2020 IEEE Intelligent Vehicles Symposium (IV)
October 20-23, 2020. Las Vegas, USA

Advanced Active Learning Strategies for Object Detection

Sebastian Schmidt[1,2], Qing Rao[2,*], Julian Tatsch[1,*] and Alois Knoll[1]

***Abstract*— Future self-driving cars must be able to perceive and understand their surroundings. Deep learning based approaches promise to solve the perception problem but require a large amount of manually labeled training data. Active learning is a training procedure in which the model itself selects interesting samples for labeling based on their *uncertainty*, with substantially less data required for training. Recent research in active learning has mostly focused on the simple image classification task. In this paper, we propose novel methods to estimate sample uncertainties for 2D and 3D object detection using *Ensembles*. We moreover evaluate different training strategies including *Continuous Training* to alleviate increasing training times introduced by the active learning cycle. Finally, we investigate the effects of active learning on imbalanced datasets and possible interactions with class weighting. Experiment results show both increased time saving around 55% and data saving rates of around 30%. For the 3D object detection task, we show that our proposed uncertainty estimation method is valid, saving 35% of labeling efforts and thus is ready for application for automotive object detection use cases.**

I. Introduction

To drive safely on highways or in urban traffic, future self-driving cars must be able to perceive and understand their environment. Object detection is a typical technology for environmental perception and has long been a challenging research topic in the field of computer vision and machine learning. In the past few years, the automotive industry and many research institutes developed deep-learning-based approaches to object detection which already produced convincing results. Among them, Faster R-CNN [1], Yolo [2], SSD [3] (2D object detection) and AVOD [4] (3D object detection) are widely used network architectures in self-driving car prototypes.

In general, the automotive industry faces three major challenges when training neural networks for self-driving cars. First, a massive amount of manually labeled data is required for training. Acquiring manual labels is only the final step of a complex data-pipeline comprising data collection, data ingestion and data cleansing. Even if manual labeling efforts are outsourced to labeling specialists, cost overruns and project delays still frequently occur. Second, the training procedure itself is time-consuming and requires considerable computational resources. This hinders fast turnaround times in the development cycle if developers have to wait for computational resources to train and evaluate their results. Last but not least, the training samples collected from public traffic scenes are often heavily class imbalanced. For example, in the KITTI dataset [5], nearly 80% of the labeled objects are *Cars*. For self-driving cars, however, the so-called corner- or edge cases are the training samples that are of real interest.

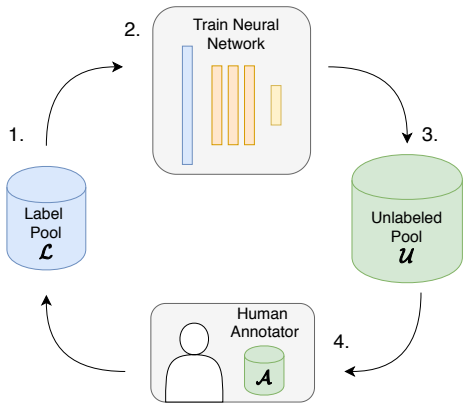


Fig. 1: The Active Learning Cycle

A recently emerging research field named *Active Learning* aims to resolve the aforementioned issues during the development of deep neural networks. The term active learning was originally used in the pedagogic context in which teaching strives to involve students in the learning process more directly. In the context of machine learning, active learning describes a training procedure where training samples are selected by the trained neural networks themselves. The underlying idea is that, if a learning algorithm can actively choose the data it wants to learn from, it can perform better than traditional passive learning methods with substantially less data for training. As illustrated in Fig. 1, active learning is an iterative training process that consists of the following three steps.

- A model is trained on labeled dataset $\mathcal{L}$.
- The resulting trained model selects new training samples $\mathcal{A}$ from an unlabeled data pool $\mathcal{U}$, typically based on their *uncertainty* of prediction.
- A human annotator labels the selected samples $\mathcal{A}$ and adds them to the labeled dataset $\mathcal{L}$.

Most existing methods to estimate the uncertainty of training samples are based on probability sampling e.g. Monte Carlo (MC) Dropout [6] or Ensembles [7], [8]. Subsequently, these methods are combined with an acquisition function that selects samples out of the unlabeled data pool. Entropy, mutual information and variation ratio (VR) are widely used

* Contributed equally to this work
[1]Chair for Robotics, Artificial Intelligence and Embedded Systems, Technical University of Munich sebastian95.schmidt@tum.de, Julian.tatsch@tum.de, knoll@in.tum.de
[2]BMW Autonomous Driving Campus, Alfred-Nobel-Str. 3, 85716 Unterschleissheim, Germany sebastian.sg.schmidt@bmw.de, Qing.Rao@bmw.de

metrics to quantify the uncertainty of a selected sample [7]–[12].

Current research in the field of active learning [7], [8], [11] is mainly focused on the simple task of image classification. For example, uncertainty estimation using ensembles has only been studied for this single use case so far [7]. There still exists a considerable research gap when it comes to extending active learning to the more complex 2D and 3D object detection tasks. Filling this research gap is the aim of our paper.

In this paper, we evaluate advanced active learning strategies for 2D and 3D object detection tasks. The key contributions of this paper are three-fold. First, we propose different methods to estimate uncertainty using ensembles for 2D object detection, which has not been extensively studied yet. Second, we examine the accuracy and data saving impact of different active learning strategies, which alleviates the computational burden incurred by retraining in the active learning loop. Last but not least, to resolve the issue of class imbalanced datasets, we examine if active learning is implicitly able to perform class weighting to some extent, or if class weighting combined with active learning shows additional improvement. In a nutshell, we show that active learning is applicable for the automotive machine learning use case.

The rest of this paper is structured as follows. In section II, we give an overview of the related work. In section III, we propose four methods including *Consensus Score*, *Consensus Score Variation Ratio*, *Region of Interest (RoI) Matching*, and *Sequential RoI Matching* to estimate uncertainty for 2D object detection. In section IV, we evaluate two active learning strategies for 2D object detection namely *Continuous Training* (Section IV-A) and *Active Class Weighting* (Section IV-B), and we extend active learning to a more complex 3D object detection architecture and evaluate it on the KITTI dataset (Section IV-C). Section VI sums up our findings and points out some possible directions for future work.

II. Related Work

Active learning is an area of current research, yet there are surprisingly few authors concerned with the vital tasks of 2D and 3D object detection. Brust et al. [13] investigated active learning for YOLO object detectors and presented an uncertainty estimation method based on the confidence score of the neural network. Their approach calculates an uncertainty score with the aggregation functions sum, mean or maximum of the margin between the highest and second highest class prediction confidence. Their active learning approach shows only slightly better results than a randomly trained baseline, for 50 to 250 sample cycles on the PASCAL VOC dataset. Kao et al. [14] proposed uncertainty measures based on the localization tightness and stability. The localization tightness measures the difference between the first stage of the network as Region of Interest (RoI) and the second stage as the final bounding box, measured by the Intersection over Union (IoU). The localization stability measures the IoU between several instances of the same sample affected by different injected Gaussian noises. While the performance of their approaches is promising, it lacks in model architecture independence and time efficiency.

Feng et al. extended their previous publication on uncertainty estimation [15] by active learning [12]. They ran a two-staged object detector based on Faster R-CNN on a LIDAR bird's eye view. They assumed a "perfect feature extractor" which was simulated with ground truth data as region proposals. Besides, they trained an object detector as "real feature extractor", based on a pre-trained model. With the feature extractors, they created a dataset for a classification and localization task containing 30000 cropped features (region proposals) each. Based on these cropped features, they applied an ensemble and a MC dropout at the second stage. By using the "perfect feature extractor", they achieved a maximum saving of 61% in the classification task. For the real proposals, their active learning method still outperformed the random sample selection but no data saving is published by the authors.

To reduce the sampling effort Huang et al. [16] used temporally sequential images as an approximation of multiple forward passes. They calculated the uncertainty of semantic segmentation for a video with MC dropout using their region-based temporal aggregation method. The authors used a single forward pass per frame and calculated the uncertainty over the time horizon. To account for the movement of the pixels, they used optical flow. They showed that their approach was ten times faster than five MC forward passes.

III. Uncertainty Estimation Using Ensembles

For 2D object detection, we prefer ensembles over MC dropout for uncertainty estimation for the following reasons. First, existing works [7], [8] already show that ensembles outperform MC dropout for the image classification task in terms of data saving rate. Second, unlike for MC dropout, the network architecture does not have to be modified when using ensembles. Compared to the already better studied active learning classification task, the active learning object detection task requires to include bounding boxes i.e. RoIs into its estimation which sample to select next for labeling. As the ensembles sub-models RoI predictions may be incongruent, they must be matched for each object before calculating the uncertainty score per object. However, for 2D object detection task, different objects detected in the same image might carry different uncertainties. To calculate an uncertainty score for the entire image, we need to aggregate uncertainties of individual objects e.g. using max, mean or sum functions as suggested in [13].

In the remainder of this section, we propose four different methods to estimate uncertainty using a Faster R-CNN ensemble network.

Consensus Score We introduce the consensus score $\gamma \in [0, 1]$ as an uncertainty metric for 2D object detection ensemble networks. This score is based on localization uncertainty. Similar to [14], we use IoU as a function to map and weight different RoIs. Let N be the number of output RoIs and M

the number of sub-models in the ensemble. We first calculate an IoU matrix $\Omega^{ij} \in \mathbb{R}^{N \times N}, i, j \in [1, M]$ by comparing each RoI from the i-th single forward pass in the Faster R-CNN ensemble against the j-th. We then identify RoI matches by finding the maximum IoU value in each row of Ω^{ij}, noted as $\max(\omega_n^{ij}), n \in [1, N]$. By subtracting it from one, the score is proportional to the uncertainty. Our proposed consensus score is calculated by finding the minimum IoU value for each RoI-match among all comparisons and taking the average of them, as expressed in Eq. 1.

$$\gamma = 1 - \frac{1}{N} \sum_{n=1}^{N} \min_{i,j \in [1,M]} \left\{ \max(\omega_n^{ij}) \right\} \tag{1}$$

Fig. 2a shows an example of the calculation of the consensus score in a ensemble with three sub-models. Each sub-model's output is visualized in different colors. The consensus score in this example is rather small ($1 - 0.85 = 0.15$), which indicates a low uncertainty of the input image.

(a) Consensus Score (b) Consensus Score VR

Fig. 2: Methods Based on Consensus Score for Ensembles

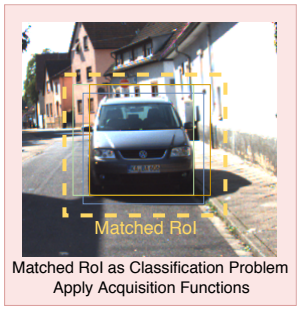
Consensus Score VR As combining localization and classification uncertainty improves the uncertainty estimation [14], we add a variation ratio (VR) term into our proposed consensus score to incorporate class information. The extended consensus score is expressed in Eq. 2, with σ being the distribution of detected classes and $\mathcal{VR}$ a variation ratio function.

$$\gamma_{vr} = 1 - \frac{1}{N} \sum_{n=1}^{N} \min_{i,j \in [1,M]} \left\{ \max(\omega_n^{ij}) \right\} \cdot [1 - \mathcal{VR}(\sigma_n)] \tag{2}$$

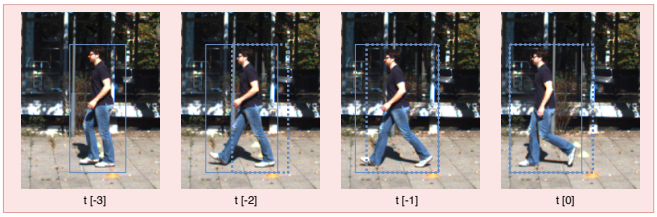
As shown in Fig. 2b, by introducing the variation ratio term, the uncertainty of the input image increased to $1 - 0.561 = 0.439$. This corresponds to the fact that two sub-models in the ensemble detect a *Car* whereas the other one detects a *Person*.

RoI Matching As Feng et al. [12] achieved better results for image classification by only incorporating class uncertainty, we also propose a method called RoI matching for 2D object detection without considering the localization uncertainty. The uncertainty score is calculated only based on the distribution of detected classes of each RoI matched by our IoU matrix Ω without using the IoU values anymore.

Fig. 3a shows an example of the RoI matching method.



(a) RoI Matching



(b) Sequence RoI Matching Example

Fig. 3: Methods Based on RoI Matching for Ensembles

Sequence RoI Matching The matching approach can also be extended to a temporal version, i.e., the model is applied to a sequence of temporally consecutive images. Over time the objects in the scene are moving and therefore their location in the images is changing. The RoIs describing these objects will differ and create different samples for uncertainty estimation, as shown in Fig. 3b. In the figure, the solid box is the prediction of the current time step whereas the dotted box is the prediction of the previous time step. Sampling multiple forward passes per image can be approximated by sampling along consecutive time steps. As objects are moving, continuous matching is necessary. We assume that the sampling rate is high enough, for the bounding box of an object to still intersect with its bounding box of the last time step. By matching the corresponding regions containing the same object, the network's uncertainty for this object can be estimated over a time horizon. As this approach only needs one sampling step per time step it allows for computationally affordable live uncertainty estimation. This makes it possible to select data during driving.

IV. Active Learning Strategies

After introducing uncertainty estimation for the object detection task using ensembles, we set up an active learning cycle as shown in Fig. 4 to compare different learning strategies. The cycle consists of a training phase, a sample selection phase and an evaluation phase. In the selection phase, the n most uncertain samples are chosen based on their uncertainty score as $\mathcal{A}$. As the step size of labels added $\mathcal{A}$ has been shown to have only a small influence on the uncertainty estimation [17], it will be fixed during our experiments. We use a Faster R-CNN model with a

pre-trained ResNet50 Feature Pyramid Network (FPN) [18] as the backbone. Similar to [12], we replace the model's classification head by three fully connected layers, each with a subsequent dropout layer. This enables us to estimate uncertainties using MC dropout. MC dropout is used in the experiments for some strategies.

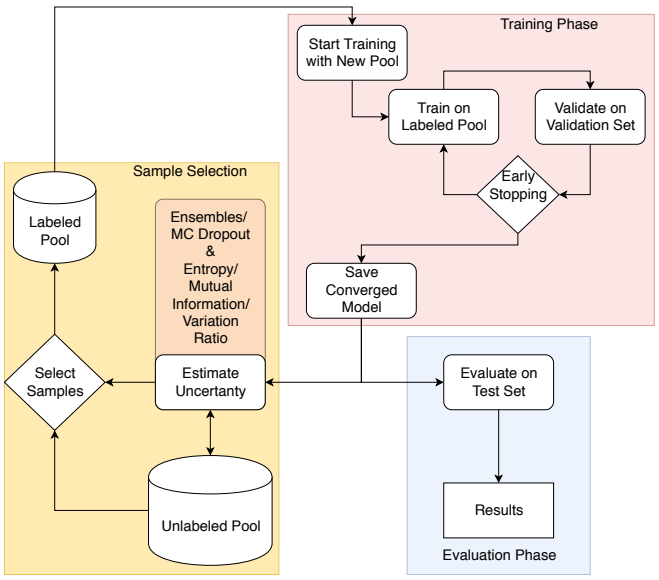
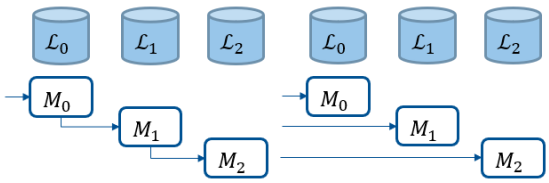


Fig. 4: Active Learning Cycle Setup

In the remainder of this section, we examine two different active learning strategies for 2D object detection and experiment with active learning for 3D object detection.

A. Continuous Training Strategy

Training from scratch in the active learning cycle is usually a time-consuming procedure. To tackle this problem, we evaluate a continuous training strategy [9] as shown in Fig. 5a, where a model $\mathcal{M}_{i-1}$ from the previous iteration is reused and fine-tuned with an extended label pool $\mathcal{L}_i$ in the current iteration. We compare this strategy against the "training from scratch" strategy where a new model $\mathcal{M}_i$ is trained in every iteration using the current labeled pool $\mathcal{L}_i$ (Fig. 5b).



(a) Continuous Training (b) Training From Scratch

Fig. 5: Continuous Training vs. Training from Scratch

B. Active Class Weighting

The datasets for autonomous driving are usually class imbalanced since in most traffic scenes, there are more cars than pedestrians or motorcycle riders. A common method to work with imbalanced datasets is to use a class weighted loss function. Active learning promises to have a similar effect, as the samples are selected based on their uncertainty and rarely occurring classes show a higher uncertainty. Therefore, we compare the method of using a class weighted loss function with active learning and we investigate the effects by combining them together, yielding "active class weighting". The weights of active class weighting are defined by the ratio of the number of total labels to that of a specific class. These weights are used for a class weighted cross-entropy loss function.

C. 3D Object Detection

3D object detection is perhaps the most important task for autonomous driving. To our best knowledge, active learning has not been applied to the automotive 3D object detection use case yet. Therefore we extend the continuous training strategy and active class weighing to 3D object detection in a two-staged AVOD model [4]. As the process of training AVOD models is quite complex, we perform active learning with a fixed number of iterations as suggested in [4] instead of training until convergence. Nevertheless, we compare our known sub-optimally trained active learning model against the fully trained AVOD model (Fig. 9).

V. Experiments & Results

We evaluate the proposed uncertainty estimation methods (Section III) and the active learning strategies (Section IV) on the KITTI dataset [5]. For 3D object detection we evaluate the classes *Person*, *Car* and *Pedestrian* which allows a direct comparison with the original AVOD model. For the 2D object detection experiments, we merge the classes *Pedestrians*, *Cyclists*, and *Sitting Persons* to the class *Person* and we use it together with the classes *Car*, *Van* and *Truck* as the target classes. This yields two classes with high occurrence and two with low occurrence. The dataset is split into a training set with labeled and unlabeled pool, a test set and a validation set used for early stopping. The labeled pool has an initial size of 2000 and will be increased by 500 ($\mathcal{A}$) at each iteration. As hyper-parameter optimization was not in the scope of our project, we use the parameters suggested by Ren et al. [1] and Ku et al. [4]. For the MC dropout models used in our experiments we used ten forward passes. To evaluate accuracy improvements and data savings, we measure the mean average precision (mAP) over the percentage of used data. We compare all our experiments with the random sample selection and a baseline model trained on the entire training set (Fully Trained).

A. Continuous Training Strategy

We evaluate the continuous training strategy for the 2D object detection task using MC dropout. Here, we compare it against the training from scratch strategy.

The results in Fig. 6 show that the continuous training method outperforms training from scratch. In addition to a faster convergence and an increased data saving rate, we also

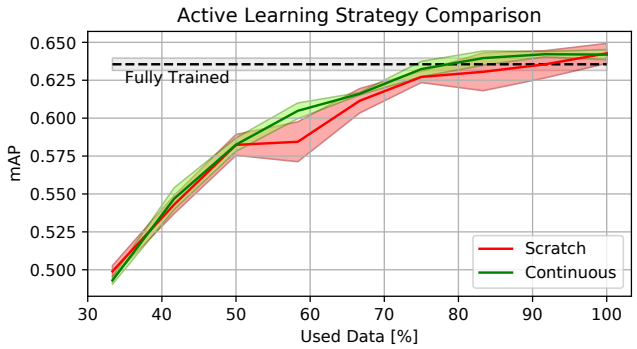


Fig. 6: Continuous Active Learning vs. Active Learning from Scratch

save a significant amount of time, which alleviates a part of the time lost due to the iterative active learning process. All time and data savings are presented in Tab. I. The continuous training strategy is about 2.35x less time-consuming than training from scratch. Based on this finding, we use this continuous training strategy exclusively for the following experiments.

Method	Training Time	Data Saving
Baseline	16 h	-
Sequence (3 Frames) Continuous Training	24 h	20 %
MC Dropout Continuous Training	34 h	25 %
MC Dropout Trained from Scratch	79,5 h	10 %
Ensembles Without parallelization	96 h	30 %

TABLE I: Training Times of Different Active Learning Strategies on the KITTI Dataset with NVIDIA Tesla V100

B. Uncertainty Estimation Using Ensembles

We compare our proposed uncertainty estimation methods using ensembles against a random sample selection and training pipeline without using active learning. We use an ensemble with three sub-models and three consecutive frames for the sequence RoI matching method. [7] showed that using more than three sub-models does not further increase the performance.

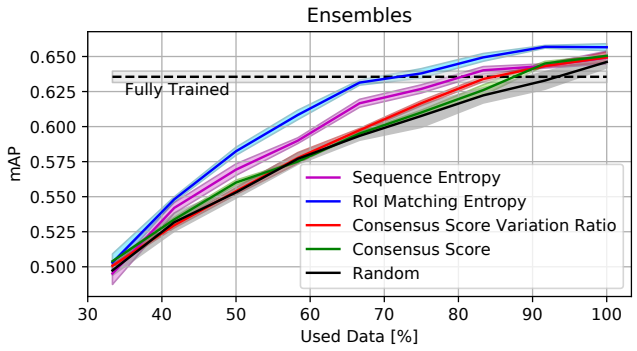


Fig. 7: Ensembles Methods

Our experiment results in Fig. 7 show that the concept of a consensus score does not outperform a random sample selection and is thus not suitable for active learning. This could be caused by the localization uncertainty, which is calculated for all RoIs, even for the background classes. Nevertheless, our proposed RoI matching and sequence RoI matching methods outperform the random sample selection. By comparing the training time and the performance with our previous experiment, we observe a trade-off between time and data saving rate. Table I shows that the faster method, sequence RoI matching, has the lowest data saving, while the method with the highest data saving, RoI Matching Ensembles, is the most time-consuming. MC dropout lies in between both with a medium time and data saving rate. Ensembles show the best data saving rate and have a high parallelization potential which reduces the time-consumption. Nevertheless, ensembles are computationally more expensive than MC dropout.

C. Active Class Weighting

We evaluate the proposed active class weighting strategy with an imbalanced dataset. For this experiment, we estimate the uncertainty using MC dropout which makes the experimental setup easier. As shown in Fig. 8, weighted classes in the loss function improve the performance of both the baseline and our active learning approach. The imbalance of the dataset seems to bias the selection based on uncertainty, which is improved through our active class weighting. Therefore, we believe that active learning should be combined with class weighting when working with imbalanced datasets.

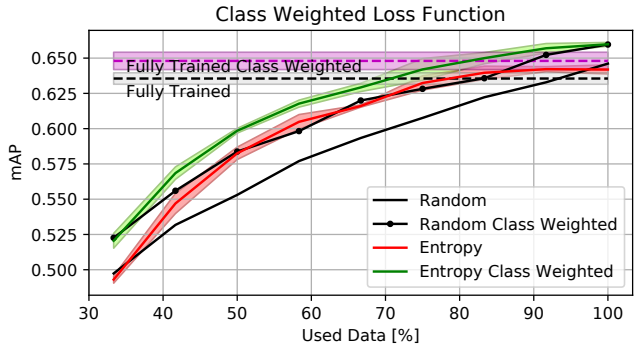


Fig. 8: Active Class Weighing

D. 3D Object Detection

For 3D object detection, we propose an active learning loop using the AVOD object detection model as a proof of concept. We show that our uncertainty estimation approach can be applied to more complex 3D object detection models. For this proof of concept, we use MC dropout for uncertainty estimation for simplicity.

Our experiment shows a data saving rate of 35% compared to the baseline approach. However, due to hardware limitations, we were not able to train each model until convergence. Therefore, we can expect that our demonstrated data saving gains to be the theoretically possible lower bound. We hope to show increased data saving rates in future work.

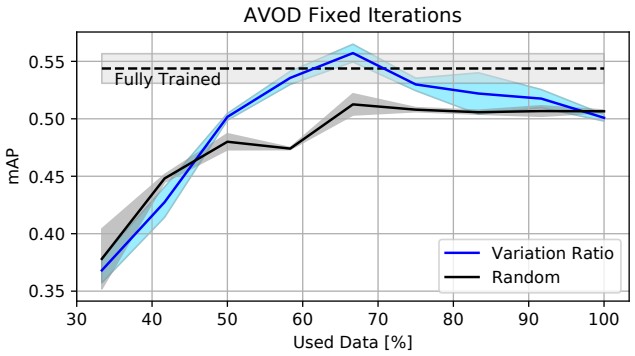


Fig. 9: 3D Object Detection - AVOD

It is worth mentioning that the AVOD training process was originally not intended to be applied for active learning. Adapting it into an active learning cycle was quite challenging.

VI. Concluding Remarks

In this paper, we examined active learning on an industrial use case: object detection for autonomous driving. We proposed different methods to estimate uncertainty using ensembles, and we evaluated two strategies for training 2D object detection networks, including continuous training and active class weighting. Using the continuous training strategy, we showed about 55% training time saving and an increased data saving rate by 15% compared with "training from the scratch". Furthermore, we showed that active learning can be combined with dataset balancing methods to further improve the data saving rate. In addition to 2D object detection, we implemented an active learning proof of concept using a more complex neural network for 3D object detection, which enables more data-efficient development of object detectors in the automotive industry.

Our future work will focus on online uncertainty estimation, i.e., estimating the uncertainty in real-time while driving. This will enable more efficient data collection as uninteresting data can be skipped during the data collection session. Moreover, we will look further into continuous learning and examine its combination with active learning. Besides, we will parallelize the training of the different sub-models.

References

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 39, no. 6, pp. 1137–1149, 2017.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-time Object Detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 21–37.

[4] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. Waslander, "Joint 3D Proposal Generation and Object Detection from View Aggregation," in *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1–8.

[5] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets Robotics: The KITTI Dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1–6, 2013.

[6] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in *Proceedings of International Conference on Machine Learning (ICML)*, vol. 48, 2016, pp. 1050–1059.

[7] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler, "The Power of Ensembles for Active Learning in Image Classification," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 9368–9377.

[8] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," in *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, 2017, pp. 6405–6416.

[9] F. Dayoub, N. Sunderhauf, and P. I. Corke, "Episode-Based Active Learning with Bayesian Neural Networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*. IEEE Computer Society, 2017, pp. 498–500.

[10] Y. Gal, "Uncertainty in Deep Learning," Ph.D. dissertation, University of Cambridge, 2016.

[11] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian Active Learning with Image Data," in *Proceedings of the International Conference on Machine Learning (ICML)*, vol. 70, 2017, pp. 1183–1192.

[12] D. Feng, X. Wei, L. Rosenbaum, A. Maki, and K. Dietmayer, "Deep Active Learning for Efficient Training of a LiDAR 3D Object Detector," in *Proceedings of IEEE Intelligent Vehicles Symposium (IV)*, 2019, pp. 667–674.

[13] C.-A. Brust, C. Käding, and J. Denzler, "Active Learning for Deep Object Detection," in *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, (VISIGRAPP)*, 2019, pp. 181–190.

[14] C.-C. Kao, T.-Y. Lee, P. Sen, and M.-Y. Liu, "Localization-Aware Active Learning for Object Detection," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2018, pp. 506–522.

[15] D. Feng, L. Rosenbaum, and K. Dietmayer, "Towards Safe Autonomous Driving: Capture Uncertainty in the Deep Neural Network for Lidar 3D Vehicle Detection," in *Proceedings of the IEEE Conference on Intelligent Transportation Systems, Proceedings (ITSC)*, 2018, pp. 3266–3273.

[16] P. Y. Huang, W. T. Hsu, C. Y. Chiu, T. F. Wu, and M. Sun, "Efficient Uncertainty Estimation for Semantic Segmentation in Videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 1, 2018, pp. 536–552.

[17] K. Chitta, J. M. Alvarez, and A. Lesnikowski, "Large-Scale Visual Active Learning with Deep Probabilistic Ensembles," in *ArXiv e-prints*, 2018.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.