Safety-aware Active Learning with Perceptual Ambiguity and Severity Assessment

Prajit T Rajendran[1,*], Guillaume Ollier[1], Huascar Espinoza[2], Morayo Adedjouma[1], Agnes Delaborde[3] and Chokri Mraidha[1]

[1]*CEA, List, F-91120, Palaiseau, France*

[2]*KDT JU, Avenue de la Toison d'Or 56-60, 1060 Brussels, Belgium*

[3]*Laboratoire National de Metrologie et d'Essais, Trappes, France*

Abstract

Deep Neural Networks (DNN) used in self-driving cars need a large data coverage and labelling to manage all potential hazards in safety-critical scenarios. Active learning approaches make use of automated data selection and labelling that can build diverse datasets, with less human costs and more accuracy. Traditional active learning methods consider uncertainty of the model predictions and diversity of the data points for query selection. However, they are not optimal in capturing many critical data points, which are potentially risky with respect to safety considerations. In this position paper, we propose a novel approach that uses human feedback related to perceptual data ambiguity and a criticality score, linked to system-level safety assessment. This approach includes a continual learning model that learns to identify corner cases and blindspots with high impact in potential risk, and combines them with uncertainty-sampling and diversity-sampling models to create a safety-aware acquisition function for active learning.

Keywords

Safety, Active learning, Autonomous driving, Human-in-the-loop learning,

1. Introduction

Self-driving cars are increasingly employing various deep learning-based components in their technology stack. These components require tremendous amounts of data to reach a significant level of performance [1]. Deep Neural Networks (DNN) generally perform poorly when they come across previously unseen data. A DNN model trained on only a homogeneous set of images from a particular scenario would perform well only in that scenario and under-perform in most other situations. This is a major concern for the safety assessment of self-driving vehicle systems [2]. In a traffic light classification task for instance, the more the diverse scenarios the DNN module encounters in training, the wider is its safe operation region [3].

Typically, the labels to train such modules are provided by humans [4]. Curating a large dataset with millions of human labels is painfully time consuming and expensive. Active learning is a powerful technique that attempts to maximize a model's performance gain while annotating the fewest samples possible. This process usually considers factors such as uncertainty and diversity to generate a query list to the human [5]. Active learning has shown impressive performance gains over random selection in many self-driving perception tasks.

While there have been emerging efforts to improve active learning for complex scenarios, little attention has been given to active learning for safety-critical features. One example of these features is the detection of ambiguous data points when the self-driving car is in a safety-critical situation. An example for ambiguity could be an image used to train a traffic light detection system wherein there is a red light for traffic intending to turn right and green light for the straight-moving traffic. This image could be delegated to the human to annotate if it is deemed to have a high impact on potential risk.

This position paper proposes a novel approach that uses human feedback related to perceptual data ambiguity and a criticality score. This criticality score, which is linked to the exposure and severity factors of a typical safety assessment, helps to characterize the criticality context of corner cases and blindspots with high impact in potential risk. In a limited query budget scenario, perceptual ambiguity level and criticality level obtained during the annotation process, along with uncertainty and diversity measurements help in selecting the images with highest impact on potential risk. This position paper is a preliminary step towards deeper research into how human-in-the-loop feedback can help in a safety-aware active learning approach.

2. Background and Related Works

2.1. Motivation

A modular driving system typically consists of several components with specific functions collaborating to achieve the intended driving behaviour. There are also end-to-end driving systems, but these are usually entirely made up of opaque blackbox models and thereby it is not feasible to certify their functional safety. Learning enabled components making use of black box machine learning models are notorious in this aspect due to their lack of transparency. Failures or unsafe behavior at the component level can potentially compromise the safety of the entire system unless there are exhaustive system level measures to tackle them, and thus it is important to ensure that the component is trained in a manner so as to minimize vulnerability to unknown situations. The presence of a human in the loop could help in mitigating some of these vulnerabilities by identifying certain blindspots undetected by the trained models and by assessing the severity level of the consequence of misprediction by the trained models. In situations of limited query budget and training time, the paradigm of active learning could assist in selecting the most safety relevant data points by analyzing the blindspot vulnerabilities of the component. In this work, we focus on improving the data selection and training of a traffic light classification component in a modular driving system.

2.2. Active Learning

Active learning is a process of eliciting training data from annotators to determine the right data to put in front of people when you don't have the budget or time for human feedback on all your data. This is especially true in datasets for autonomous driving, which could have millions of hours of data available for training. More than the raw quantity of the data used, the quality, diversity and usability of the data are the important parameters to assure optimum performance and safety of the deployed models. The deep neural networks responsible for self-driving functions require exhaustive training, and the data needs to cover new and uncertain situations in order to tackle the problems of unknown unknowns. Unknown unknowns are data points for which the AI model provides a wrong prediction with a high degree of accuracy. Such points are dangerous because they are immune to detection by uncertainty measures, which are often used as a proxy metric to test models' weaknesses. The combination of data annotation and curation poses a major challenge to deploying deep learning models in autonomous systems and active learning helps by automatically finding the relevant data points to query the human, to build better datasets in a fraction of the time,

with less cost and more accuracy [6]. In this work, we focus on pool-based active learning, where we have a small set of labelled data available and a large set of unlabelled data which needs to labelled within a certain query budget.
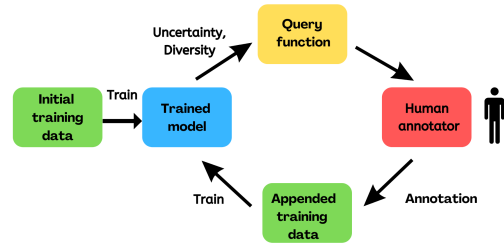

Figure 1: Block diagram of the active learning process

Some of the sampling strategies in active learning are as follows [7]:

- Random sampling is a strategy where we pick random samples from the unlabeled pool of data as query points for the human to label. This is usually used just as a baseline as it does not have an intelligent strategy to select the query points.
- Uncertainty sampling is the set of strategies for identifying unlabeled items that are near a decision boundary in the trained model. This approach basically picks out the data points with a higher predictive uncertainty, and is thereby reflective of the blindspots of the trained model.
- Diversity sampling is the set of strategies for identifying unlabeled items that are underrepresented or unknown to the machine learning model (for instance, features that are not common in the training data, or are under-represented in real world demographics)

The simplest approach in literature as illustrated in [8] is to select examples based on distances in the feature space. In [9], diversity is measured using a similarity matrix made using the Gaussian kernel of the distance between two points. [10] makes use of entropy as a metric of uncertainty.[11] makes use of information density of the candidate instance obtained from the input space for the remaining unlabeled in- stances. [12] and [13] use ensemble and Bayesian methods to approximate uncertainty respectively. [14] proposes heuristic methods to balance between the uncertainty and the represen- tativeness of the selected sample, considering the redundancy between selected samples. [15] argues that the initial model does not have a good performance so the queries generated by it are also likely to be inefficient. In [16], it

is proposed to include knowledge from unlabeled images, by adding unsupervised and semi-supervised methods to enhance the performance. The authors in [17] proposed to use a binary classifier to predict if an image is from the labeled or unlabeled pool using the concept of adversarial learning. In [18], a semi-supervised active learning approach is proposed wherein contention points are determined by making use of both the informativeness and adaptive probabilistic label of the unlabelled points based on the hypothesis of the current model.

2.3. Blindspots and Corner Cases

Blindspots are the deficiencies that are present in a model which may be detrimental to its performance and adaptability to unknown and uncertain situations [19]. In active learning, data points falling under these blindspots can be specifically picked to query to a human oracle. There can be various categories of blindspots:

- **Model Blindspots:** The set of data points, and the feature regions they enclose, in which the model is highly uncertain about or unsure of its predicted label constitute the model blindspots. It is possible to identify model blindspots using the prediction uncertainty of data points. Data points for which the model has a prediction with a high entropy fall under this category.

- **Data Blindspots:** The areas of the feature space that are not covered in the training set constitute the data blindspots. Diversity is one of the aspects that help in uncovering these blindspots. An example could be a dataset with images only recorded in daytime. An image taken at night time would be very distant from the images that the model has seen before, and even if the model's output prediction has a low entropy, it can not be fully trusted.

- **Human-identified Blindspots:** The model blindspots reveal the underconfidence and knowledge gaps of the trained model, and the data blindspots explore the diversity of the data. However, there may be more conceptual aspects in the dataset which are not covered under both the above categories of blindspots. For example, consider an image in the training set of a traffic light classification system wherein there are two visible traffic lights- one for left moving traffic, and the other for straight-moving traffic. If the ego vehicle is in the rightmost lane, a human looking at the image can see that the vehicle could not possibly turn left so only the signal light for straight-moving traffic is relevant for the scene. This however is an ambiguous situation that could be potentially difficult to classify

without conceptual knowledge about the traffic, which a blackbox model may not necessarily possess. Such blindspots can be identified with the help of a human-in-the-loop.

- **Safety Blindspots:** Datapoints whose misclassification by the specific trained model at the component level could compromise the safety of the system which the component is a part of, constitute safety blindspots.

3. Proposed Method

In contexts which are subjective in nature or when human contextual knowledge plays a major role, current active learning methods based purely on model knowledge do not tend to perform well [2]. Safety in particular is a complex concept involving other environmental and situational factors. Since the onus in active learning is on a particular component, one can not discuss safety as it is a system-level concept. However, it is possible to think about the safety implications of a mislabelled or ambiguous data point. A human-in-the-loop can help in identifying certain conceptual blindspots which are not covered under the model and data blindspots as discussed in the section above. Although human-in-the-loop involves effort in terms of labelling, active learning acquisition functions ensure that only a fraction of the data points which are most critical according to the chosen criterion have to be labelled by the humans, thereby solving the scalability issue. Human bias is always a factor in labelling but classic methods in active learning such as inter-annotator agreement can be used to mitigate this problem.

3.1. Perceptual Ambiguity

Data points which the annotator perceives to be potentially ambiguous could be rejected and removed from the training set. However, a black and white approach of reject and accept is not suitable in many cases, such as traffic related tasks. Many data points could be slightly ambiguous yet interesting to include in the dataset for diversity and task relevance. Conservatively rejecting all data points the annotator perceives to be slightly ambiguous leads to lesser diversity in the training set. These constitute human-identified blindspots and provide additional information for data selection. Thus, it would be useful to quantify the level of ambiguity and underconfidence that the annotator feels for each data point as very low, low, medium, high or very high. A secondary model can be trained to predict the level of perceptual ambiguity with the help of human feedback and this could assist in better data selection for active learning querying under a limited budget. We propose table 1 as reference for the

Table 1
Perceptual ambiguity levels

Level	Explanation
Very low	Unambiguous image, label easy to identify
Low	Distracting features but easy to classify
Medium	Some ambiguities in identifying the label
High	Occlusions and ambiguities, hard to classify
Very high	Corner case with safety implications

annotators:

Figure 2: Ambiguous class labels and distracting features

Figure 3: Ambiguous class labels and conceptual understanding

Consider figure 2 from the traffic light detection dataset presented in [20]. There are two traffic lights visible in the image, which is a source of ambiguity. Additionally, at night the tail lights of traffic ahead may constitute distracting features which may affect the label prediction. In figure 3 also from the same dataset, one can see that once again there is an ambiguity in the class label on first sight. However considering that the ego vehicle is in the middle lane, with proper conceptual knowledge it can be presumed that the traffic light for straight-moving traffic is the relevant one.

3.2. Criticality Assessment

We consider the safety awareness of the data labeling process through the concept of criticality assessment and thereby aim to tackle safety blindspots discussed above. The idea behind it is to estimate the importance of a specific image regarding a task according to the global risk it could represent on a system facing that task. The global risk is here the combination of two factors. These two factors are the severity, i.e., estimated safety consequences if the system fails the task, and the exposure, i.e., the estimated probability of this fail. In the context of traffic light classification, this severity concerns the expected consequences if this traffic light is misclassified, and it will depend on which class is misclassified (i.e., green light misclassified as red/orange light, or red/orange light misclassified as green light) and the different visible environmental parameter which can participate in the possible accidents (e.g., pedestrian crossing, road intersection). The exposure is estimated by detecting the different visible factors that could cause the misclassification (e.g., camera obstruction or corruption, weather conditions). We focus here on the risk assessment at the component level without considering the whole system's capabilities and interactions with the other components and subsystems.

To include this active learning approach in a complete safety engineering process, the requirements identified in the preliminary analysis shall be considered to adapt this score on it. A first question to estimate the severity level is presented to the human annotator. We formulate the question as *"How do you estimate the consequences on accidents risk if the automated driving system misclassifies this traffic light?"*(As shown in Figure 5) with the possible answers *"Negligible"*, *"Light"*, *"Severe"*, and *"Fatal"*. We associate each of these answers to a value (zero for *"Negligible"*) and if the human rater do not select the answer *"Negligible"* i.e., if the severity score is higher than zero, we ask another question for the exposure estimation: "Can you see any factor that might bother this traffic light identification?" with the answers *"Yes"*, *"No"*. If the rater answers "No", the exposure value is zero. Else we ask additional questions to identify these factors. Each factor is associated with an exposure value defined in amount by the expert and not visible by the human rater. We can then compute the criticality score with the formula $\left(\sum_{k=1}^{n} f_k * e_k\right) * s$ where n is the number of identified factor, f is a boolean vector which represent the presence/absence of a factor, e is a vector that represent the exposure value for each factor and s is the severity score.

3.3. Continual Learning Model for Perceptual Ambiguity and Criticality

A continual learning approach would be suitable in a human-in-the-loop environment when the human can initially provide labels and eventually a simple model (different from the main component that is being trained) would be able to replace the human when it reaches a sufficient level of performance. Note that before the continual learning model's misclassifications would only affect the data selection and not the predictions of the main component directly. Along with providing the class labels, the human annotator can be asked to provide the perception ambiguity and severity level associated with the data point. Thus there can be two separate continual learning models attached to the main component model- one to predict perceptual ambiguity and one to predict the severity level of the data point. The model used here could be a shallow neural network with the intermediate features from the main component model.

An issue with the continual learning approach is catastrophic forgetting when the model updates itself constantly and forgets what it learnt before. To avoid this, it is necessary to maintain the best representation set of what the model already knows so that when the model is re-trained it can also include this representation set. In this work, we make use of a buffer called the familiarity buffer for this purpose. The familiarity buffer holds a representation of the data points where the model predicts the perceptual ambiguity or the criticality of the data point accurately. When the model encounters data points wherein there is mismatch between the model prediction and the human feedback, that data point is populated in the unfamiliarity buffer. When the unfamiliarity buffer is full, the continual learning model is retrained with the contents of both the familiarity and unfamiliarity buffers. After the re-training, the familiarity buffer of size 'n' is updated. From the contents of both the buffers, the most diverse 'n' data points are chosen to repopulate the familiarity buffer. Finally, the unfamiliarity buffer is emptied.

3.4. Uncertainty and Diversity

The model blindspots and data blindspots can be captured by uncertainty and diversity respectively. They can be calculated as follows:

- **Uncertainty-based querying:** In uncertainty-based querying, the model's uncertainty about its predictions is used as a metric for selecting query points [10]. The model predictions typically contain probability scores associated with each class label. In the ideal scenario, the model should allocate a probability of one to the correct label and zero to all the incorrect labels. Thus, entropy

can be used as a measure of the self-evaluated confidence of the model in its own predictions. Zero entropy means that the model is perfectly confident in its prediction while an entropy of one is the level of maximum doubt. Entropy of a model with 'c' classes with each class 'i' having a probability p_i is defined as follows:

$$Entropy = -p_i \sum_{n=1}^{c} p_i log(p_i) \qquad (1)$$

Thus, data points with a higher entropy are those with a higher level of uncertainty attached. The queries can be generated such that the most uncertain data points are shown to the human for review. In this work, we use an ensemble of models as in [21] to generate the average predictive entropy.

- **Diversity-based querying:** The diversity-based querying approach aims to include the data points most different from what the model has previously seen [9]. For this, one should store a representation of the training data that the model has been trained on. An ideal candidate for this is the distribution of the features at an intermediate layer of the prediction model. The distribution of features of a fully connected (FC) layer in the later layers of a convolutional neural network for the training data points could be computed and then compared with each new data point to obtain a distance score. In this work, we consider an FC layer with 'N' neurons and compute the means and variances of the output values from that layer for all training data points as a new variant of the existing distance based acquisition functions for diversity such as in [8]. Then, for each new data point, we calculate the Z-score for each of the 'N' features f_1 to f_N and consider their average.

$$Z - score = \frac{\sum_{n=1}^{N} \frac{f_i - \mu_i}{\sigma_i}}{N} \qquad (2)$$

The higher the Z-score, the more distant the new data point from the known distribution. In this approach, the queries would be generated such that the data points with a higher average Z-score are shown to the human for labelling.

4. Proposed Evaluation Framework

4.1. Planned Experiment

The first step in the active learning process is training the initial model using the available pool of labelled data.
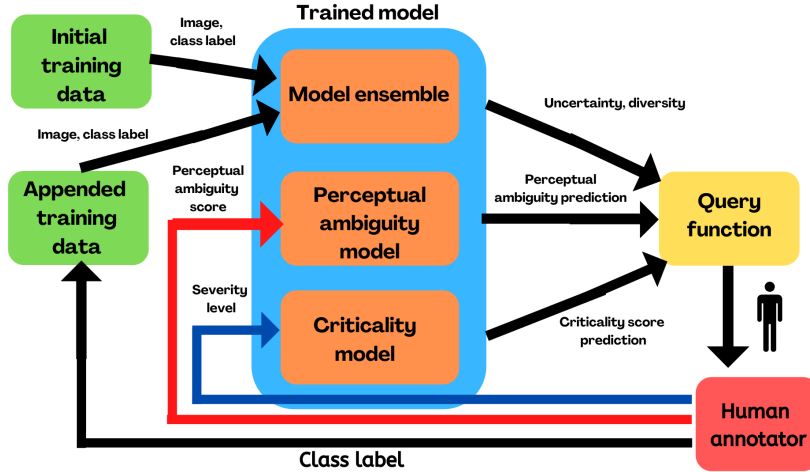
Figure 4: Detailed block diagram of the proposed approach

This model would serve as a starting point to generate queries from the unlabelled set. The large pool of unlabelled data is divided randomly into various chunks. Each of these chunks shall be labelled in a particular round of active learning [22]. In the first round of active learning, the pre-trained model is made use of to generate a query list of the data points to be reviewed and labelled by the human. The selection criteria of the query points is the major challenge in active learning and it depends on the mode of active learning selected as explained above. After all the data points in the first round of active learning are labelled successfully, the model is re-trained with the updated set of labelled data and the next chunk of unlabelled data is selected for the second round of active learning. This process continues till all data points are labelled.

During the labelling process, the annotators are tasked at providing the class label, perceptual ambiguity level and severity level of each data point on a graphical user interface as shown in figure 5. If the data point has a high severity and ambiguity level, additional questions can be asked of the annotators to determine the associated criticality score as mentioned above.

4.2. Active Learning Acquisition Functions

In order to demonstrate the effectiveness of the proposed approach, we propose to perform the experiment with the following combinations of acquisition functions:

- **Random:** In this mode, N% of images are randomly selected from the subset of unlabelled data in a particular round, and are assigned to the human to label
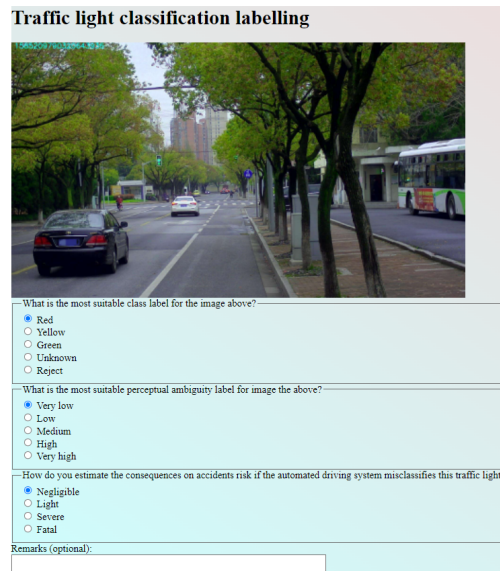

Figure 5: Graphical user interface

- **Uncertainty:** In this mode, the top N% of images with the highest average entropy are assigned to the human to label
- **Diversity:** In this mode, the top N% of images with the highest average Z-score from the current distribution are assigned to the human to label
- **Perceptual Ambiguity:** In this mode, the top N% of images with the highest perceptual ambiguity scores are assigned to the human to label
- **Criticality:** In this mode, the top N% of images

with the criticality scores are assigned to the human to label

- **Combined:** In this mode, the top N% of images with the highest combined average of entropy, Z-score, criticality and perceptual ambiguity score are assigned to the human to label

4.3. Evaluation Metrics

We propose to use the following evaluation metrics to compare the safety and performance of the proposed approach with that of the pre-existing ones:

4.3.1. F1-score

When there is an imbalance in the number of data points in different classes, accuracy might not be a good metric for prediction performance. In this case, F1-score which accounts for both type-I and type-II errors would be a better metric:

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \qquad (3)$$

4.3.2. Uncertainty Reduction

The goal of training a model is to generalize its knowledge over the assigned task and therefore perform well on the unseen test set. As mentioned above, entropy is a good measure of prediction power of a model when the label probabilities are available. Therefore, we can use entropy over the test set as one of the measures of how the model uncertainty is reduced. Note that while the reduction of uncertainty is good, it has to be viewed in tandem with other metrics such as accuracy, precision, recall or F1-score.

4.3.3. Query Relevance

In each round of active learning, N% of the data is selected as query points to be shown to the human. It is necessary to measure if the selected points are indeed the best ones. One of the ways to do this is to measure the difference in the relevant scores (an average of the uncertainty, diversity, criticality and perceptual ambiguity scores for each point) in predictions of the human labelled points and those of the auto-labelled points. The larger the difference between these sets, the more relevant are the selected query points. For the random mode, the query relevance is expected to be the least because the points are selected randomly without considering their relevance in active learning.

4.3.4. Safety-weighted Accuracy

To consider the importance of each input data on the safety relevancy for a machine learning model training,

we can reuse the accuracy metric used to evaluate the performance of classification models and adapt it to criticality aspects. Given the safety requirements identified through the Hazard Analysis and Risk Assessment (HARA) methods and all the relevant Operating Conditions (OCs) visible on the input data, a safety expert identifies the possible hazardous scenarios that could be caused by misclassification of this input data (with a minimum probability of occurrence), and weight the score associated to this input on the visible risk. The OCs are any relevant parameters to describe the system's usage scenarios, including environmental conditions, dynamic elements, and scenery. As the criticality assessment presented in section 3.2, the risk evaluation is decomposed into severity and exposure factors. We estimate for each input the Safety Integrity Level (SIL), presented in the IEC 61508[23] standard, with the severity and exposures scores and the risk matrix. We then give each input an integer score between one to four, and we compute the model safety-weighted accuracy as follows:

$$\frac{\sum_{k=1}^{n} sil_k * c_k}{\sum sil}$$

With n the number of predictions, sil the vector with the SIL scores for all inputs, and c a vector with the values of the classification correctness (one if the classification is correct and zero else).

5. Conclusion

In this paper, we introduced the concepts of perceptual ambiguity and criticality, and proposed a model which learns to predict these through continuous feedback from a human in the loop. The proposed approach is aimed at tackling blindspots not covered under current approaches dealing with the uncertainty and diversity sampling methods. An experiment was designed with the goal of testing such a model trained to perform traffic light detection. The work is still in an early stage and the next steps include performing an active learning experiment on a large scale with several volunteers, linking the definition of criticality to concrete safety metrics in the industry, development of other evaluation metrics and testing alternate designs of the continual learning model.

Acknowledgments

References

[1] H. M. Eraqi, M. N. Moustafa, J. Honer, End-to-end deep learning for steering autonomous vehicles considering temporal dependencies, arXiv preprint arXiv:1710.03804 (2017).

[2] S. Mohseni, M. Pitale, V. Singh, Z. Wang, Practical solutions for machine learning safety in autonomous vehicles, CoRR abs/1912.09630 (2019). URL: http://arxiv.org/abs/1912.09630. arXiv:1912.09630.

[3] D. Wang, X. Ma, X. Yang, Tl-gan: Improving traffic light recognition via data synthesis for autonomous driving, arXiv preprint arXiv:2203.15006 (2022).

[4] J. Geary, H. Gouk, S. Ramamoorthy, Active altruism learning and information sufficiency for autonomous driving, arXiv preprint arXiv:2110.04580 (2021).

[5] E. Haussmann, M. Fenzi, K. Chitta, J. Ivanecky, H. Xu, D. Roy, A. Mittel, N. Koumchatzky, C. Farabet, J. M. Alvarez, Scalable active learning for object detection, in: 2020 IEEE intelligent vehicles symposium (iv), IEEE, 2020, pp. 1430–1435.

[6] C.-C. Kao, T.-Y. Lee, P. Sen, M.-Y. Liu, Localization-aware active learning for object detection, in: Asian Conference on Computer Vision, Springer, 2018, pp. 506–522.

[7] R. Monarch, Human-in-the-Loop Machine Learning, Manning Publications Co., 2021.

[8] Y. Geifman, R. El-Yaniv, Deep active learning over the long tail, CoRR abs/1711.00941 (2017). URL: http://arxiv.org/abs/1711.00941. arXiv:1711.00941.

[9] G. Wang, J.-N. Hwang, C. Rose, F. Wallace, Uncertainty sampling based active learning with diversity constraint by sparse selection, in: 2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP), IEEE, 2017, pp. 1–6.

[10] A. J. Joshi, F. Porikli, N. Papanikolopoulos, Multi-class active learning for image classification, in: 2009 ieee conference on computer vision and pattern recognition, IEEE, 2009, pp. 2372–2379.

[11] X. Li, Y. Guo, Adaptive active learning for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 859–866.

[12] W. H. Beluch, T. Genewein, A. Nürnberger, J. M. Köhler, The power of ensembles for active learning in image classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9368–9377.

[13] Y. Gal, R. Islam, Z. Ghahramani, Deep bayesian active learning with image data, in: International Conference on Machine Learning, PMLR, 2017, pp. 1183–1192.

[14] T. He, S. Zhang, J. Xin, P. Zhao, J. Wu, X. Xian, C. Li, Z. Cui, An active learning approach with uncertainty, representativeness, and diversity, The Scientific World Journal 2014 (2014).

[15] O. Siméoni, Robust image representation for classification, retrieval and object discovery, Ph.D. thesis, Université rennes1, 2020.

[16] O. Siméoni, M. Budnik, Y. Avrithis, G. Gravier, Rethinking deep active learning: Using unlabeled data at model training, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 1220–1227.

[17] D. Gissin, S. Shalev-Shwartz, Discriminative active learning, arXiv preprint arXiv:1907.06347 (2019).

[18] I. Muslea, S. Minton, C. A. Knoblock, Active+ semi-supervised learning= robust multi-view learning, in: ICML, volume 2, Citeseer, 2002, pp. 435–442.

[19] R. Ramakrishnan, E. Kamar, B. Nushi, D. Dey, J. Shah, E. Horvitz, Overcoming blind spots in the real world: Leveraging complementary abilities for joint execution, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 6137–6145.

[20] X. Yang, J. Yan, X. Yang, J. Tang, W. Liao, T. He, Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing, arXiv preprint arXiv:2004.13316 (2020).

[21] R. Rahaman, et al., Uncertainty quantification and deep ensembles, Advances in Neural Information Processing Systems 34 (2021).

[22] R. Ganti, A. Gray, Upal: Unbiased pool based active learning, in: Artificial Intelligence and Statistics, PMLR, 2012, pp. 422–431.

[23] International Electrotechnical Commission, Functional safety of electrical/electronic/programmable electronic safety-related systems, 2010.