Deep Active Learning for Autonomous Perception

Navjot Singh, Håkon Hukkelås, and Frank Lindseth

Abstract. Traditional supervised learning requires significant amounts of labeled training data to achieve satisfactory results. As autonomous perception systems collect continuous data, the labeling process becomes expensive and time-consuming. Active learning is a specialized semi-supervised learning strategy that allows a machine learning model to achieve high performance using less training data, thereby minimizing the cost of manual annotation. We explore active learning for autonomous vehicles, and propose a novel deep active learning framework for object detection and instance segmentation. We review prominent active learning approaches, study their performances in the aforementioned computer vision tasks, and perform several experiments using state-of-the-art R-CNN-based models for datasets in the self-driving domain. Our empirical experiments on a number of datasets reflect that active learning reduces the amount of training data required. We observe that early exploration with instance-rich training sets leads to good performance, and that false positives can have a negative impact if not dealt with appropriately. Furthermore, we perform a qualitative evaluation using autonomous driving data collected from Trondheim, illustrating that active learning can help in selecting more informative images to annotate.

Keywords: Active Learning · Object Detection · Instance Segmentation · Autonomous Perception

1 Introduction

We have come a long way with the development of autonomous driving in the past years. Today, several companies are working with state-of-the-art technologies for object detection, and image segmentation, to reach the common goal of developing the first fully Autonomous Vehicle (AV).

Currently, convolutional neural networks are the prominent machine learning models for these tasks; however, they are known to require a large amount of labeled data to achieve satisfactory results. Acquiring labeled data for computer vision tasks is extremely expensive and time-consuming as it is often done manually by an expert human annotator.

Active Learning (AL) is a semi-supervised strategy that aims to minimize the annotation effort and maximize the usage of "highly informative" learning data (Settles, 2009). "Informative" data contains new information that is beneficial for the learner's understanding of the current environment. AL gives the learner a chance to choose its

This paper was presented at the NIK-2020 conference; see http://www.nik.no/.

own training data, where the goal is to improve performance using less data. This strategy has received much attention in recent years and fits well in scenarios where data is easy to obtain but expensive to label (Settles, 2009).

The goal of this paper is to identify if AL can be used to achieve at least the same model performance using a smaller, carefully assembled dataset compared to using all of the data available. Furthermore, we explore the potential of current AL methods in the setting of autonomous driving.

To evaluate AL, we implement a pool-based AL Framework to train a "learner" (a convolutional neural network) to perform object detection and instance segmentation. We assess various AL query strategies based on their performance and compare to a random-selection baseline.

2 Related Work

Roy et al., 2018 implements different AL strategies in both a black-box and white-box setting using the object detection network Single Shot Multibox Detector (SSD) (Liu et al., 2016). They present an effective explore vs. exploit framework to achieve the best of both techniques by selecting samples with high and low informativeness; the explore vs. exploit trade-off is a known dilemma in AL (Bondu et al., 2010). Our primary takeaways from their approach are the two black-box query strategies, maximum entropy and sum entropy. In addition, we use a similar explore vs exploit technique.

Brust et al., 2018 combines AL on object detection with incremental learning that enables continuous exploration. In this scenario, new data and classes are added to the dataset over time. Furthermore, they propose various aggregation metrics (e.g., sum, maximum and average), query strategies for object detection models, and present an approach to better handle class imbalances during sample selection. Our experiments use the aggregation techniques presented in their approach to compare if we notice any similar performance patterns.

Morrison et al., 2019 address the task of instance segmentation, by considering both spatial and semantic uncertainty of a prediction using dropout sampling. They do not perform AL; however, their proposed approach presents a way to measure the prediction uncertainty of a model. By adding dropout layers to the fully-connected layers of Mask R-CNN (He et al., 2017), they run inference over the same image multiple times to measure the segmentation uncertainty. This approach can be seen as a committee of models where their disagreement is measured, and is used in one of our experiments.

Sörsäter, 2018 proposes a Query by Committee (QbC) AL approach using Monte Carlo dropout to perform semantic segmentation on roads. Monte Carlo dropout is used to measure a model's uncertainty similar to Gal and Ghahramani, 2015. Query strategies such as least confident, margin sampling, entropy, and Monte Carlo dropout are being used and compared with a baseline random learner. Their Monte Carlo dropout approach runs inference over every sample N times, and collect the predictions; however, the uncertainty is calculated differently compared to Morrison et al., 2019. The work presented in this paper is highly relevant in a autonomous self-driving setting.

3 Method

In this section, we describe various AL strategies to measure the informativeness of samples [1] for object detection and instance segmentation. We make no assumptions

[1] In this context, a sample is a single RGB image taken of a traffic scene.

on the underlying model and focus on using black-box query strategies. To measure a sample's informativeness, the method quantifies the uncertainty considering exclusively the model's inputs and outputs. Thereby making our method highly generalizable for other architectures of choice and easily usable for further development. We implement each method in a modular pool-based AL Framework [2] to be used with Detectron2 (Wu et al., 2019).

Our framework uses a *query scenario* [3] (i.e., sampling technique) called pool-based sampling. In pool-based sampling, the informativeness of all samples in an unlabeled set is measured, and a small set of highly informative samples are selected to be labeled by an annotator.

As a baseline, we use a random-selection learner (RAND) that selects samples randomly from the unlabeled set; as done in past works (Roy et al., 2018, Brust et al., 2018, Sörsäter, 2018).

The informativeness of each sample is measured with a *query strategy*, and is given as a score. This score tells us how certain or uncertain a model is about a sample. If a model is uncertain about a sample, it considers the sample as being highly informative.

Since an image can contain multiple detections with different scores, aggregation techniques are required to score the entire image. We use the aggregation techniques presented by Brust et al., 2018 with entropy for measuring the informativeness. The following learners require a probability distribution over all classes K in the form of a softmax vector. The entropy for a single detection in an image is given in Equation 1, where K is the number of classes, and x is a detection.

$$E_i = -\sum_{k}^{K} P(y_k|x) \log P(y_k|x) \tag{1}$$

We aggregate the entropy for each detection to produce a final image score, using the following aggregation techniques (a higher image score indicates more informative samples):

Sum Entropy (SUMENT) Sum of entropy is set as the image score. Prefers images containing many informative detections. See E_{SUM} in Equation 2

Max Entropy (MAXENT) Highest entropy score is set as the image score. Prefers images containing a single highly informative detection, and is not affected by the number of detections. See E_{MAX} in Equation 2.

Average Entropy (AVGENT) Average of entropy scores is set as the image score. Can prefer images containing many informative samples as much as images containing a single informative sample. See E_{AVG} in Equation 2.

$$E_{SUM} = \sum_{i=1}^{n} E_i \qquad E_{MAX} = \max_i E_i \qquad E_{AVG} = (\sum_{i=1}^{n} E_i)/n \tag{2}$$

Dropout (DROPOUT) We use Monte-Carlo dropout to form a "virtual" QbC framework, similar to Morrison et al., 2019. Model uncertainty is measured by adding dropout layers in the final layers of the model, and run inference over the same image multiple times. This generates a set of different detections, where we can measure the uncertainty by considering the disagreement. Overlapping detections are

[2] The code is available in a public repo: github.com/RovelMan/active-learning-framework

[3] A query scenario specifies how the learner asks an annotator for labels on specific highly informative samples (Settles, 2009)

grouped and defined as observations. For each observation, we calculate the mean softmax, mean bbox and mean mask, and use these values to calculate the semantic uncertainty (u_sem), spatial bounding box uncertainty (u_spl_bbox), spatial mask uncertainty (u_spl_mask), and a number of appearances of each detection out of a number of inferences (u_n); following the formulas given by Morrison et al., 2019. Take notice, we add a fourth uncertainty (i.e., u_spl_bbox), since we believe that the disagreement in bounding box prediction should be accounted for to get an overall uncertainty. The informativeness of an image is calculated using Equation 3, where $|O|$ is the number of observations. This virtual QbC framework is computationally lighter compared to having several models to form the committee.

$$DROPOUT = \sum_{i=1}^{|O|}(u\_sem * u\_spl\_bbox * u\_spl\_mask * u\_n) \tag{3}$$

Approximate Sampling: Using a pool-based sampling technique, the learner has to search through the whole unlabeled set to find informative samples. As this process is repeated, it becomes computationally intensive if the unlabeled set is large. We use the faster sample selection method proposed by Ertekin et al., 2007, where the framework search through a smaller set of randomly selected samples from the unlabeled set (i.e., 5000 samples) to then find informative samples.

Spectrum vs No-Spectrum: In some experiments, we select samples from the entire informativeness spectrum. This method (i.e., Spectrum) is slightly different from what is proposed by Roy et al., 2018. We refer to "Spectrum" as both exploring and exploiting by selecting samples with high, medium, and low informativeness scores. "No-Spectrum" is referred to as only exploring by selecting top k samples having the highest informativeness.

4 Experiments

In this section, we first present experiments with object detection and instance segmentation on the Apollo Synthetic dataset (team, 2019). Furthermore, we extend our object detection experiments to the Waymo Open dataset (“Waymo Open Dataset: An autonomous driving dataset”, 2019). Finally, we present qualitative experiments on data collected from Trondheim.

Evaluation Details: Each active learner is compared to the random-selection baseline (RAND) to see if they can reach a similar performance using a smaller training set. The performance of the learners can vary based on the initial set. Therefore, we run EXP 1 three times, using different initial training sets [4]. All reported results are computed from the test set. The model performance is measured as the prediction accuracy on the test set, and the results for each learner are shown in a graph as the performance over training set size.

Active Learning Setup: Each experiment follows a general AL flow, which lasts for a number of Active Learning Iterations (ALIs). Initially, samples are selected randomly from an unlabeled set to create an initial training set and to train an initial model. The learners then use this initial model as a starting point. During each ALI, a learner selects a new set of informative samples from the unlabeled set, adds it to its training set, does a full training, and evaluates itself on a test set. A learner is not trained from scratch each ALI; it uses the weights from a previous ALI.

[4] EXP 2 and 3 are only run once due to computational limitations.

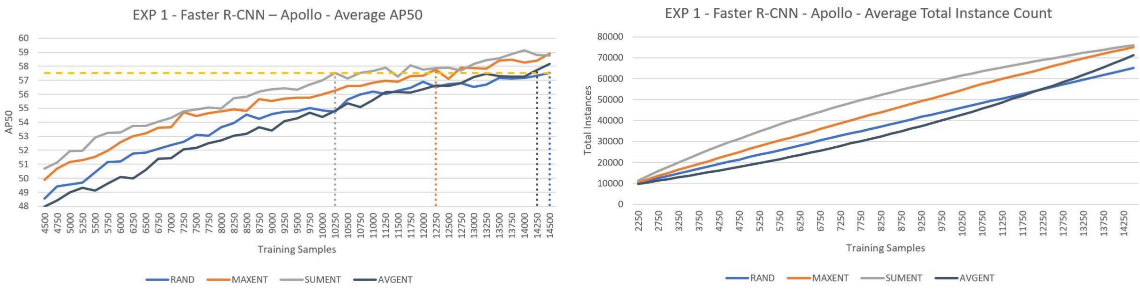


Figure 1: EXP 1 - **Left:** Performance results using AL with Object Detection on Apollo Synthetic. The performance is measured on a test set of 3000 samples, and averaged over three independent runs from ALI 10 to 50. **Right:** Shows the number of class instances added to the training set during each ALI for each learner. SUMENT favors images containing large number of instances.

EXP 1 - Object Detection - Apollo Synthetic

We use the photo-realistic Apollo Synthetic dataset (team, 2019), and the active learners SUMENT, MAXENT, and AVGENT for comparison.

Experimental Details: The dataset is pre-processed and contains 21244 samples that are randomly split into an initial training set of 2000, a test set of 3000, and an unlabeled set of 16244. We use all available thing classes in the dataset: Sedan, SUV, Hatchback, Van, PickupTruck, Truck, Bus, Cyclist, Motorcyclist, and Pedestrian.

We use our AL Framework with Detectron2's Faster R-CNN model for object detection. RAND selects n samples randomly, and the active learners use the faster sample selection method. They select 5000 samples randomly from the unlabeled set, run inference over them, give them an informativeness score, and return n samples from the entire informativeness spectrum using the method "Spectrum".

We train the model for 2500 iterations during the initial ALI using the initial training set. The trained model is then used as a starting point for the learners. In all the following ALIs, each learner is trained for 500 iterations. The AL process is repeated for 50 ALIs. During each ALI, the learner is evaluated, and 250 new images are added to the training set based on their informativeness score.

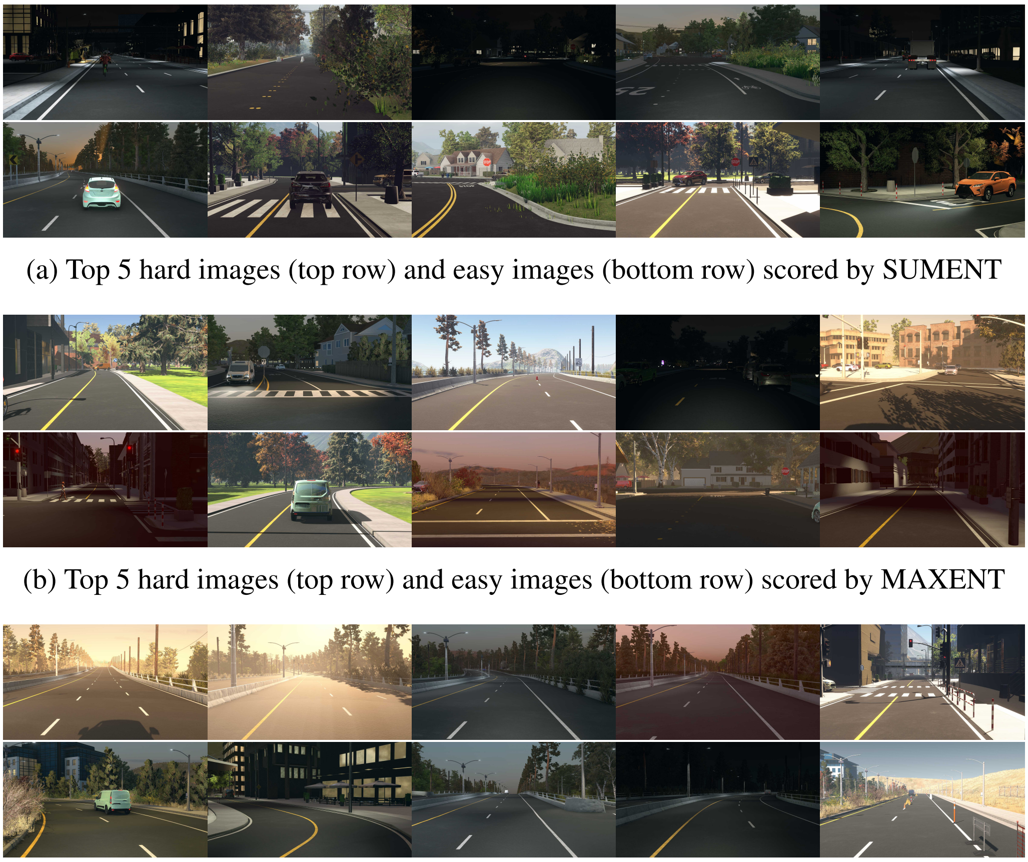
We sample from the entire spectrum, where we select the top 150 samples with high informativeness, bottom 25 with low informativeness, and 75 randomly with medium informativeness. This creates a set of 250 samples that is added to the training set by each active learner [5]. This experiment was run using early stopping as well (See "Early Stopping" in Appendix A)

Evaluation: Figure 1 (left) presents the AP50 [6] for each learner, and Figure 2 presents the hard/easy images chosen by each learner. Both active learners are able to perform better than RAND, using fewer samples. The performance results indicate that having an instance-rich training set can be beneficial for an active learner, and that the aggregation technique SUM fulfills this requirement.

We notice that AVGENT performs poorly. The Apollo Synthetic dataset can contain images having no classes or no objects of interest. As seen in Figure 2 (bottom 2 rows), AVGENT tends to select images containing few or no classes. In addition, images given

[5] This distribution (i.e., 150 high, 75 mid, 25 low) is favoring exploration, since we believe it might be beneficial for a learner to explore as early as possible. Nevertheless, we do not focus on tweaking this distribution as this is out of our scope.

[6] Average Precision score calculated at Intersection over Union (IoU) threshold of .50



(a) Top 5 hard images (top row) and easy images (bottom row) scored by SUMENT

(b) Top 5 hard images (top row) and easy images (bottom row) scored by MAXENT

(c) Top 5 hard images (top row) and easy images (bottom row) scored by AVGENT

Figure 2: EXP 1 - Images from Apollo Synthetic scored by the active learners from ALI 23. See "EXP 1" in Appendix A for more results.

high informativeness often contain a single false positive. Figure 1 (right) shows that AVGENT ends up with a lower number of total instances in the training set compared to other learners.

EXP 2 - Instance Segmentation - Apollo Synthetic

We implement a QbC Framework using dropout layers following the approach proposed by Morrison et al., 2019. We examine if this type of framework can improve on simple black-box uncertainty-based techniques. We use Detectron2's implementation of Mask R-CNN, the Apollo Synthetic dataset from EXP 1, and the active learners SUMENT, MAXENT, and DROPOUT for comparison.

Experimental Details: We use the two query strategy frameworks, QbC and Uncertainty Based, with the sample selection method No-Spectrum. For the QbC Framework, we use the learner DROPOUT. We only select 750 samples randomly from the unlabeled set, instead of 5000 in EXP 1, as DROPOUT is computationally heavy. The AL process is identical to EXP 1, except, models are trained for 3000 iterations in the initial ALI and 750 in the following ALIs. We run this experiment only once due to its complexity and the computationally heavy QbC Framework. For the DROPOUT learner, we follow the optimal thresholds [7] found by Morrison et al., 2019 and use them with these

[7] Thresholds are used to neglect highly informative detections that might contain false positives. The authors find these thresholds by performing a grid search on their uncertainty metrics. See Morrison et al.,

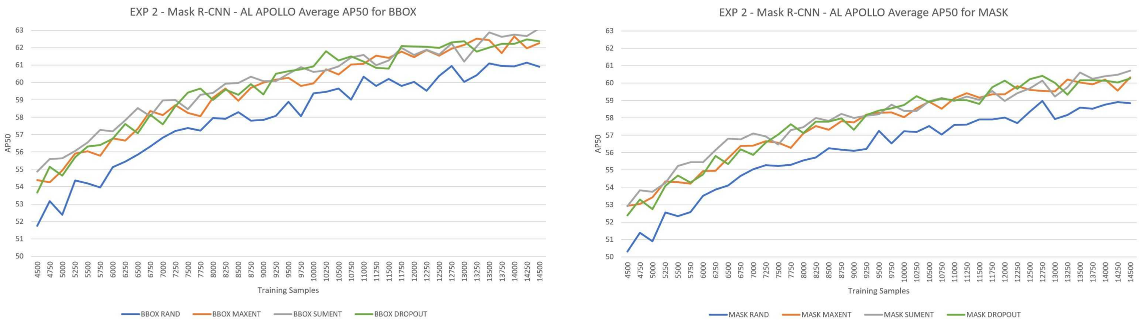


Figure 3: EXP 2 - Performance results using AL with Object Detection and Instance-based Segmentation on Apollo Synthetic. The values are from a single run from ALI 10 to 50. **Left:** AP50 is given for the bounding boxes. **Right** AP50 is given for the masks.

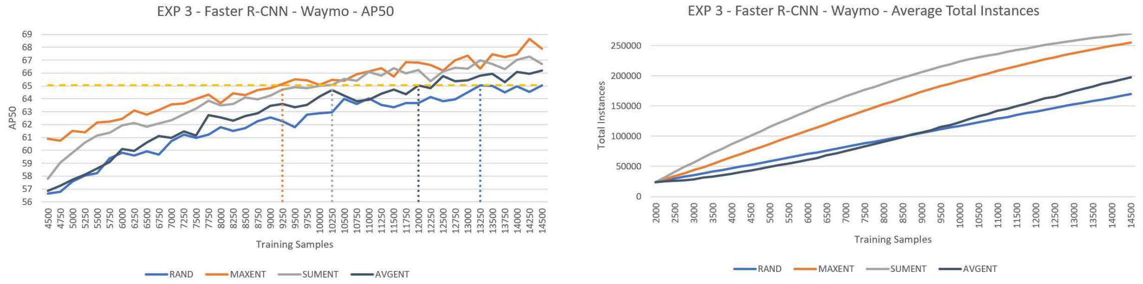


Figure 4: EXP 3 - **Left:** Performance results from a single run using AL with Object Detection on Waymo Open. The performance is measured on a test set of 3000 samples from ALI 10 to 50. **Right:** Shows the number of class instances added to the training set during each ALI for each learner. SUMENT favors images containing large number of instances.

uncertainty metrics.

Evaluation: The AP50 of bounding boxes and masks for each learner is shown in Figure 3. We see that all learners perform better than the baseline RAND learner, and SUMENT performs slightly better than MAXENT. However, the computationally heavy learner DROPOUT did not perform any better than the uncertainty-based learners. Even though, its performance can be further improved by tweaking and optimizing the thresholds on the uncertainty measures; this is outside the scope of this paper.

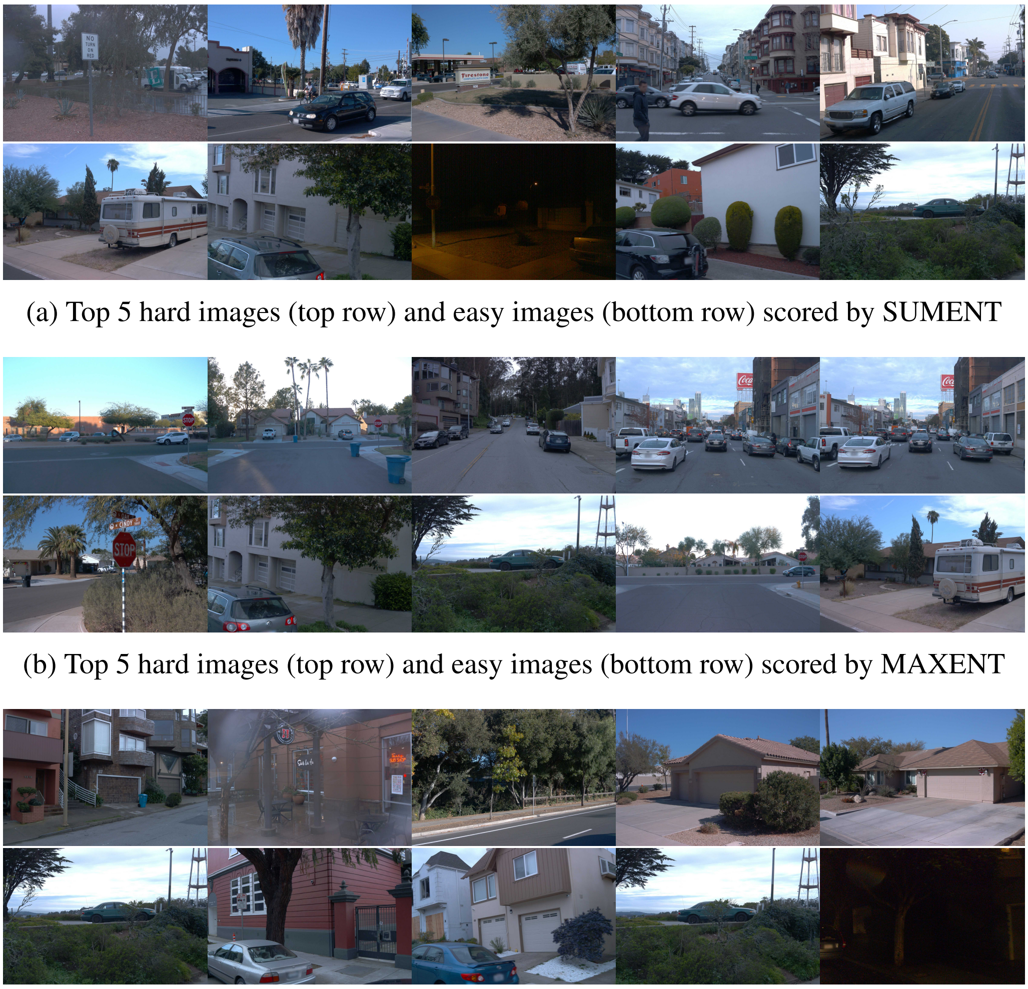
EXP 3 - Object Detection - Waymo Open

We explore how AL behaves in a realistic autonomous setting using the real-life dataset Waymo Open (“Waymo Open Dataset: An autonomous driving dataset”, 2019). Detectron2’s Faster R-CNN is used for this task, and the active learners SUMENT, MAXENT, and AVGENT are used for comparison.

Experimental Details: We use the identical setup as to EXP 1, except the following: For the dataset, we use an initial training set of size 2000, a test set of size 3000, and an unlabeled set of size 24 154. Also, all learners use the No-Spectrum selection method.

Evaluation: The AP50 of each learner is presented in Figure 4 (left) and the hard/easy samples chosen by each learner in Figure 5. The results are different from what we observe in the earlier experiments. MAXENT has the best overall performance, AVGENT

2019 for more details.



(a) Top 5 hard images (top row) and easy images (bottom row) scored by SUMENT

(b) Top 5 hard images (top row) and easy images (bottom row) scored by MAXENT

(c) Top 5 hard images (top row) and easy images (bottom row) scored by AVGENT

Figure 5: EXP 3 - Images from Waymo Open scored by the active learners from ALI 23. See "EXP 3" in Appendix A for more results.

performs the worst out of all active learners; it is highly sensitive in collecting images containing false positives as discussed in EXP 1. This experiment was also run using early stopping; see "Early Stopping" in Appendix A.

5 Discussion

Our results demonstrate that most active learners can achieve a better overall performance than the baseline learner. In addition, most active learners can reach the same maximum performance as the baseline learner using a smaller training set. However, we observe that the number of total class instances in the training set has a great impact on a learner's performance.

Instance-rich Samples: We notice that the object detection learners (e.g., SUMENT, MAXENT, AVGENT) end up with varying numbers of total instances in their training sets. This can be seen in Figure 1 (right) and Figure 4 (right) for EXP 1 and EXP 3, respectively. The SUM learners, having the highest instance-count, benefit from selecting images containing many objects, as most of these might be highly informative. In addition, these might contain hard-to-detect and rare objects that result in diverse scenarios in the context of autonomous perception.

False Positives: The number of false positives predicted by a learner can be profoundly affected by the complexity of the dataset. AVGENT has the worst overall performance as it tends to score images containing a single, low confidence, false positive, as highly informative. In addition, most of the images selected by AVGENT contain few true positives, as observed in EXP 1. Our results are reflected by Brust et al., 2018, where SUM is performing best, while maximum and average achieve similar performance. However, their average learner performs significantly better on the PASCAL VOC dataset (Brust et al., 2018) compared to our results. We believe this is due to all images in PASCAL VOC contains at least 1 object.

Performance: The final performance of a learner is heavily affected by the samples selected in the initial training set. In our image classification experiments, we use a balanced initial set containing an equal amount of samples from each class. In the later experiments, we follow typical AL (i.e., initially selecting samples randomly from the unlabeled set) to create an initial training set. We believe that creating a good initial training set is essential to take advantage of AL fully.

Annotation Time: AL reduces the annotation job, since a human annotator ends up annotating a smaller set of informative samples, instead of annotating the entire available dataset. However, the cost of annotating an informative sample itself can vary. For instance, we notice that learners using SUM tend to select images containing multiple instances. For learners using MAX, this number can be less. Annotating informative images selected by SUM might take more time than an image selected by MAX. We leave this trade-off analysis for further work.

6 Ablation Studies

We perform some ablation studies to see how adding samples from the whole informativeness spectrum, using diverse training data and transfer learning affects the AL process.

Spectrum vs. No-Spectrum: We analyze the impact of exploiting (No-Spectrum) vs. exploring the dataset (Spectrum). We re-run EXP 1 and EXP 3 with these two sampling techniques. We compare the active learners MAXENT, SUMENT, and AVGENT. All learners use the same initially trained model as a starting point.

Figure 6 shows the learners using Spectrum vs No-Spectrum. We notice that SUMENT significantly improves using No-Spectrum. We believe this performance increase is due to the SUM learner being able to select samples containing a large number of instances and exploit the dataset. We notice no significant difference for the AVGENT and MAXENT learners.

Data Diversity: Waymo Open contains video segments that are not pre-processed. When we select a frame from one of these segments, the surrounding frames will be very similar. If a frame is given a high informativeness score, a surrounding frame will likely be given the same score. In other words, there is a high chance that similar images are selected for labeling during each ALI; which affects the data diversity.

To improve data diversity, we remove similar images by comparing images and removing similar images from the dataset. We find similar images by using Learned Perceptual Image Patch Similarity (LPIPS) metric (Zhang et al., 2018).

From Figure 6 (right), we notice no significant difference by using diverse samples. This might be due to that the model is able to exploit similar images and improve convergence. We notice that the MAX learners perform slightly worse by not using diverse samples.

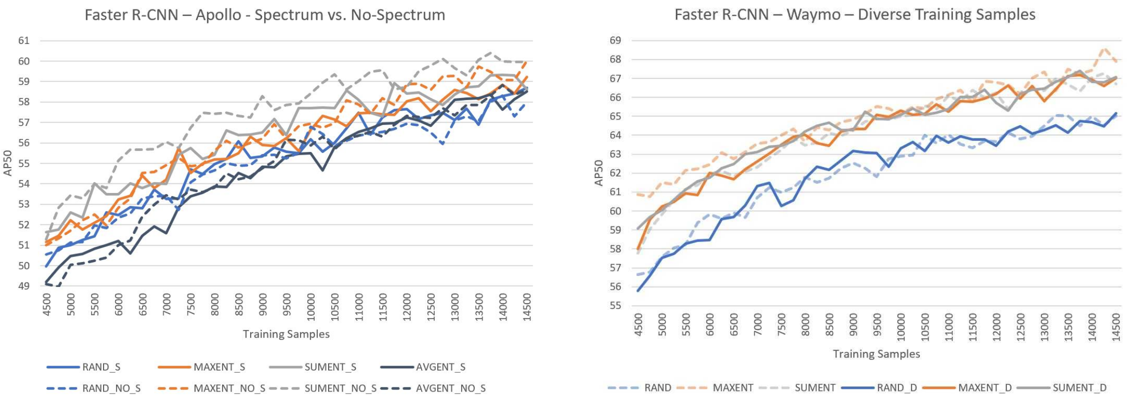


Figure 6: **Left:** Spectrum vs. No-Spectrum on Apollo - Performance results using the selection methods Spectrum (solid lines) and No-Spectrum (dashed lines). From ALI 10 to 50. **Right:** Diversity vs. No Diversity on Waymo - Performance results using diverse training sets (solid lines) and non-diverse datasets (dashed lines). From ALI 5 to 50.

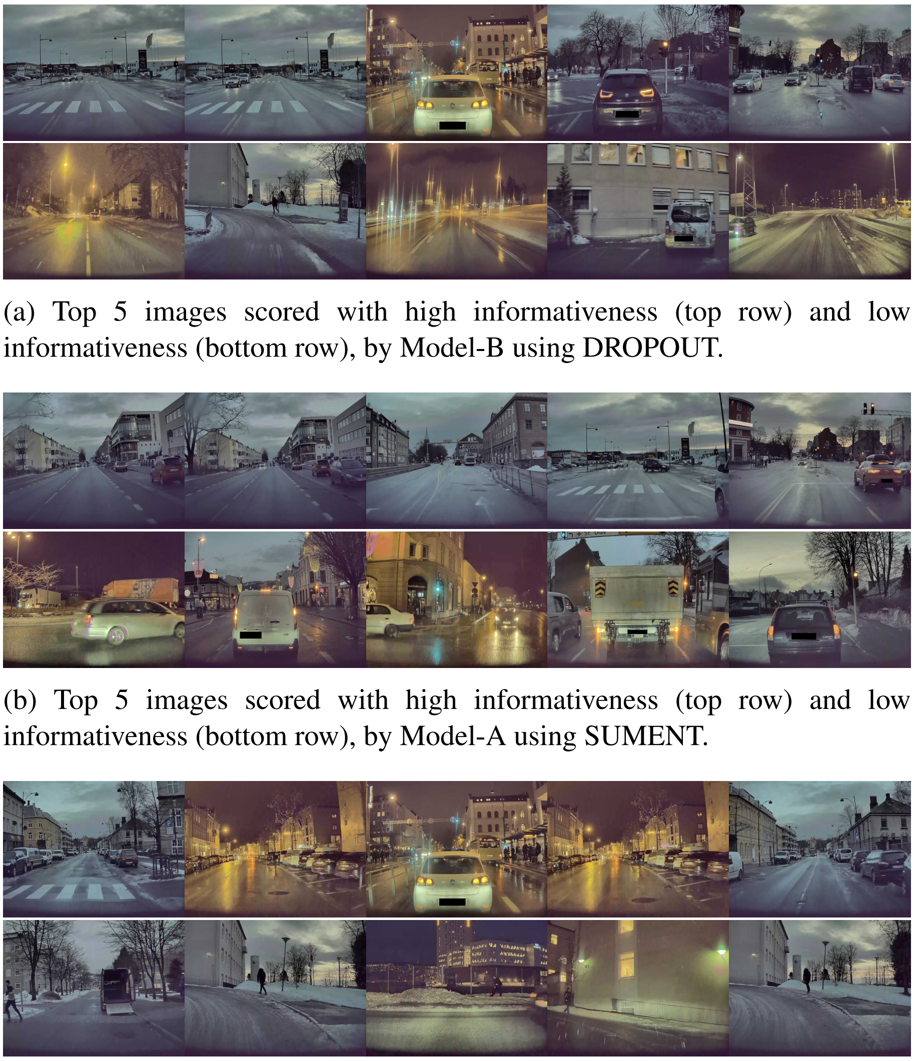
Transfer Learning - NAP-Set: AL can be accelerated, by initializing it with transfer learning (Kale and Liu, 2013). Our goal is to find an initial set of informative samples from a never-before-seen dataset using transfer learning. For this experiment, we use two different models that are previously trained on self-driving datasets, to search for informative images on the in-house NTNU Autonomous Perception Lab's (NAPLab's) raw traffic data (NAP-Set). The first model (Model-A) is the best performing Faster R-CNN MAXENT model from EXP 3. As the second model (Model-B), we use one from Detectron2's model zoo that is trained on CityScapes using Mask R-CNN (Wu et al., 2019). Keep in mind, Model-B is pre-trained using the entire CityScapes dataset, while Model-A is trained using 50% of the Waymo Open dataset in EXP 3.

As seen in Figure 7, we observe that Model-B is able to select images that are more instance-rich and complex compared to Model-A. This is mostly due to Model-B being trained for longer on a full dataset, while Model-A being trained for less using AL in EXP 3. Other factors might be that both models are trained on different datasets. In this case, Model-A is trained on Waymo Open, which contains images from the USA, and Model-B is trained on CityScapes, which contains images from Germany; Germany has a more similar city structure as Norway. We notice that Model-B, gives images containing high density and overlapping objects high informativeness, while images with low informativeness often contain few objects.

7 Conclusion

We explore Active Learning (AL) and evaluate its effectiveness on object detection, and instance-based segmentation in an autonomous domain. We propose a novel pool-based AL Framework based on state-of-the-art Faster-R-CNN and Mask R-CNN. Our results demonstrate that active learning outperforms the random selection baseline by selecting highly informative samples for training. Our results indicate that AL reduces the amount of training data required, and can potentially minimize the annotation job required for autonomous vehicles.

Our AL Framework is modular and can be configured to run custom experiments. It includes various uncertainty-based active learners that are trained and compared to a random-selection baseline. Other interesting findings indicate that learners using SUM as



(a) Top 5 images scored with high informativeness (top row) and low informativeness (bottom row), by Model-B using DROPOUT.

(b) Top 5 images scored with high informativeness (top row) and low informativeness (bottom row), by Model-A using SUMENT.

(c) Top 5 images scored with high informativeness (top row) and low informativeness (bottom row), by Model-B using SUMENT.

Figure 7: Images from the NAP-Set scored by the models. See "Transfer Learning Results" in Appendix A for more results.

aggregation technique have increased performance as they select instance-rich samples, learners using AVG have decreased performance when introduced to false positives, and transfer learning can be used to create an informative initial training set.

Future Work

Techniques and ideas that might be worth looking into as future work.

Co-Operative Learning: With co-operative learning, both a model and a human annotator can be used for annotating samples. Over time, a model using AL will become better at differentiating between samples with high and low informativeness. Samples having low informativeness contain detections with low informativeness. These detections can either be used to automatically annotate images or give annotation suggestions that can be approved by a human annotator for quality assurance. Doing so can save more time in annotating and make the labeling process even faster.

White-Box Query Strategies: A white-box query strategy can be used to make more complex measurements by considering other metrics than just the output of the model. However, this requires that you understand the model's architecture. Using a white-box query strategy had a better overall performance compared to black-box query strategies, as seen in Roy et al., 2018.

Balanced Training Sets: Methods can be added to prevent unbalanced training sets. Brust et al., 2018 counts minority and majority classes to make sure there is a balance when selecting new, highly informative samples.

References

Bondu, A., Lemaire, V., & Boullé, M. (2010). Exploration vs. exploitation in active learning : A bayesian approach, In *The 2010 international joint conference on neural networks (ijcnn)*. https://doi.org/10.1109/IJCNN.2010.5596815

Brust, C.-A., Käding, C., & Denzler, J. (2018). Active learning for deep object detection.

Ertekin, S., Huang, J., Bottou, L., & Giles, L. (2007). Learning on the border: Active learning in imbalanced data classification, In *Proceedings of the sixteenth acm conference on conference on information and knowledge management*, Lisbon, Portugal, Association for Computing Machinery. https://doi.org/10.1145/1321440.1321461

Gal, Y., & Ghahramani, Z. (2015). Dropout as a bayesian approximation: Representing model uncertainty in deep learning.

He, K., Gkioxari, G., DollÃ¡r, P., & Girshick, R. (2017). Mask r-cnn, In *2017 ieee international conference on computer vision (iccv)*. https://doi.org/10.1109/ICCV.2017.322

Kale, D., & Liu, Y. (2013). Accelerating active learning with transfer learning, In *2013 ieee 13th international conference on data mining*.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. *Lecture Notes in Computer Science*, 21–37. https://doi.org/10.1007/978-3-319-46448-0_2

Morrison, D., Milan, A., & Antonakos, E. (2019). Uncertainty-aware instance segmentation using dropout sampling, In *Computer vision and pattern recognition (cvpr)*.

Roy, S., Unmesh, A., & Namboodiri, V. P. (2018). Deep active learning for object detection, In *British machine vision conference (bmvc)*.

Settles, B. (2009). *Active learning literature survey* (Computer Sciences Technical Report No. 1648). University of Wisconsin–Madison.

Sörsäter, M. (2018). *Active learning for road segmentation using convolutional neural networks* (Master's thesis). Linköping University.

team, B. A. (2019). Apollo synthetic - photo-realistic dataset for autonomous driving.

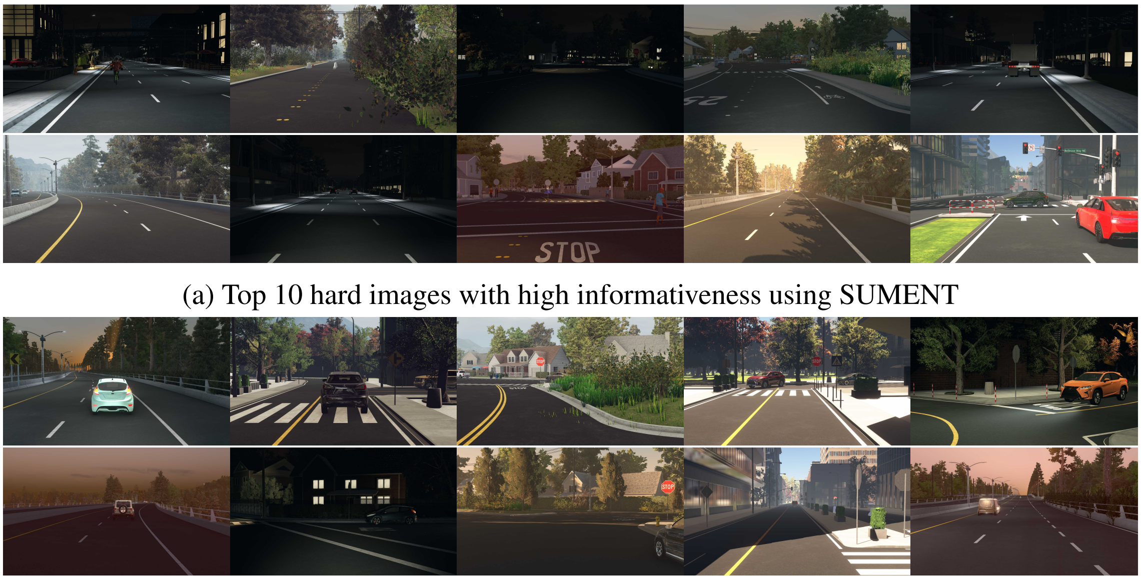
Waymo open dataset: An autonomous driving dataset. (2019).

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. (2019). Detectron2. GihHub.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric, In *Cvpr*.

A Experiment Results

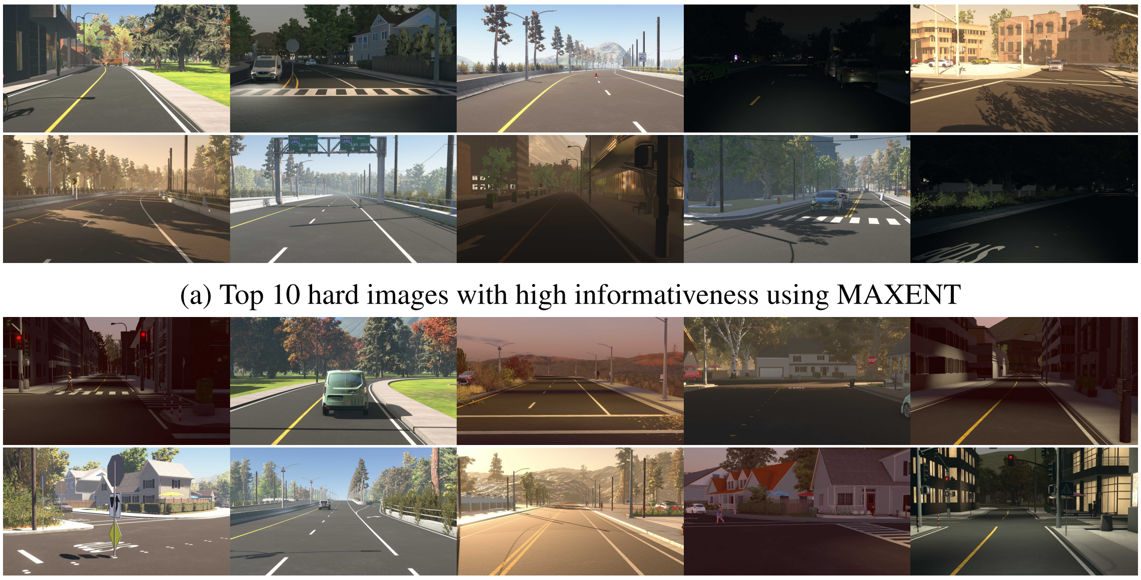
EXP 1



(a) Top 10 hard images with high informativeness using SUMENT

(b) Top 10 easy images with low informativeness using SUMENT

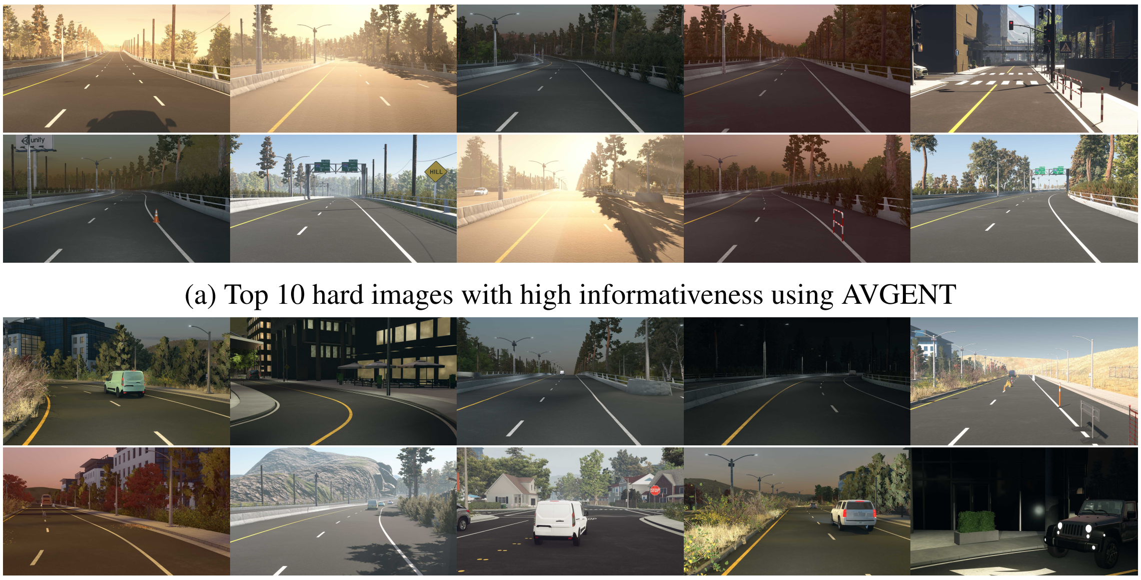
Figure 8: EXP 1 - Apollo - SUMENT Top 10 Hard/Easy Images - Scoring results on ALI 23 from using SUMENT on Apollo Synthetic.



(a) Top 10 hard images with high informativeness using MAXENT

(b) Top 10 easy images with low informativeness using MAXENT

Figure 9: EXP 1 - Apollo - MAXENT Top 10 Hard/Easy Images - Scoring results on ALI 23 from using MAXENT on Apollo Synthetic.

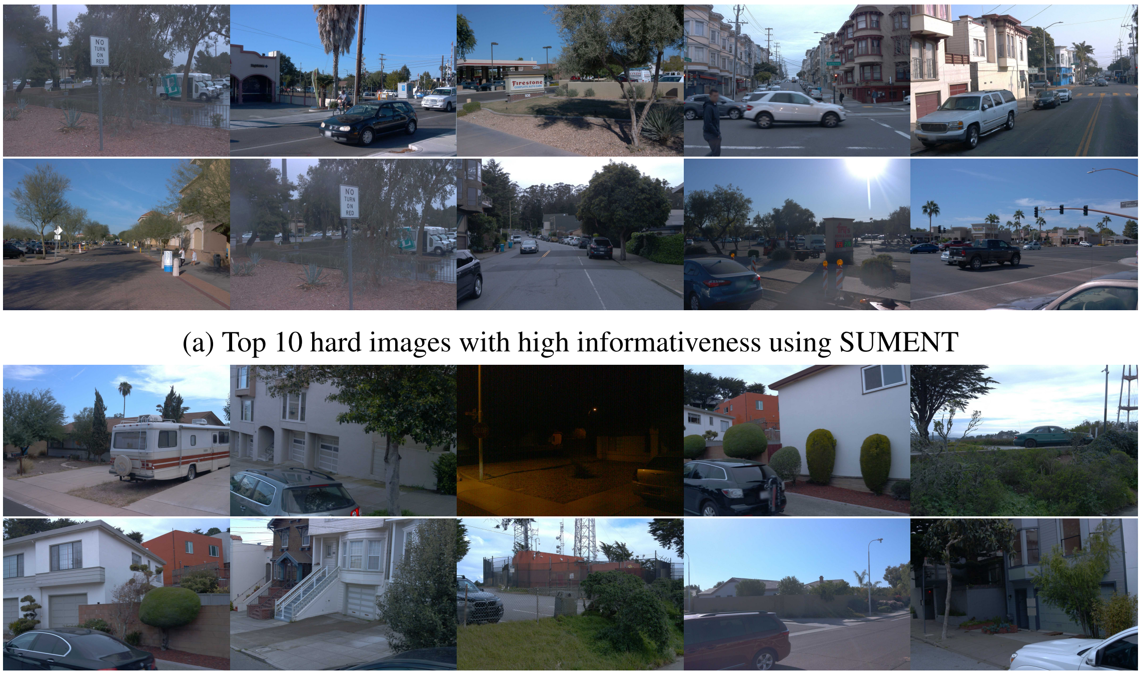


(a) Top 10 hard images with high informativeness using AVGENT

(b) Top 10 easy images with low informativeness using AVGENT

Figure 10: EXP 1 - Apollo - AVGENT Top 10 Hard/Easy Images - Scoring results on ALI 23 from using AVGENT on Apollo Synthetic.

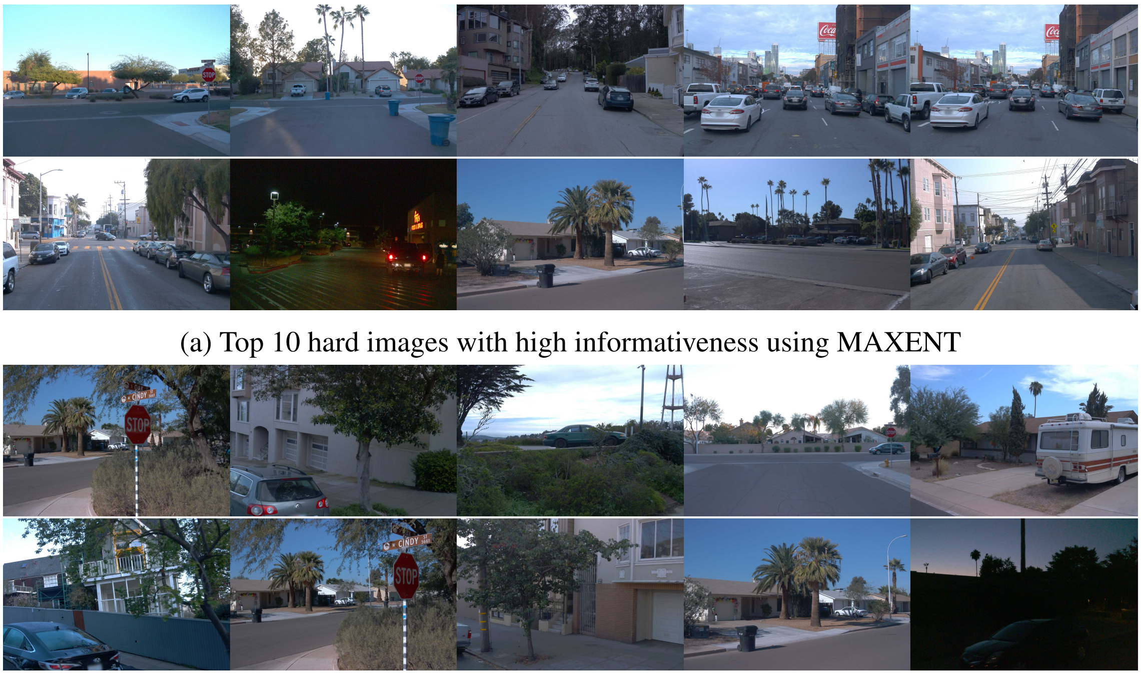
EXP 3



(a) Top 10 hard images with high informativeness using SUMENT

(b) Top 10 easy images with low informativeness using SUMENT

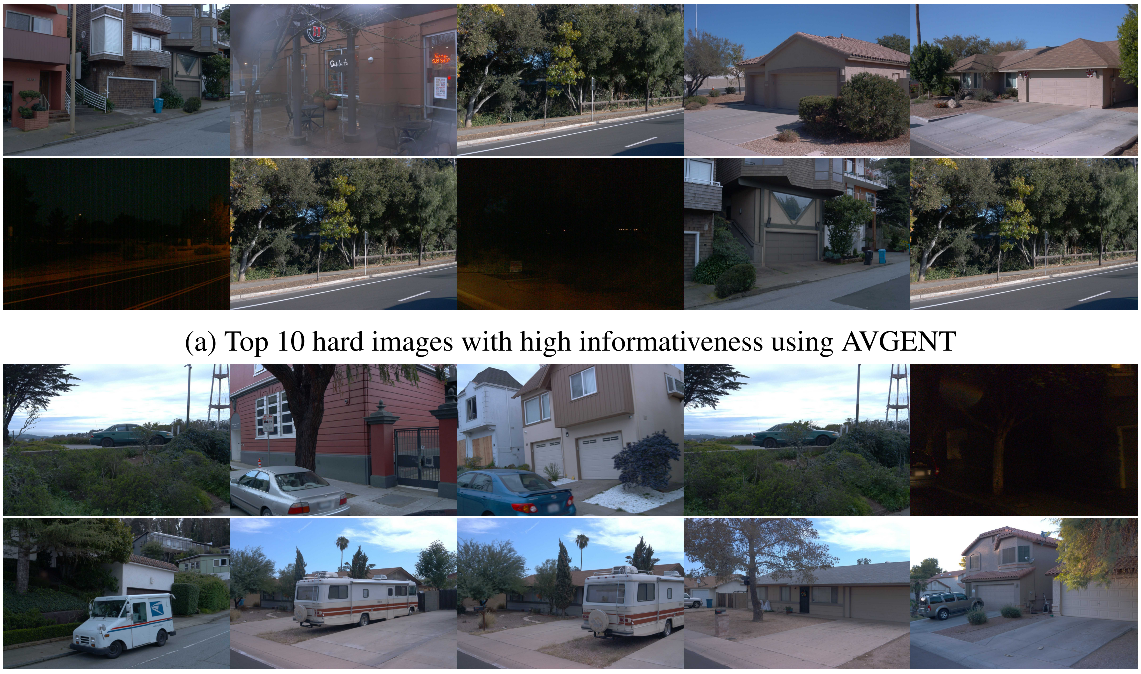
Figure 11: EXP 3 - Waymo - SUMENT Top 10 Hard/Easy Images - Scoring results on ALI 23 from using SUMENT on the Waymo Open dataset.



(a) Top 10 hard images with high informativeness using MAXENT

(b) Top 10 easy images with low informativeness using MAXENT

Figure 12: EXP 3 - Waymo - MAXENT Top 10 Hard/Easy Images - Scoring results on ALI 23 from using MAXENT on the Waymo Open dataset.



(a) Top 10 hard images with high informativeness using AVGENT

(b) Top 10 easy images with low informativeness using AVGENT

Figure 13: EXP 3 - Waymo - AVGENT Top 10 Hard/Easy Images - Scoring results on ALI 23 from using AVGENT on the Waymo Open dataset.

Early Stopping

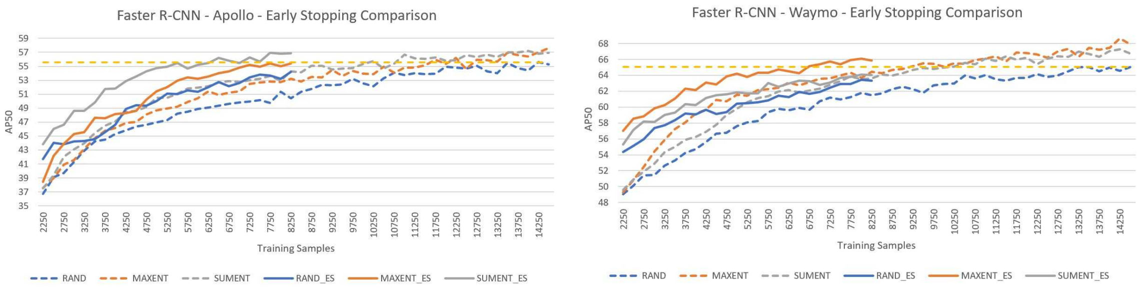
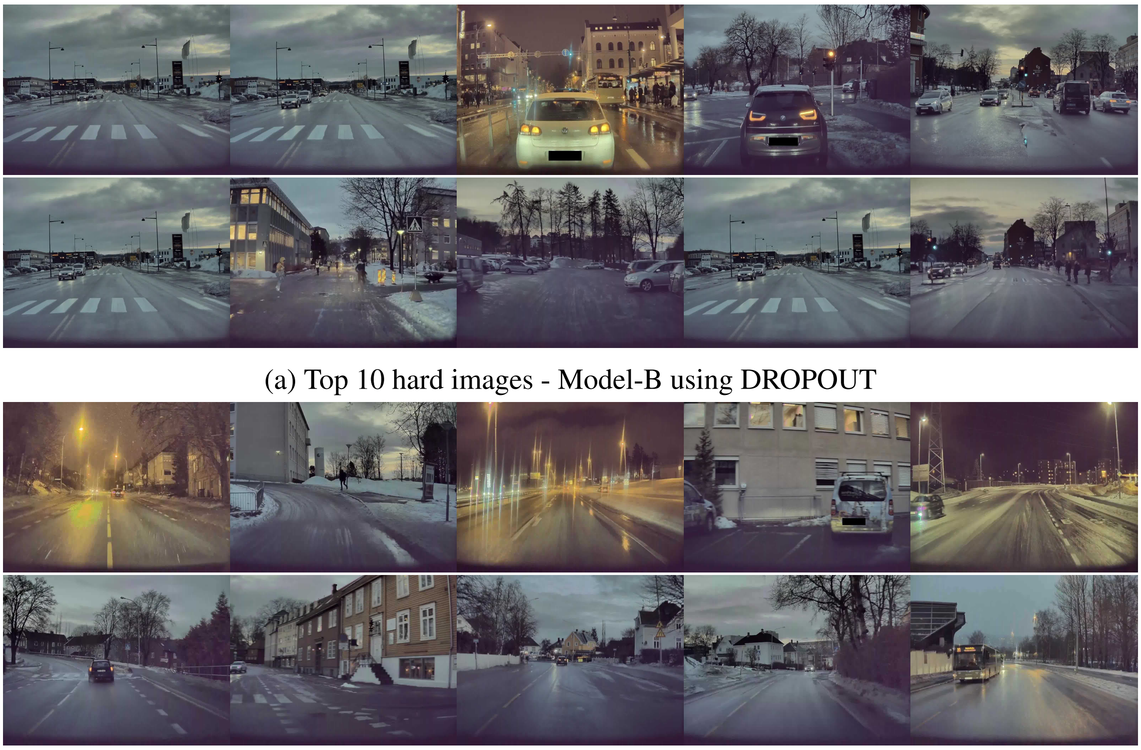


Figure 14: Using Early Stopping. Left: Apollo Synthetic - A run with early stopping (ALI 1 to 25) is compared to EXP 1 (ALI 1 to 50). Right: Waymo Open - A run with early stopping (ALI 1 to 25) is compared to EXP 3 (ALI 1 to 50).

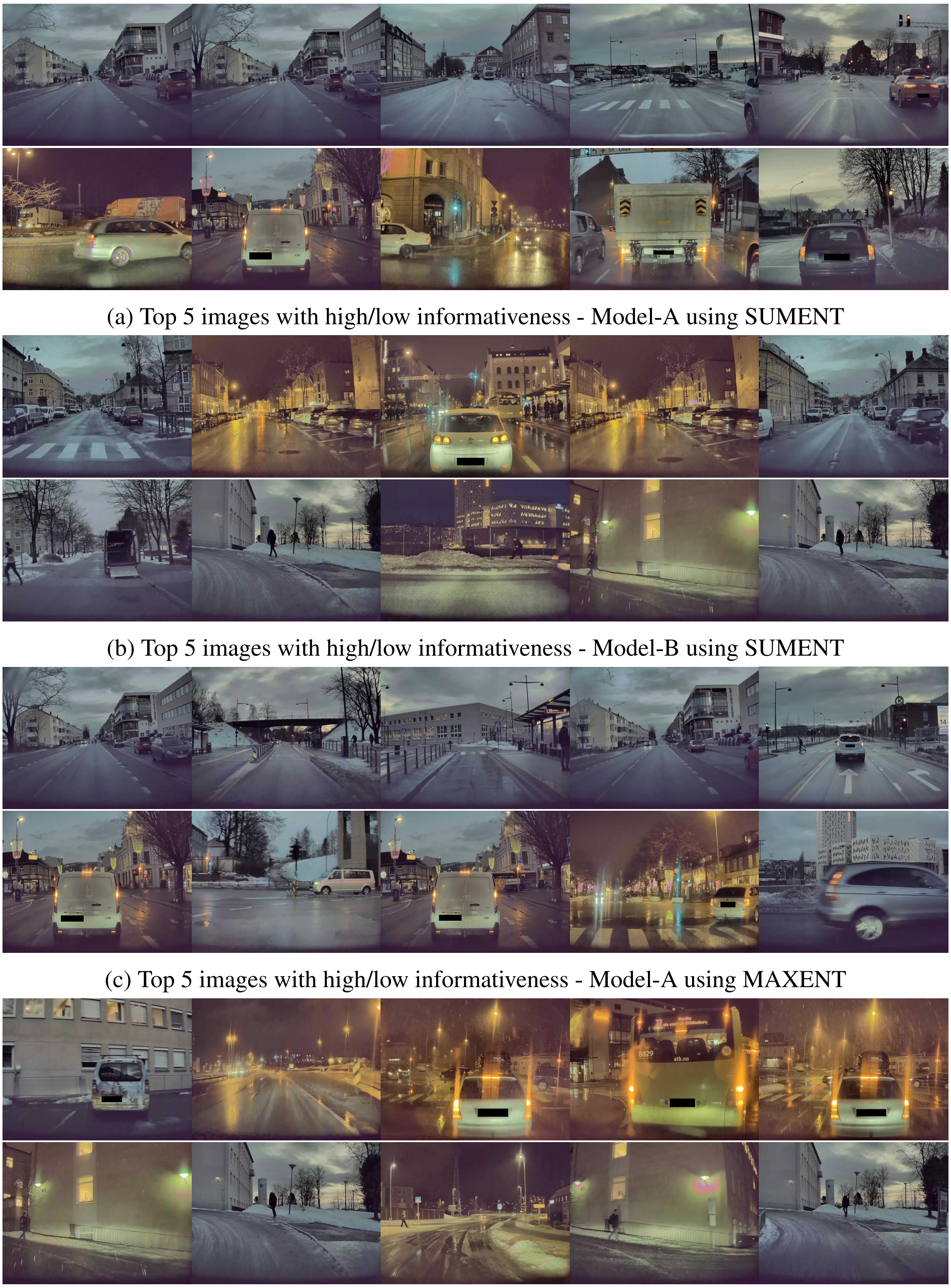
Transfer Learning Results



(a) Top 10 hard images - Model-B using DROPOUT

(b) Top 10 easy images - Model-B using DROPOUT

Figure 15: Images selected from the Trondheim-area by only using Model-B: Fully trained CityScapes model from Detectron2 (Mask R-CNN). **15a:** Top 10 hard images starting with the hardest. **15b:** Top 10 easy images starting with easiest.



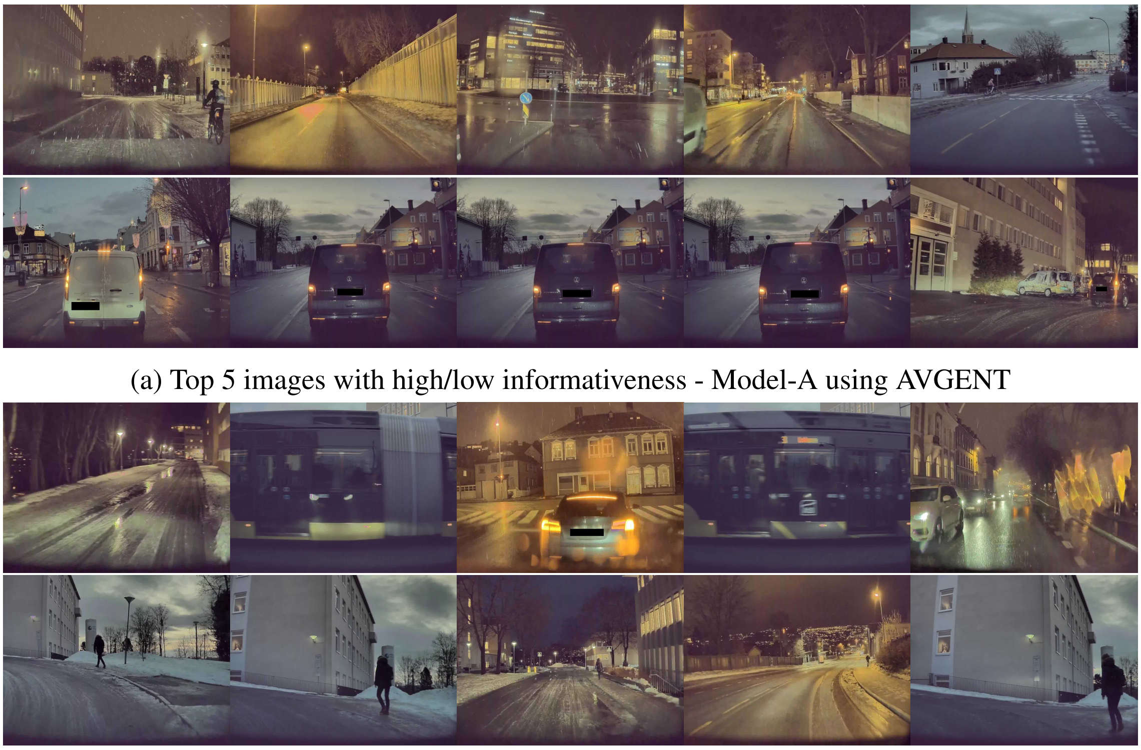
(a) Top 5 images with high/low informativeness - Model-A using SUMENT

(b) Top 5 images with high/low informativeness - Model-B using SUMENT

(c) Top 5 images with high/low informativeness - Model-A using MAXENT

(d) Top 5 images with high/low informativeness - Model-B using MAXENT

Figure 16: Images selected from the Trondheim-area. Model-A: Best performing MAXENT learner from ALI 50 in EXP 5 (Faster R-CNN). Model-B: Fully trained CityScapes model from Detectron2 (Mask R-CNN). **Top row in each subfigure:** Top 5 images with high informativeness starting with the highest informative. **Bottom row in each subfigure:** Top 5 images with low informativeness starting with the least informative.



(a) Top 5 images with high/low informativeness - Model-A using AVGENT

(b) Top 5 images with high/low informativeness - Model-B using AVGENT

Figure 17: Images selected from the Trondheim-area. Model-A: Best performing MAXENT learner from ALI 50 in EXP 5 (Faster R-CNN). Model-B: Fully trained CityScapes model from Detectron2 (Mask R-CNN). **Top row in each subfigure:** Top 5 images with high informativeness starting with the highest informative. **Bottom row in each subfigure:** Top 5 images with low informativeness starting with the least informative.