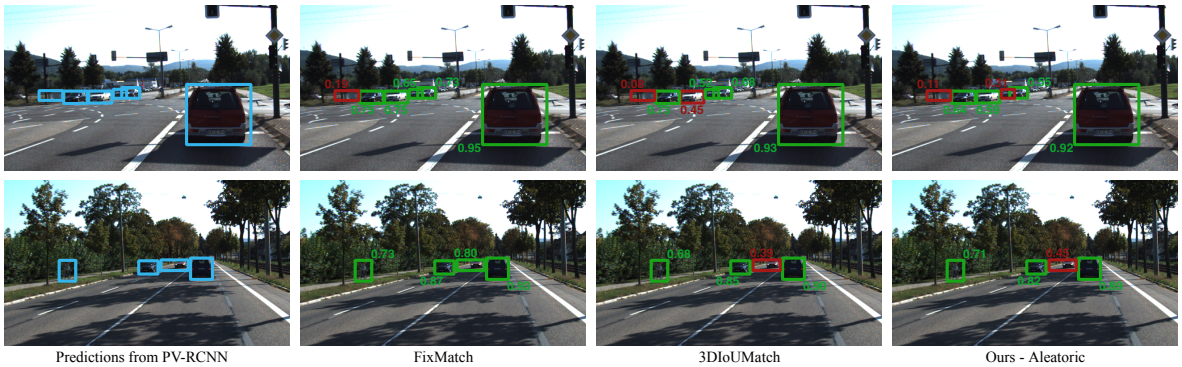


Figure 5. Qualitative comparisons of pseudo-label confidence algorithms (FixMatch, 3DIoUMatch, Ours) on KITTI. Blue boxes represent predictions from the PV-RCNN detector. Red and green boxes represent boxes with a confidence score below and above 0.5 threshold for the corresponding pseudo-label confidence approach.

uncertainty. For these types of objects, utilizing a strategy that also identifies localization uncertainty is essential.

Approaches	Mod.	Time (ms)	Memory
DD3D [28]	16.92	53.4	-
No Confidence	20.14	53.4	-
FixMatch [42]	22.59	53.4	-
3DIoUMatch [47]	23.15	62.9(+17.8%)	696 MB
Ours - Aleatoric	**24.74**	54.2(+1.6%)	184 MB

Table 2. Comparison of different pseudo-labeling confidence algorithms on KITTI *val*.

4.4. Ablation studies

Ablation on scoring. Next, we study the effect of each component in our MonoLiG framework. We present our ablation study with different combinations for our selection in the supervised-learning setting in Table 3. Our findings indicate that, among the single scores, i_{ts} performs best in the initial cycles, highlighting the effectiveness of utilizing teacher predictions when the student uncertainty is not yet well-learned. When all three scores are combined, we obtain the highest performing selection score.

$s(x)$	40%	50%	70%	90%
u_t^{al}	9.12	10.07	11.44	12.94
u_s^{tv}	8.86	10.51	12.64	14.12
i_{ts}	9.53	11.13	12.68	13.89
$u_s^{tv} + i_{ts}$	9.22	10.73	13.39	14.44
$(u_s^{tv} + i_{ts}) * u_t^{al}$	**10.18**	**11.43**	**13.76**	**14.65**

Table 3. Ablation study on selection scores and training strategy on KITTI *val*. The percentage indicates the ratio of labeled data.

Different teacher-student architectures. To show the robustness of MonoLiG to the architecture choice of the teacher and the student model, we try with two different monocular detectors: DD3D [28] and MonoFlex [54] as our student model and two different LiDAR detectors as our teacher model: PV-RCNN [38] and PointPillars [23]. In Table 4, we see that MonoLiG boosts the performance of both student models compared to the random sampling strategy. We observe a performance gain of 1.24 AP at 80% data percentage for DD3D and even a further 1.85 AP for the MonoFlex. We also observe that for the choice of teacher model, the better the teacher model is, the higher the performance gain when MonoLiG is used.

	DD3D [28]		MonoFlex [54]	
	40%	80%	40%	80%
Base	7.58	12.47	5.22	10.08
PointPillars [23]	8.91	12.89	7.35	10.92
PV-RCNN [38]	**9.36**	**13.71**	**7.81**	**11.33**

Table 4. Performance comparison of MonoLiG with two different LiDAR and monocular detectors on KITTI *val*. The percentage indicates the ratio of labeled samples.

5. Conclusion

We introduced a novel SSAL framework. MonoLiG consists of a novel training phase that uses aleatoric uncertainty weighted pseudo-labels from the LiDAR detector to guide the training of the monocular detector and a selection phase with a novel acquisition function based on the inconsistency between predictions from the LiDAR and the monocular detector. Our extensive experiments validate the effectiveness of MonoLiG compared to both AL and SSL baselines. We further showed that MonoLiG could easily be adapted to any monocular detector. Pseudo-labels' quality is essential for our framework; we will further explore how to generate more precise pseudo-labels by adding more modalities, e.g., radar and tracking over time.

References

[1] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *ECCV*, 2020. 2, 6

[2] Hamed H Aghdam, Abel Gonzalez-Garcia, Joost van de Weijer, and Antonio M López. Active learning for deep detection neural networks. In *ICCV*, 2019. 2

[3] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *IJCNN*, 2020. 3

[4] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *CVPR*, 2018. 2, 5

[5] Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection. In *ICCV*, 2019. 7

[6] Benjamin Caine, Rebecca Roelofs, Vijay Vasudevan, Jiquan Ngiam, Yuning Chai, Zhifeng Chen, and Jonathon Shlens. Pseudo-labeling for scalable 3d object detection. *arXiv preprint arXiv:2103.02093*, 2021. 3

[7] Cong Chen, Shouyang Dong, Ye Tian, Kunlin Cao, Li Liu, and Yuanhao Guo. Temporal self-ensembling teacher for semi-supervised object detection. *Transactions on Multimedia*, 2021. 3

[8] Hansheng Chen, Yuyao Huang, Wei Tian, Zhong Gao, and Lu Xiong. Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation. In *CVPR*, 2021. 7

[9] Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clément Farabet, and José Manuel Álvarez. Active learning for deep object detection via probabilistic modeling. In *ICCV*, 2021. 2, 5

[10] Zhiyu Chong, Xinzhu Ma, Hong Zhang, Yuxin Yue, Haojie Li, Zhihui Wang, and Wanli Ouyang. Monodistill: Learning spatial features for monocular 3d object detection. In *ICLR*, 2022. 3, 4, 7

[11] 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 *CVPR*, 2022. 2

[12] Di Feng, Ali Harakeh, Steven L Waslander, and Klaus Dietmayer. A review and comparative study on probabilistic object detection in autonomous driving. *T-ITS*, 2021. 2

[13] Di Feng, Xiao Wei, Lars Rosenbaum, Atsuto Maki, and Klaus Dietmayer. Deep active learning for efficient training of a lidar 3d object detector. In *IV*, 2019. 2

[14] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *ICML*, 2017. 2, 5

[15] Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö Arık, Larry S Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *ECCV*, 2020. 3

[16] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1, 6

[17] Elmar Haussmann, Michele Fenzi, Kashyap Chitta, Jan Ivanecky, Hanson Xu, Donna Roy, Akshita Mittel, Nicolas Koumchatzky, Clement Farabet, and Jose M Alvarez. Scalable active learning for object detection. In *IV*, 2020. 2, 5

[18] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011. 2

[19] Kuan-Chih Huang, Tsung-Han Wu, Hung-Ting Su, and Winston H. Hsu. Monodtr: Monocular 3d object detection with depth-aware transformer. In *CVPR*, 2022. 7

[20] Siyu Huang, Tianyang Wang, Haoyi Xiong, Jun Huan, and Dejing Dou. Semi-supervised active learning with temporal output discrepancy. In *ICCV*, 2021. 3

[21] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *NeurIPS*, 2019. 3

[22] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017. 3

[23] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, 2019. 6, 8

[24] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICMLW*, 2013. 3

[25] Xinzhu Ma, Yinmin Zhang, Dan Xu, Dongzhan Zhou, Shuai Yi, Haojie Li, and Wanli Ouyang. Delving into localization errors for monocular 3d object detection. In *CVPR*, 2021. 7

[26] Dimity Miller, Feras Dayoub, Michael Milford, and Niko Sünderhauf. Evaluating merging strategies for sampling-based uncertainty techniques in object detection. In *ICRA*, 2019. 5

[27] Daniele Mugnai, Federico Pernici, Francesco Turchini, and Alberto Del Bimbo. Soft pseudo-labeling semi-supervised learning applied to fine-grained visual classification. In *ICPR*, 2021. 3

[28] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *ICCV*, 2021. 6, 7, 8

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 6

[30] Liang Peng, Fei Liu, Zhengxu Yu, Senbo Yan, Dan Deng, Zheng Yang, Haifeng Liu, and Deng Cai. Lidar point cloud guided monocular 3d object detection. In *ECCV*, 2022. 1, 3, 4, 7

[31] Liang Peng, Xiaopei Wu, Zheng Yang, Haifeng Liu, and Deng Cai. Did-m3d: Decoupling instance depth for monocular 3d object detection. In *ECCV*, 2022. 7

[32] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *CVPR*, 2021. 6

[33] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *ICLR*, 2021. 3

[34] Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. In *ICML*, 2001. 5

[35] Sebastian Schmidt, Qing Rao, Julian Tatsch, and Alois Knoll. Advanced active learning strategies for object detection. In *IV*, 2020. 2

[36] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018. 2, 6

[37] Claude Elwood Shannon. A mathematical theory of communication. *Mobile Computing and Communications Review*, 2001. 2

[38] 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 *CVPR*, 2020. 6, 8

[39] Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng. Transductive semi-supervised deep learning using min-max features. In *ECCV*, 2018. 3

[40] Oriane Siméoni, Mateusz Budnik, Yannis Avrithis, and Guillaume Gravier. Rethinking deep active learning: Using unlabeled data at model training. In *ICPR*, 2021. 3

[41] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *ICCV*, 2019. 2

[42] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020. 3, 7, 8

[43] Shuang Song, David Berthelot, and Afshin Rostamizadeh. Combining mixmatch and active learning for better accuracy with fewer labels. *arXiv preprint arXiv:1912.00594*, 2019.

[44] 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 *CVPR*, 2020. 6

[45] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 3

[46] Evgenii Tsymbalov, Maxim Panov, and Alexander Shapeev. Dropout-based active learning for regression. In *AIST*, 2018. 2, 5

[47] He Wang, Yezhen Cong, Or Litany, Yue Gao, and Leonidas J Guibas. 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. In *CVPR*, 2021. 3, 7, 8

[48] Jun Wang, Shaoguo Wen, Kaixing Chen, Jianghua Yu, Xin Zhou, Peng Gao, Changsheng Li, and Guotong Xie. Semi-supervised active learning for instance segmentation via scoring predictions. In *BMVC*, 2020. 3

[49] Li Wang, Liang Du, Xiaoqing Ye, Yanwei Fu, Guodong Guo, Xiangyang Xue, Jianfeng Feng, and Li Zhang. Depth-conditioned dynamic message propagation for monocular 3d object detection. In *CVPR*, 2021. 7

[50] Li Wang, Li Zhang, Yi Zhu, Zhi Zhang, Tong He, Mu Li, and Xiangyang Xue. Progressive coordinate transforms for monocular 3d object detection. In *NeurIPS*, 2021. 7

[51] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *CVPR*, 2019. 2, 5, 6

[52] Weiping Yu, Sijie Zhu, Taojiannan Yang, and Chen Chen. Consistency-based active learning for object detection. In *CVPR*, 2022. 2

[53] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *CVPR*, 2021. 2

[54] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *CVPR*, 2021. 7, 8