2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Joint Semi-Supervised and Active Learning via 3D Consistency for 3D Object Detection

Sihwan Hwang, Sanmin Kim, Youngseok Kim, and Dongsuk Kum

***Abstract*— Autonomous driving powered by deep learning requires large-scale, high-quality training data from diverse driving environments to operate effectively worldwide. However, collecting and annotating such data is costly and time-consuming. To address this challenge, active learning methods have been explored to select the most informative data samples for training. Nevertheless, most existing methods focus on 2D tasks and do not fully exploit the value of unlabeled data. In this paper, we propose a semi-supervised active learning approach for 3D object detection tasks that leverages the potential of collected data and reduces annotation costs. Our method considers the 3D consistency of bounding box predictions in both semi-supervised and active learning processes, thereby improving the performance of point cloud-based 3D object detection models. Our framework specifically utilizes self-supervision to decrease bounding box uncertainties. Moreover, it selects objects that are either occluded or distant and still exhibit high uncertainty for annotation even after semi-supervised training has decreased their uncertainty. Experiments on the KITTI dataset demonstrate that our semi-supervised active learning approach selects objects with high measurement uncertainties and enhances the model's ability to detect occluded objects. Our approach improves the baseline by more than 60% (+17.12 mAP) when using only 1500 annotated frames.**

I. INTRODUCTION

Detecting objects in 3D space is a fundamental and essential task in autonomous driving. Since the performance of a learning-based 3D detection network depends on large-sized and high-quality training data, massive human efforts are required to collect and annotate data. Accordingly, several active learning approaches have been proposed to select and label data that is effective for model training to efficiently reduce the annotation cost.

Active learning approaches either select data points with high prediction uncertainty of the model [1], [12], [16] or data samples that are most distant from previously selected samples for diversity [7], [8], [9]. However, most of these active learning methods are for image-based 2D classification and object detection, which has several distinctions from LiDAR-based 3D object detection for autonomous driving.

First, the number of training data samples for 3D detection is limited compared to 2D tasks, even though the 3D detection task demands a large number of data samples due to its complexity. The difficulty of collecting and annotating 3D data limits the availability of labeled data. On this account, *active learning may encounter challenges at the initial training stages*, as the number of training data samples is insufficient, leading to inaccurate measurement of uncertainty when selecting data. Second, localization uncertainty plays a more critical role than classification uncertainty in 3D detection tasks compared to 2D tasks. Since the number of classes in autonomous driving datasets is only a few (*e.g.*, KITTI [42] and nuScenes [43] have 3 and 10 classes) compared to image 2D detection datasets (*e.g.*, Pascal VOC [40] and COCO [41] have 20 and 80 classes), the class imbalance problem is less significant in 3D detection. Also, the number of attributes representing a 2D bounding box is smaller (4DoF: 2D offset, 2D size) than the 3D bounding box (7DoF: 3D offset, 3D size, rotation). Therefore, *the majority of errors in 3D detection arise from localization*. However, existing 2D-based methods primarily emphasize measuring uncertainty from classification, given the significance of ensuring class diversity.

To address the discrepancies between 2D and 3D detection in active learning, we propose a semi-supervised active learning method for LiDAR-based 3D object detection that leverages unlabeled data to overcome data shortage and accounts for the uncertainty of bounding box regression. Our semi-supervised active learning is inspired by [21], which originally targets image-based 2D detection, while ours focuses on the 3D detection task. We define self-supervised consistency using diverse input data augmentation strategies in order to utilize both labeled and unlabeled samples in the training stage. Furthermore, we introduce an active selection method that can handle both class and 3D bounding box uncertainty. The proposed method combines active learning and semi-supervision from the measurement of 3D consistency by training the detection model to predict objects consistently and selecting data with low consistency. Our method presents a significant improvement of the baseline on the KITTI autonomous driving dataset.

II. RELATED WORK

Active Learning for 2D Object Detection: Active learning methods for object detection tasks formulate a scoring function to measure the importance of each detected object and aggregate these scores into frame-level. The approaches are usually divided into either diversity-based or uncertainty-based approaches based on the objective of the scoring function. The diversity-based approaches [9], [8] focus on selecting new and different samples to make the training set as diverse as possible. However, the diversity-based

*This work was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) and the National Research Foundation of Korea(NRF) funded by Korea Government (Ministry of Science and ICT) under Grants 2021-0-00951 and 2022R1A2C200494412.

Authors are with Cho Chun Shik Graduate School of Mobility, KAIST, Daejeon 34051, Republic of Korea. E-mail: {shhwang0129, sanmin.kim, youngseok.kim, dskum}@kaist.ac.kr

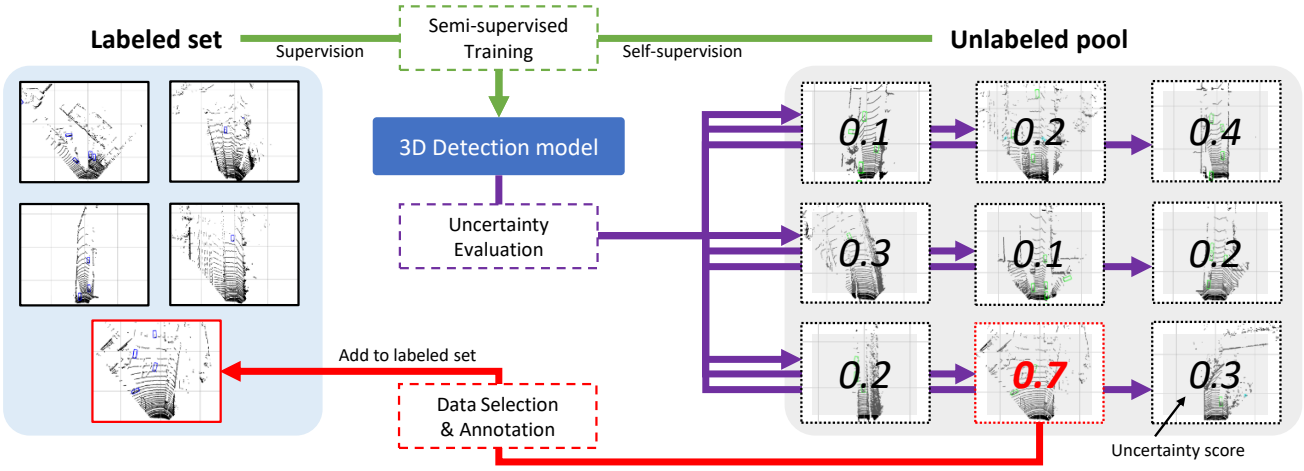


Fig. 1. Overall framework of our method. The framework consists of two alternating stages of semi-supervised training (green) and active data selection stage (purple, red). In the first semi-supervised training, the model trains on both labeled and unlabeled datasets by utilizing additional self-supervised loss to predict 3D boxes consistent with data transformations. In the following active data selection, each unlabeled data is evaluated using the uncertainty in the model prediction. Unlabeled samples with the highest uncertainties are selected for annotation by humans, which are then used for supervision in the next training stage.

approaches only consider the data distribution while ignoring the difficulty of samples.

The uncertainty-based approaches [10]−[16], on the other hand, focus on selecting samples with the highest uncertainty in the model prediction by measuring the entropy [34], [11] or probability of class [10]. These approaches can be categorized into two: single-model-based and multiple-model-based approaches. 'Query by Committee' approach [14] estimates the uncertainty from the difference in class probability between feature layers, whereas the learning loss [12] proposes a loss prediction module that attaches to the feature layers to predict the training loss that can be used as the uncertainty. Choi et al. [16] utilizes mixture density networks [33] to estimate aleatoric and epistemic uncertainties. Even though single model-based uncertainty approaches show fast inference time, their performances are inferior to those of multiple model-based approaches. Multiple model-based approaches apply perturbations either to the model parameters via Monte Carlo (MC) Dropout [36] or Deep Ensembles [24], [15] or to the data via augmentation [13], to generate multiple predictions on a single frame and calculate predictive uncertainty on the matched predictions. These multiple model-based methods generally perform better than single model-based methods in exchange for computation time or memory.

Semi-supervised Active Learning: Recently, several works leverage unlabeled data to additionally train the model in a semi-supervised manner, in classification task [19], [18], [20], 2D object detection task [17], [21], and segmentation task [22]. Following the seminal work [19], which introduced the combining method of semi-supervision and active learning in text classification, semi-supervised active learning is adopted in the image domain. In the image classification task, Gu et al. [20] propose using Local and Global Consistency for semi-supervised learning. Gao et al. [18] utilized a perturbation scheme in both the training and data selection stage to obtain multiple predictions, which allows consistency regularization in semi-supervision and uncertainty measurement from predicted variance. The recent work in 2D object detection [17], [21] additionally utilizes unlabeled data for training. MI-AOD [17] trains on unlabeled data via minimizing and maximizing the discrepancy between class predictions and selects data using the prediction discrepancy. Elezi et al. [21] trains via consistency regularization and pseudo-labeling and select data by defining the multiplication of entropy and KL-divergence as the class consistency loss as the uncertainty.

Our work is inspired by [21] to train the 3D object detector in a semi-supervised manner in the training stage. However, these methods measure uncertainty from class predictions which is insufficient for the case of 3D object detection since uncertainty in regression is more significant than that of 2D object detection.

Active Learning for 3D Object Detection: There is a couple of active learning approaches for 3D object detection tasks [25], [23], [24] that utilize 3D LiDAR point clouds. Segal et al. [25] simply utilize the entropy to measure uncertainty but propose using partial labeling and training to annotate only informative instances. Both [23] and [24] utilize multiple models by giving perturbation to the model parameters via ensemble [24], [23] or MC-Dropout [23] to obtain multiple set of predictions per each data point. From the matched set of predictions, uncertainties are measured from class predictions. Schmidt et al. [24] propose four different uncertainty estimation methods for 2D object detection, including the uncertainty based on bounding box overlap. Still, only classification uncertainty of variation ratio is used for 3D object detection as the proof of concept. Our uncertainty measurement considers 3D box uncertainty to

select hard samples, such as objects with high occlusions. Since 3D box uncertainty of easy samples can be minimized from self-supervised training, this results in selecting only objects that require supervision from annotation.

III. METHOD

A. Overview

Our semi-supervised active learning framework consists of two stages: the semi-supervised training and active data selection, as illustrated in Fig. 1. In the semi-supervised training stage, we train the detection model with labeled and unlabeled data, utilizing consistency regularization. To achieve this, we fed two augmented data points with stochastic transformations to the model, which enables consistency regularization and multiple-model-based uncertainty measurement. In the active data selection stage, we evaluate objects detected by the trained model in terms of entropy and 3D bounding box uncertainty. Subsequently, we aggregate the uncertainty scores of each frame to obtain a frame-level score to select frames with the highest uncertainty for annotation. These two stages iterate alternatively until the label budget is reached, starting with the randomly selected initial training set. Our proposed method effectively selects data samples with high uncertainty for annotation, thereby improving detection accuracy.

B. Semi-supervised Training

To take advantage of both labeled and unlabeled data during the training stage, a consistency regularization approach [26] is adopted for both labeled and unlabeled data, in addition to the supervision of labeled data. The consistency regularization makes the model generate consistent predictions when a small perturbation is applied to the input or the model parameters. To this end, we give stochastic transformations to input as a perturbation which consists of flipping along the x-axis, rotation, and scaling with a probability of $\mathcal{F}$, θ, and α, respectively. After the detection model predicts 3D objects from both the original and transformed input, predicted 3D objects are matched for consistency calculation similar to [28].

For the matching, we first apply an inverse transformation to predictions from transformed inputs so that two original and transformed outputs are in the same coordinate system. Afterward, predictions with confidence less than τ_{conf} are filtered to prevent false positives that might damage the training stability. Then IoU (Intersection over Union) between the remaining predictions is measured, and pairs with a larger IoU than a threshold τ_{IoU} are considered as a matched pair for consistency regularization.

For every matched prediction pair, consistency loss $\mathcal{L}_{cons}$ is defined as inconsistency between the class and box predictions. We define the class consistency loss $\mathcal{L}^{cls}_{cons}$ with Kullback-Leibler (KL) divergence to minimize the difference between the predicted class probability distributions as:

$$\mathcal{L}^{cls}_{cons} = \frac{1}{|P^o|}\sum_{p^t} D_{KL}(p^t||p^o), \tag{1}$$

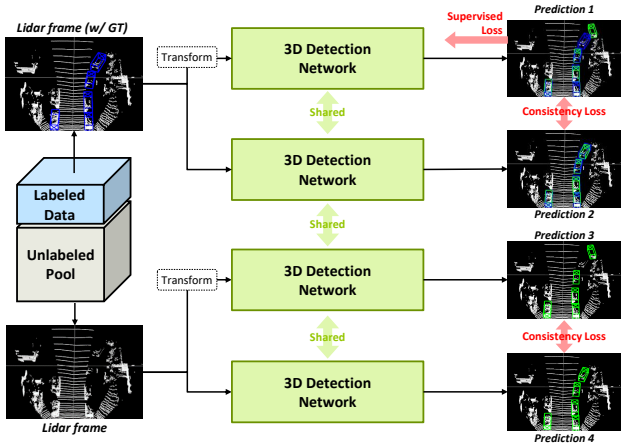


Fig. 2. Semi-supervised training process. Our method utilizes self-supervised consistency loss to train on unlabeled data. Each data point is augmented using stochastic transformations before passing to the model to predict pairs of predictions per data point.

where $P^i = \{p^i\},\ i \in \{t,o\}$ denotes class probability of transformed and original bounding boxes.

For box consistency loss, we use Smooth-L1 $\mathcal{L}^{box}_{cons}$ to minimize misalignment between two predicted bounding boxes. The aligned set of 3D bounding box $B^i = \{b^i\},\ i \in \{t,o\}$ consists of 7 parameters $b^i = \{x,y,z,w,h,l,r\}$, where (x,y,z), (w,h,l), and r denote location, dimension, and orientation.

$$\begin{aligned}\mathcal{L}^{reg}_{cons} &= \frac{1}{|B^o|}\sum_{b^t}\sum_{e}\frac{1}{7}\text{smooth}_{L_1}(\delta_e(b^t,b^o)) \\ \delta_e(b^t,b^o) &= \begin{cases} |e^t - e^o| & \text{if } e \in \{x,y,z,w,l,h\} \\ |\sin(e^t - e^o)| & \text{if } e \in \{r\}. \end{cases}\end{aligned} \tag{2}$$

Once the overall consistency loss $\mathcal{L}_{cons}$ is calculated by taking a weighted sum of regression and classification consistency losses, then the total loss is obtained by adding the supervised loss $\mathcal{L}_{super}$ following CenterPoint [30]:

$$\begin{aligned}\mathcal{L}_{cons} &= \lambda_1\mathcal{L}^{reg}_{cons} + \lambda_2\mathcal{L}^{cls}_{cons} \\ \mathcal{L}_{total} &= \mathbb{1}\mathcal{L}_{super} + \sigma\lambda_3\mathcal{L}_{cons}\end{aligned} \tag{3}$$

$$\mathbf{x}^* = \arg\min_{\mathbf{x}}\sum \mathbf{e}_i^T(\mathbf{x})\Omega_i\mathbf{e}_i(\mathbf{x}) \tag{4}$$

where $\lambda_1,\lambda_2,\lambda_3$ are weights to modulate the importance of each loss term, and $\mathbb{1}$ indicates the existence of annotation. Additionally, the weight of consistency loss $\mathcal{L}_{cons}$ is gradually increased by multiplying the sigmoid-shaped ramp-up function $\sigma = e^{-5(1-T)^2}$ as T increases from 0 to 1 linearly in the earlier training epochs as suggested in [29].

In the case of the unlabeled data, we use only consistency loss to train the model since annotations to compute supervised loss are unavailable. We stack both labeled and unlabeled data points into mini-batches with a ratio of $1:N_u$. We apply only a flip transformation to the unlabeled data points to ease the perturbation and increase the probability of matching.

TABLE I

COMPARISON OF ENTROPY ($\mathcal{H}$) AND MUTUAL INFORMATION ($\mathcal{MI}$).

	Probability of each class car	ped.	cyc.	$\mathcal{H}(\mathbf{p}^i)$	$\mathcal{H}(\bar{\mathbf{p}})$	$\mathcal{MI}(\mathbf{p}^t, \mathbf{p}^o)$
$\mathbf{p}^t$	0.01	0.98	0.01	0.1614	1.0708	0.9093
$\mathbf{p}^o$	0.01	0.01	0.98	0.1614		
$\mathbf{p}^t$	0.32	0.35	0.33	1.5839	1.5847	0.0003
$\mathbf{p}^o$	0.33	0.33	0.34	1.5848		

C. Active Data Selection

In the active data selection stage, the informativeness of each data point in the unlabeled pool has to be evaluated. To this end, we measure the uncertainty between prediction pairs that are matched in the previous section. The difference from the training stage is that only the flip operation is applied to the transformed input with 100% probability to keep the transformation fixed for evaluation. Considering the fact that the performance of LiDAR-based 3D object detection depends on localization error than classification error, we measure the 3D bounding box uncertainty in addition to the class uncertainty.

For measuring the class uncertainty, we first examine the entropy and mutual information used in [23], which are expressed in (5) and (6), respectively. The p_c refers to the class probability distribution of class c.

$$\mathcal{H}(\bar{\mathbf{p}}) = -\sum_c \bar{\mathbf{p}}_c \log \bar{\mathbf{p}}_c, \text{ where } \bar{\mathbf{p}}_c = \frac{1}{2}\sum_{i\in\{t,o\}} \mathbf{p}_c^i \tag{5}$$

$$\mathcal{MI}(\mathbf{p}^t, \mathbf{p}^o) = \mathcal{H}(\bar{\mathbf{p}}) - \frac{1}{2}\sum_{i\in\{t,o\}} \mathcal{H}(\mathbf{p}^i) \tag{6}$$

As shown in Table I, entropy and mutual information quantify different types of uncertainties. The first two rows show the case where two predictions are very confident but highly disagreed, and the others are the case where two predictions are less confident and yield similar class distributions. The mutual information captures the amount of disagreement between two predictions but fails to find predictions with low confidence. In contrast, the entropy can measure the lack of confidence in each prediction and capture the disagreement of matched pairs by taking an average of the two predictions. Therefore, we employ the entropy of matched predictions as class uncertainty measurement.

For the 3D bounding box uncertainty, we apply the loss-based uncertainty. Since the 3D center position or orientation errors come from a tricky situation, such as object occlusion or sparsity of point cloud, significant bounding box errors represent difficult samples. Therefore, loss-based 3D bounding box uncertainty effectively selects data for 3D object detection. For simplicity, we reuse training loss defined in (2) as the 3D bounding box uncertainty. The total uncertainty is defined as the multiplication of the two uncertainties. After obtaining the final uncertainties of detected objects, we aggregate uncertainties of the objects in each frame into frame-level by taking the sum. The frames with the top N highest uncertainties are selected for labeling and used for supervised training in the following cycle as illustrated in Fig. 3.

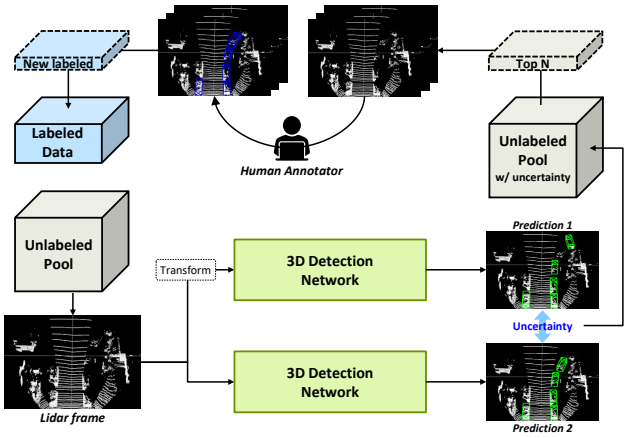


Fig. 3. Active data selection process. Following the perturbation scheme in the training stage to obtain multiple predictions per data, we evaluate the uncertainty of each prediction by measuring the difference between the matched predictions. After aggregation of the object-level uncertainties, we select the frames with the highest uncertainties for annotation.

IV. EXPERIMENTS

A. Implementation Details

We tested the proposed semi-supervised active learning framework on the KITTI 3D object detection dataset [42]. Following 3DOP [32], we split the training set containing 7,481 frames into train and validation splits of 3,712 and 3,769. We regard the training set as unlabeled even if the annotations exist unless selected at the selection stage. We start by randomly selecting 300 samples and adding 300 new ones at every active learning cycle until 1,500. We measure the detection performance at each active learning cycle using 3D mean Average Precision (mAP) of three classes (car, pedestrian, and cyclist) with moderate difficulty.

We used CenterPoint [30] as the baseline detector and discarded GT-AUG strategy [31] to eliminate the effect of using pre-labeled annotation. The network is trained for 40 epochs by an AdamW optimizer with a one-cycle learning

TABLE II

HYPER-PARAMETERS OF THE IMPLEMENTATION.

		Training stage	Selection stage
Augmentation	$\mathcal{F}$	0.5	1.0
	θ	45	0
	α	0.05	0
Matching	τ_{conf}	0.1	0
	τ_{IoU}	0.25	0.25
Consistency	λ_1	1	-
	λ_2	20	-
	λ_3	4	-
Unlabeled	N_u	1	-

TABLE III

COMPARISON OF UNCERTAINTY MEASUREMENTS ON ACTIVE LEARNING.

Cycles	Number of Frames	mAP (the higher is better) Random	Entropy [10]	Ensemble [23]	Ours (w/o SSL)	Ours
1	300	7.96 ±0.78	7.85 ±1.96	8.90 ±1.37	8.75 ±0.71	10.66 ±1.10
2	600	14.26 ±1.37	18.24 ±2.05	18.28 ±1.30	21.31 ±0.46	29.33 ±3.04
3	900	20.58 ±0.65	26.39 ±0.80	28.36 ±1.53	29.58 ±0.67	37.29 ±1.54
4	1200	24.79 ±1.04	30.13 ±0.58	33.45 ±0.94	32.87 ±1.24	40.71 ±1.75
5	1500	27.68 ±0.75	33.76 ±0.54	34.70 ±1.53	35.27 ±0.38	44.80 ±1.16

TABLE IV

ABLATION STUDY OF THE PROPOSED METHODS.

Uncertainty Class	Uncertainty 3D Box	Consistency D^L	Consistency D^U	mAP (the higher is better) 300 (Cycle 1)	600 (Cycle 2)	900 (Cycle 3)	1200 (Cycle 4)	1500 (Cycle 5)
				7.96 ±0.78	14.26 ±1.37	20.58 ±0.65	24.79 ±1.04	27.68 ±0.75
		✓		9.70 ±0.80	19.76 ±1.59	27.30 ±0.94	31.34 ±1.70	34.64 ±0.71
		✓	✓	10.23 ±1.32	23.03 ±0.21	29.96 ±1.87	32.11 ±1.27	38.61 ±2.35
✓				8.16 ±0.73	18.82 ±2.04	25.11 ±2.20	31.27 ±1.41	33.98 ±1.47
	✓			8.35 ±0.98	21.90 ±1.18	28.63 ±1.41	31.13 ±0.61	35.04 ±0.96
✓	✓			8.75 ±0.71	21.31 ±0.46	29.58 ±0.67	32.87 ±1.24	35.27 ±0.38
✓	✓	✓	✓	10.66 ±1.10	29.33 ±3.04	37.29 ±1.54	40.71 ±1.75	44.80 ±1.16

rate policy, with a max learning rate of 0.001, weight decay of 0.01, and momentum of 0.85 to 0.95. We conducted all experiments on an Intel Xeon Gold 5220 CPU and three Tesla V100 GPUs. We set the hyperparameters defined in the previous section as Table II. For training stage in Sec. III-B, we follow the augmentations proposed in CenterPoint [30] for transformation parameters $\{\mathcal{F}, \theta, \alpha\}$. At the selection stage in Sec. III-C, only the flip augmentation is applied with 100% probability while discarding others. For matching, we set τ_{conf} to filter objects with low confidence to 0.1 at training and 0 at selection, regarding the greater sensitivity of self-supervised training to false positives than uncertainty estimation. We set the τ_{conf} for both the training and selection stages to consider small object classes fully. We empirically find that hyper-parameters in Table II help stable and effective convergence in our experimental setting.

B. Comparisons with Active Learning Methods

To show the effectiveness of our joint semi-supervised active learning, we compare the performance with the previous methods in Fig. 4. Random refers to the model trained without active learning, where each frame is selected randomly. The entropy [10] measures the Shannon Entropy [34] of the single model's class probability distribution. The ensemble [23] uses ensembles of models to obtain multiple predictions and measure the entropy and mutual information of the matched predictions. In implementing the ensemble methods, we set the number of ensembles as two to keep the number of multiple predictions the same as ours. We also include the result without using consistency loss for a fair comparison with the baselines so that the only difference is the selected training data. As shown in Fig. 4, our method

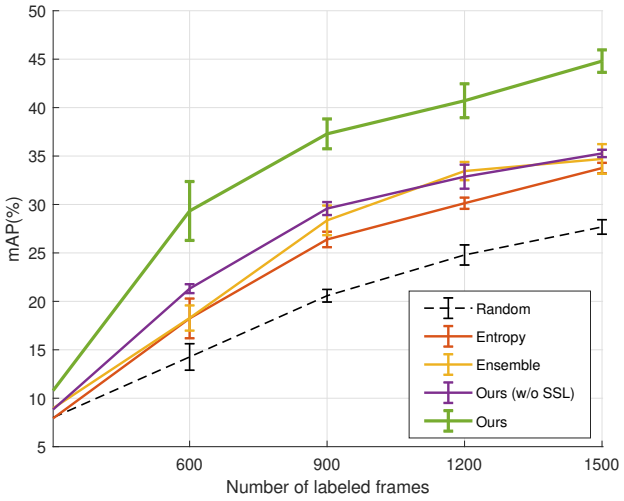
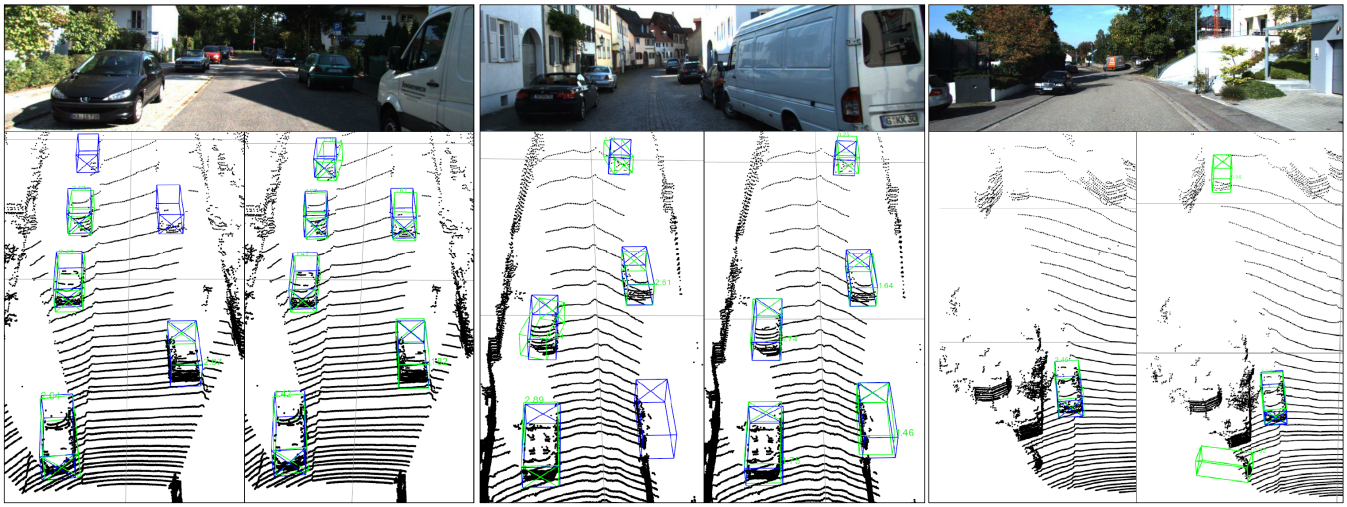


Fig. 4. 3D object detection performance at each active learning cycle with different uncertainty measurements. (Best viewed in color)

outperforms other methods from the initial cycle to the last cycle. Specifically, our method with semi-supervision achieves better performance than the ones without, which means semi-supervision can boost detection performance in the 3D object detection task. Even without semi-supervision, our approach (Ours w/o SSL) achieves higher performance, especially at the initial cycles, due to taking advantage of 3D bounding box uncertainty.

C. Ablation Study

We conduct an ablation study to show the effect of each proposed component on the KITTI Val set as shown in Table IV. We compare the performance of each method in the second cycle and the last cycle to show the effect of the first



(a) (b) (c)

Fig. 5. Visualization of detection results on KITTI validation set. In each case, the left is the baseline result, and the right is the result of the proposed method. Blue and green boxes denote ground truth and prediction, respectively.

and last actively selected data. For a fair comparison and to reduce the performance fluctuation by training randomness, we train and evaluate each model three times with different initial data and report the mean and standard deviation. Note that the oracle performance trained supervised on full training split without GT-AUG yields 40.65 ($\pm$1.68) mAP. The uncertainty column shows which uncertainty measurement is used in active data selection in section III-C, and the consistency column refers to the use of consistency loss on labeled data (D^L) and unlabeled data (D^U) explained in section III-B.

The first three rows are models trained without active learning that select data randomly, from which we can observe the effect of consistency loss. Adding consistency loss to the training improves the detection performance, even using only labeled data. Conversely, the following three rows are models trained without consistency loss that show each uncertainty measurement's effect. Comparing the result of class and 3D box uncertainty, we notice that using 3D box uncertainty shows better performance, especially at the initial selection when class distribution is most imbalanced. It is because when using the class uncertainty, objects of less frequent class tend to have much higher uncertainty than vehicles. Also, by observing the standard deviation, we find that class uncertainty has a much higher variance since it relies on the randomly selected initial training set. On the other hand, 3D box uncertainty selects objects that are hard to predict boxes, such as occluded vehicles. When combining class and 3D box uncertainties, performance at the initial stages slightly decreased due to the class imbalance but increased as the class distribution balances.

The last row displays the performance of our proposed semi-supervised active learning approach. Our method outperformed the baseline model by 17.12%, the semi-supervised model by 6.19%, and the active learning model by 9.53%. These results demonstrate that active learning and semi-supervised learning mutually benefit each other.

D. Qualitative Results

We visualize the detection results of the proposed method and baseline (random sampling without consistency loss) at the last cycle to show how semi-supervised active learning can improve the 3D detector. Fig. 5a shows that our method improves the detection performance of distant and highly occluded objects. These highly uncertain samples are more likely to be sampled in the following cycle and improved by active learning. Moreover, object confidence and localization become more accurate with the help of semi-supervision, as shown in Fig. 5b. Interestingly, the model trained using our approach detected objects with no annotation due to high truncation and low point density correctly in Fig. 5c.

V. CONCLUSION

In this research, we presented a joint semi-supervised active learning method for LiDAR-based 3D object detection that can leverage both labeled and unlabeled data and take 3D box uncertainty into account. We utilize semi-supervision in the training stage to take full advantage of available data points regardless of annotation, which allows the trained model to measure more informative uncertainty in the active data selection stage. Moreover, we also proposed measuring uncertainty in the 3D bounding box predictions to make the model can select samples challenging to detect, such as objects with a high measurement uncertainty of occlusion or point sparsity. We demonstrated our approach using the KITTI dataset to show that using 3D consistency for both training and data selection, the trained model can detect objects with high occlusion or low point density. In future work, we plan to evaluate our approach using other datasets such as nuScenes [43] and Waymo Open [44] and explore different semi-supervision strategies such as self-ensemble [27]–[29].

REFERENCES

[1] Beluch, W., Genewein, T., Nürnberger, A. & Köhler, J. The power of ensembles for active learning in image classification. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 9368-9377 (2018)

[2] Houlsby, N., Huszár, F., Ghahramani, Z. & Lengyel, M. Bayesian active learning for classification and preference learning. *ArXiv Preprint ArXiv:1112.5745*. (2011)

[3] Gal, Y., Islam, R. & Ghahramani, Z. Deep bayesian active learning with image data. *International Conference On Machine Learning*. pp. 1183-1192 (2017)

[4] Kirsch, A., Van Amersfoort, J. & Gal, Y. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances In Neural Information Processing Systems*. **32** (2019)

[5] Zhu, J. & Bento, J. Generative adversarial active learning. *ArXiv Preprint ArXiv:1702.07956*. (2017)

[6] Tran, T., Do, T., Reid, I. & Carneiro, G. Bayesian generative active deep learning. *International Conference On Machine Learning*. pp. 6295-6304 (2019)

[7] Sinha, S., Ebrahimi, S. & Darrell, T. Variational adversarial active learning. *Proceedings Of The IEEE/CVF International Conference On Computer Vision*. pp. 5972-5981 (2019)

[8] Sener, O. & Savarese, S. Active learning for convolutional neural networks: A core-set approach. *ArXiv Preprint ArXiv:1708.00489*. (2017)

[9] Agarwal, S., Arora, H., Anand, S. & Arora, C. Contextual diversity for active learning. *European Conference On Computer Vision*. pp. 137-153 (2020)

[10] Brust, C., Käding, C. & Denzler, J. Active learning for deep object detection. *ArXiv Preprint ArXiv:1809.09875*. (2018)

[11] Aghdam, H., Gonzalez-Garcia, A., Weijer, J. & López, A. Active learning for deep detection neural networks. *Proceedings Of The IEEE/CVF International Conference On Computer Vision*. pp. 3672-3680 (2019)

[12] Yoo, D. & Kweon, I. Learning loss for active learning. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 93-102 (2019)

[13] Kao, C., Lee, T., Sen, P. & Liu, M. Localization-aware active learning for object detection. *Asian Conference On Computer Vision*. pp. 506-522 (2018)

[14] Roy, S., Unmesh, A. & Namboodiri, V. Deep active learning for object detection.. *BMVC*. pp. 91 (2018)

[15] Haussmann, E., Fenzi, M., Chitta, K., Ivanecky, J., Xu, H., Roy, D., Mittel, A., Koumchatzky, N., Farabet, C. & Alvarez, J. Scalable active learning for object detection. *2020 IEEE Intelligent Vehicles Symposium (iv)*. pp. 1430-1435 (2020)

[16] Choi, J., Elezi, I., Lee, H., Farabet, C. & Alvarez, J. Active learning for deep object detection via probabilistic modeling. *Proceedings Of The IEEE/CVF International Conference On Computer Vision*. pp. 10264-10273 (2021)

[17] Yuan, T., Wan, F., Fu, M., Liu, J., Xu, S., Ji, X. & Ye, Q. Multiple instance active learning for object detection. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 5330-5339 (2021)

[18] Gao, M., Zhang, Z., Yu, G., Arık, S., Davis, L. & Pfister, T. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. *European Conference On Computer Vision*. pp. 510-526 (2020)

[19] Zhu, X., Lafferty, J. & Ghahramani, Z. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. *ICML 2003 Workshop On The Continuum From Labeled To Unlabeled Data In Machine Learning And Data Mining*. **3** (2003)

[20] Gu, Y., Jin, Z. & Chiu, S. Combining active learning and semi-supervised learning using local and global consistency. *International Conference On Neural Information Processing*. pp. 215-222 (2014)

[21] Elezi, I., Yu, Z., Anandkumar, A., Leal-Taixe, L. & Alvarez, J. Not All Labels Are Equal: Rationalizing The Labeling Costs for Training Object Detection. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 14492-14501 (2022)

[22] Golestaneh, S. & Kitani, K. Importance of self-consistency in active learning for semantic segmentation. *ArXiv Preprint ArXiv:2008.01860*. (2020)

[23] Feng, D., Wei, X., Rosenbaum, L., Maki, A. & Dietmayer, K. Deep active learning for efficient training of a lidar 3d object detector. *2019 IEEE Intelligent Vehicles Symposium (IV)*. pp. 667-674 (2019)

[24] Schmidt, S., Rao, Q., Tatsch, J. & Knoll, A. Advanced active learning strategies for object detection. *2020 IEEE Intelligent Vehicles Symposium (IV)*. pp. 871-876 (2020)

[25] Segal, S., Kumar, N., Casas, S., Zeng, W., Ren, M., Wang, J. & Urtasun, R. Just label what you need: fine-grained active selection for perception and prediction through partially labeled scenes. *ArXiv Preprint ArXiv:2104.03956*. (2021)

[26] Jeong, J., Lee, S., Kim, J. & Kwak, N. Consistency-based semi-supervised learning for object detection. *Advances In Neural Information Processing Systems*. **32** (2019)

[27] Zhao, N., Chua, T. & Lee, G. Sess: Self-ensembling semi-supervised 3d object detection. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 11079-11087 (2020)

[28] Zheng, W., Tang, W., Jiang, L. & Fu, C. SE-SSD: Self-ensembling single-stage object detector from point cloud. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 14494-14503 (2021)

[29] Tarvainen, A. & Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances In Neural Information Processing Systems*. **30** (2017)

[30] Yin, T., Zhou, X. & Krahenbuhl, P. Center-based 3d object detection and tracking. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 11784-11793 (2021)

[31] Yan, Y., Mao, Y. & Li, B. Second: Sparsely embedded convolutional detection. *Sensors*. **18**, 3337 (2018)

[32] Chen, X., Kundu, K., Zhu, Y., Berneshawi, A., Ma, H., Fidler, S. & Urtasun, R. 3d object proposals for accurate object class detection. *Advances In Neural Information Processing Systems*. **28** (2015)

[33] Bishop, C. Mixture density networks. (Aston University,1994)

[34] Shannon, C. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing And Communications Review*. **5**, 3-55 (2001)

[35] Kingma, D. & Welling, M. Auto-encoding variational bayes. *ArXiv Preprint ArXiv:1312.6114*. (2013)

[36] Gal, Y. & Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *International Conference On Machine Learning*. pp. 1050-1059 (2016)

[37] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. Generative adversarial networks. *Communications Of The ACM*. **63**, 139-144 (2020)

[38] Odena, A., Olah, C. & Shlens, J. Conditional image synthesis with auxiliary classifier gans. *International Conference On Machine Learning*. pp. 2642-2651 (2017)

[39] Larsen, A., Sønderby, S., Larochelle, H. & Winther, O. Autoencoding beyond pixels using a learned similarity metric. *International Conference On Machine Learning*. pp. 1558-1566 (2016)

[40] Everingham, M., Van Gool, L., Williams, C., Winn, J. & Zisserman, A. The pascal visual object classes (voc) challenge. *International Journal Of Computer Vision*. **88**, 303-338 (2010)

[41] Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. Microsoft coco: Common objects in context. *European Conference On Computer Vision*. pp. 740-755 (2014)

[42] Geiger, A., Lenz, P. & Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. *2012 IEEE Conference On Computer Vision And Pattern Recognition*. pp. 3354-3361 (2012)

[43] Caesar, H., Bankiti, V., Lang, A., Vora, S., Liong, V., Xu, Q., Krishnan, A., Pan, Y., Baldan, G. & Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 11621-11631 (2020)

[44] Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B. & Others Scalability in perception for autonomous driving: Waymo open dataset. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 2446-2454 (2020)