Active Learning for Traffic Sign Recognition: Overcoming Data Scarcity in Autonomous Vehicle Development

Phuong-Nam Tran[1], Nam Van Hai Phan[2], Duc Tai Phan[2], Nhat Truong Pham[3], and Duc Ngoc Minh Dang *[2]

[1] Department of Computer Science and Engineering, Kyung Hee University, Yongin-si, 17104, Republic of Korea
namphuongtran9196@gmail.com
[2] AiTA Lab, Dept. of Computing Fundamental, FPT University, Ho Chi Minh City, Vietnam
nampvhse182309@fpt.edu.vn, phantaiduc2005@gmail.com,ducdnm2@fe.edu.vn
[3] Sungkyunkwan University, Suwon, 16419, Gyeonggi-do, Republic of Korea
truongpham96@skku.edu Springer Heidelberg, Tiergartenstr. 17, 69121 Heidelberg, Germany lncs@springer.com
http://www.springer.com/gp/computer-science/lncs
[4] ABC Institute, Rupert-Karls-University Heidelberg, Heidelberg, Germany

Abstract. Deep neural networks and data-driven intelligence play a crucial role in advancing automated driving, specifically in recognizing traffic signs. However, the dataset in this field is still limited due to the differences in traffic signs across countries. This issue hinders the utilization of artificial intelligence models in automated driving. On the other hand, creating a traffic sign dataset is costly and time-consuming due to the large number and variations of signs worldwide. To address this problem, researchers have introduced a technique called active learning, which aims to reduce labeling costs in training methods. However, applying active learning to the object detection field is challenging due to the unknown number of traffic signs in a given sample. In this paper, we propose an end-to-end system that combines a pre-trained object detection model as a supervisor with the active learning technique and hybrid labeling to address the need for low-budget datasets. Our method offers an efficient labeling approach that significantly reduces human effort by up to two-thirds of the time. Furthermore, our approach achieves performance comparable to using a dataset fully labeled by humans. The source code is available at https://anonymous.4open.science/r/TransferAL-8A42.

Keywords: Traffic signs recognition, traffic signs dataset, autonomous vehicles, object detection, active learning

* Corresponding author: Duc Ngoc Minh Dang (ducdnm2@fe.edu.vn)

1 Introduction

Traffic sign detection and recognition play a crucial role in ensuring road safety and efficient traffic management in the self-driving era. With the rise of autonomous driving technology, it is essential for vehicles to accurately interpret traffic signs in order to travel safely [1, 2]. These signs provide important information such as speed limits and warnings, which help prevent accidents and improve the overall driving experience. As the number of autonomous cars and smart city infrastructure increases, it becomes even more important to develop effective and accurate traffic sign detection systems. However, achieving a high level of performance in these systems requires a high-quality dataset. The lack of a comprehensive traffic sign dataset and the variations in datasets across different countries present significant real-world challenges, as presented in [3].

Deep neural networks and data-driven techniques have shown promise in enhancing traffic sign perception [4]. However, training these networks requires large and diverse datasets, which can be costly and time-consuming to annotate. To address this challenge, active learning [5–7] (AL), a specialized semi-supervised machine learning technique, offers a potential solution. AL supports users in the labeling process, strategically selecting specific instances for labeling to maximize accuracy while minimizing labeling effort. By focusing on the most informative data points, active learning can improve the model's accuracy and potentially reduce the overall annotation effort. However, active learning is mostly applied to the classification problem, with only a few methods available for the object detection case. In the context of object detection, factors such as the quantity and quality of the dataset need to be considered.

To lower the expenses associated with labeling and decrease the amount of human work required in creating an object detection dataset, this paper presents an active learning system designed for traffic sign recognition. The system automatically labels objects in samples with a supervisor model which is a pre-trained model in the same domain as the unlabeled dataset. The supervisor model acts as a guide to help the model decide whether to use machine-generated labels or labels provided by humans. By implementing our system, human workload can be reduced by up to two-thirds, while maintaining a similar level of performance as a dataset that has been entirely labeled by humans. The main contributions of this study can be summarized as follows:

1. Designing an active learning system for traffic sign recognition reduces the human work on creating the traffic sign dataset.
2. Conducting experiments with various combinations of activate learning settings to assess the object detection model's performance on the VTSDB46 dataset.
3. Analyzing the contributions of active learning in creating the dataset and the affect of hyperparameters on the performance of the model.

The subsequent sections of this paper are organized in the following manner: Section 2 offers an overview of the existing object detection and active learning

methods in traffic sign recognition. Section 3 presents our proposed active learning system for traffic sign recognition, detailing its components and processes. Section 4 evaluates the effectiveness of our object detection model and active learning system in various scenarios. Finally, Section 5 summarizes the findings, discusses implications, and outlines future research directions for this work.

2 Related Works

2.1 Object Detection

With the advancement of artificial intelligence, computer vision has experienced remarkable progress in various fields. Object detection, as a crucial aspect of computer vision, has witnessed the introduction of numerous methods and techniques each year. Over time, strategies and models have been developed to enhance the accuracy and efficiency of object identification. The emergence of deep learning has significantly transformed this field, leading to successful models in both one-stage and two-stage object detection. The two-stage object detection method, Faster R-CNN [8], introduced region proposal networks, which greatly accelerated the identification process and enabled real-time performance. This architecture revolutionized the speed of two-stage object detection. In one-stage object detection, the introduction of You only look once (YOLO) and its variants have brought a paradigm shift in researchers' approaches. YOLO offers significantly faster processing compared to the two-stage architecture, although its initial performance was limited. Through iterative improvements across multiple versions, YOLO has evolved to become more accurate and faster than its predecessors. One of the remarkable versions of the YOLO series is YOLOv8 [9] which significantly improves detection performance through the integration of novel features and enhanced architecture design. YOLOv8 introduces anchor-free detection and incorporates advanced techniques such as multi-scale predictions, adaptive anchor computation, and transformer-based modules. These advancements allow YOLOv8 to strike a remarkable balance between speed and accuracy. Consequently, YOLOv8 shows really good performance over a broad spectrum of object identification tasks, including sophisticated detection situations and real-time applications.

In addition to YOLOv8, other methods have been introduced in object detection to address specific challenges, such as detecting small objects and detecting speed. Three notable examples are RetinaNet [10], EfficientNet [11], and SSD [12]. RetinaNet tackles the class imbalance problem commonly encountered in object detection by introducing focal loss. This loss function specifically enhances the detection of hard-to-classify objects. By assigning higher penalties to difficult samples while reducing the influence of easier samples, focal loss effectively balances the backpropagation between foreground and background, resulting in improved object detection performance. EfficientNet, on the other hand, employs a compound scaling technique to achieve high performance with fewer parameters. This novel approach enables the model to achieve a balance

between accuracy and efficiency. By optimizing the model's architecture, EfficientNet achieves faster processing speeds while maintaining excellent accuracy in object detection tasks. SSD is another typical device that does detection in a single pass generating a set of default boxes from bounding box output space that span numerous aspect ratios and scales per place in a feature map, which have made significant strides in pushing the boundaries of object detection performance.

2.2 Traffic Sign Recognition

Traffic sign recognition plays a vital role in enhancing road safety and is a key component of advanced driver assistance systems and intelligent transportation systems. Its objective is to enable vehicles to accurately perceive and interpret traffic signs, thereby facilitating safer and more efficient navigation. Several steps are involved in traffic sign recognition, including detecting signs within images or cameras and identifying their meanings, such as speed limits or stop signs. [13]

A novel method for developing traffic sign recognition systems has been presented in [3], utilizing the GTSDB [14] and GTSRB [15] datasets. However, this approach has limitations in terms of its application to diverse real-world scenarios, as the dataset utilized in this system does not cover all traffic signs in Germany. Another work introduces the VTSDB46 [16] dataset, which aims to enable traffic sign recognition for self-driving cars in Vietnam. This dataset consists of over 80,000 raw and augmented samples, encompassing 46 common traffic signs in Vietnam. It was created in various environments in Ho Chi Minh City to facilitate the development of a robust traffic sign recognition system in Vietnam. Nonetheless, the dataset is still limited due to the absence of other types of traffic signs beyond the 46 common ones included.

These works demonstrate important efforts in traffic sign recognition, but there is a need for further exploration and development of comprehensive datasets that cover a wider range of traffic signs in various regions to ensure the effectiveness and reliability of traffic sign recognition systems in the real world.

2.3 Active Learning

Active learning [5] is a sub-field of machine learning that aims to improve the efficiency of the learning process by allowing the algorithm to select the most informative data for labeling [17]. Active learning's basic theory is that, given less labeled training cases, a machine learning system can reach better accuracy if it has the ability to choose the data from which it learns. In active learning, the algorithm actively selects specific instances from a pool of unlabeled data and presents them as queries to a human annotator for labeling. By focusing on the most informative instances, the algorithm aims to maximize the learning outcome while minimizing the labeling effort. Active learning is particularly advantageous in situations where unlabeled data is abundant or easily accessible, but obtaining labels for these instances is challenging, time-consuming, or expensive. By actively selecting the most relevant samples for labeling, active learning

can significantly reduce the labeling cost while maintaining or even improving the overall accuracy of the learning model.

In different scenarios, there are many different query strategies that can be used. Some common query strategies include uncertainty sampling [7], query-by-committee [18], density-weighted methods [19], and diversity-based instance selection [20]. In uncertainty sampling, the model focuses on selecting instances that have the least certainty score. This strategy can be extended to non-probabilistic classifiers by considering factors like proximity to the decision boundary. On the other hand, query-by-committee strategy employs a committee of models in which each model represents a distinct hypothesis. The instance that elicits the most disagreement among committee members is subsequently selected for labeling. This strategy aims to reduce the version space defined as the set of all hypotheses consistent with the labeled training data. Training a diverse committee model that reflects different areas of the version space and figuring out suitable metrics of disagreement are two difficulties in implementation [21]. Density-weighted methods select instances in areas of high density in attempt to create labels for instances that better reflect the dataset as a whole. This strategy address a drawback of approaches like uncertainty sampling. Density can be used to improve the efficacy of other query algorithms, such as query-by-committee [19, 21]. The last one is diversity-based instance selection was employed to select a diverse set of instances, ensuring that the labeled data accurately represents the variety in the dataset [5, 22]. However, a challenge is a balance between usefulness and variety so as not to pick instances that are too difficult or irrelevant for the model [22]. The choice of which query strategy to use depends on the specific task and the characteristics of the data.

Active learning approach has been widely applied has applied this approach in various domains, including text classification, image recognition, and data annotation tasks, where there is a large pool of unlabeled data available, and the labeling process is resource-intensive as shown in [6, 21, 23]. Active learning provides a valuable strategy to make efficient use of resources and improve the performance of machine learning models.

3 Methodology

3.1 Datasets

Germany Traffic Sign Detection Benchmark The Germany traffic sign detection benchmark (GTSDB) [14] is an evaluation benchmark specifically designed for researchers interested in computer vision, pattern recognition, and image-based driver assistance systems. The GTSDB provides a comprehensive assessment of single-image detection approaches, allowing researchers to evaluate and compare their algorithms. The dataset consists of high-resolution photos, enabling detailed study and accurate recognition of traffic signs. It contains a total of 900 images, with 600 images designated for training and 300 images for evaluation. The availability of the GTSDB dataset allows researchers to train and test their models on real-world traffic sign images, facilitating the development

and improvement of traffic sign detection algorithms. By providing a standardized benchmark, the GTSDB supports the advancement of image-based driver assistance systems and contributes to enhancing road safety in Germany.

Tsinghua-Tencent 100K The Tsinghua-Tencent 100K (TT100K) dataset is designed to address the challenge of simultaneous traffic sign detection and classification in the real world, specifically focusing on the Chinese context. This dataset is created from 100,000 Tencent Street View panoramas, providing a rich source of images for training and evaluation. Within the TT100K dataset, there are 30,000 instances of traffic signs, offering a diverse set of examples for algorithm development. The dataset includes photos that exhibit significant variations in illuminance and weather conditions, accurately representing the challenging and realistic scenarios encountered in real-world environments. Each traffic sign in the TT100K dataset is annotated with a class label, bounding box, and pixel mask. This level of annotation detail enables comprehensive training and evaluation of models, allowing researchers to develop algorithms that can accurately detect and classify traffic signs. The TT100K dataset, particularly the 2021 edition, contains 17,000 photos, with annotations available for 10,000 of those photographs. This dataset is a valuable resource for researchers working on traffic sign detection algorithms that can handle real-world complications, such as changing illumination and weather conditions, specifically in the Chinese context.

Vietnamese Traffic Sign Detection Recognition 46 The Vietnamese traffic sign detection recognition 46 dataset (VTSDB46) is a dataset specifically designed to address the traffic sign detection needs in Vietnam. The dataset has a size of approximately 90 gigabytes, and it consists of over 95,000 raw and augmented images. These images capture a diverse range of traffic signs commonly found on Vietnamese roadways. The signs have been categorized into 49 distinct classes and annotated using the YOLO format, which is a widely used annotation format for object detection tasks. This format facilitates easy integration with various detection algorithms and frameworks. The images in the VTSDB46 dataset are obtained from real-life driving situations using a full HD camera, ensuring that the captured traffic signs contain detailed information. This enables the constructed models to detect traffic signs in real-world scenarios, accounting for factors such as varying illumination, atmospheric conditions, and partially obstructed signs. The VTSDB46 contributes to the advancement of safer and more efficient autonomous driving technologies in Vietnam, ultimately improving road safety and transportation efficiency.

3.2 Active Learning for Traffic Sign Recognition

Our active learning system, depicted in Figure 1, comprises two phases. In the first phase, an object detection model is trained using a small set of predefined and labeled datasets which is manually created by humans. Once the training

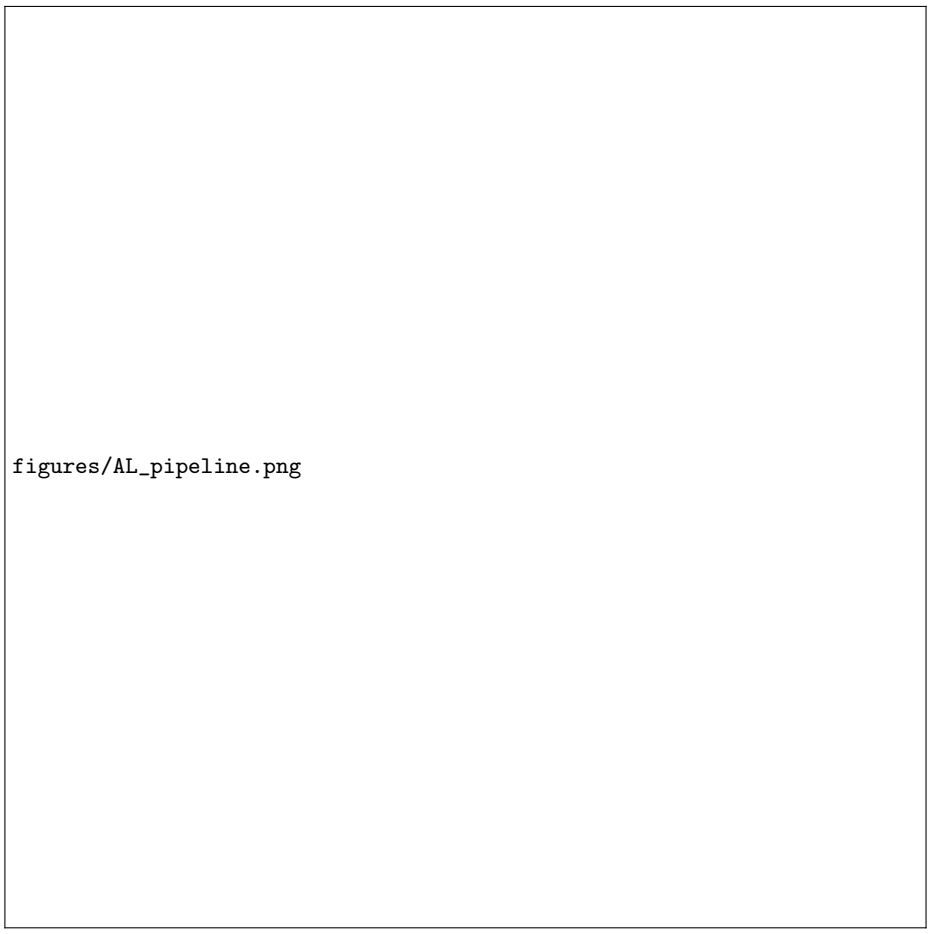


Fig. 1: System pipeline of the Active Learning for Traffic system.

process is complete after several iterations, the model's weight will be utilized in the second phase of active learning. As a progress, a small dataset continues to be added more samples to this dataset from the unlabeled pool, labeling them based on predefined criteria using both human and machine labeling.

In the second phase, a random subset of samples is selected from the unlabeled pool. These samples are then processed through both the training model and a supervisor model to obtain predictions for bounding box positions and class confidences. The supervisor model is a pre-trained model in the same domain as the training model. However, it only focuses on determining the number of objects present in a sample, without predicting the specific types of objects. This can be considered as the "support ground truth" for the task. In the context of traffic sign detection, the supervisor model identifies the number of traffic signs present in a sample, without specifying the exact types of traffic signs. The combined count of bounding boxes from the supervisor model and the training model is used to evaluate the box criterion. This information helps the training model assess whether it has accurately predicted enough objects in its predictions. In this study, a soft criterion is applied to the number of bounding boxes, which means that the human labeling process is only required if the training model predicts fewer bounding boxes than the supervisor model. The math formula of the bounding boxes label condition ($\mathcal{C}_{bbox}$) can be expressed as follows:

$$\mathcal{C}_{bbox} = \begin{cases} True, & n_{model} >= n_{supervisor} \\ False, & \text{otherwise} \end{cases}, \tag{1}$$

where n_{model} is the number of bounding boxes predicted by the training model and $n_{supervisor}$ is the number of bounding boxes predicted by the supervisor model.

Furthermore, the class confidence is utilized to evaluate the confidence criterion using a configurable threshold. If the condition stated in Equation 1 is satisfied, the system proceeds to evaluate the confidence scores of all the predicted bounding boxes. For each bounding box, its confidence score must exceed or equal the predefined threshold. If any bounding box fails to meet this criterion, the sample requires human labeling. The class confidence label condition ($\mathcal{C}_{conf}$) formula for determining the labeling process can be defined as follows:

$$\mathcal{C}_{conf} = \begin{cases} True, & \sum_{i=0}^{n} c_i >= n \\ False, & \text{otherwise} \end{cases}, where\ c_i = \begin{cases} 1, & c_i^{conf} > t \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

where n represents the number of the bounding box in the model prediction, c_i is an indicator variable, where $c_i = 1$ indicates that the bounding box at the i^{th} object meets the criteria, and $c_i = 0$ indicates that it does not meet, c_i^{conf} is the confidence score of bounding box at the i^{th} bounding boxes and t is the predefined threshold for accepted the bounding which has the value in range $[0, 1]$. If both Equations 1 and 2 are satisfied, the prediction of the training model will be considered as the ground truth for the next training round. This means

that the model's predictions are accurate enough to be used as reliable labels for further training. However, if either Equations 1 and 2 do not meet the criterion, it indicates that the model's predictions are not sufficiently accurate. In such cases, human labeling becomes necessary to manually update the ground truth for the sample. This ensures that the training data is corrected and improved with minimal of error.

4 Experiments and Discussion

4.1 Evaluation metrics

Intersection over Union (IoU) is a critical evaluation metric in object detection tasks. It is calculated by comparing the overlap between the predicted bounding boxes ($B1$) and the ground truth boxes ($B2$), as expressed in the equation 3:

$$\mathrm{IoU}(B1, B2) = \frac{|B1 \cap B2|}{|B1 \cup B2|}. \tag{3}$$

It quantifies localization accuracy by computing the ratio of the intersection area to the union area of these bounding boxes. A higher IoU indicates a greater spatial alignment between the predicted and ground true bounding boxes. The model's predictions are compared to the actual objects in the scene using IoU values, which range from 0 (no overlap) to 1 (perfect overlap).

In this study, the performance of the object detection model is assessed using the mean average precision (mAP) metric, which determines the true positive and false positive classifications based on the IoU criterion. The mAP metric is widely embraced in the computer vision research community to evaluate object recognition and segmentation systems. It is commonly used in benchmark challenges such as Pascal VOC and COCO. The formula of AP and mAP can be defined as follows:

$$\mathrm{AP} = \int_0^1 P(\tau)\mathrm{d}\tau, \tag{4}$$

where τ is the decision threshold and P(t) is the value of precision at the threshold τ.

$$mAP = \frac{1}{N}\sum_{n=1}^{N} AP_n, \tag{5}$$

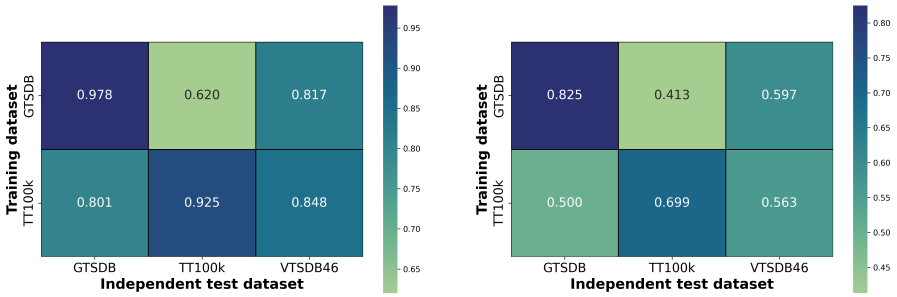
where N represents the number of classes in the problem and AP_n represents the average value of the precision values corresponding to each class.

Additionally, two commonly used variations of mAP are mAP@50 and mAP@50:95. mAP@50 calculates the mAP by considering model's predictions with an IoU threshold of 0.5, indicating the model's ability to identify objects with moderate overlap compared to ground truth boxes. On the other hand, mAP@50:95 computes the mAP across a range of IoU thresholds from 0.5 to 0.95 with a step size

of 0.05. This metric provides a comprehensive assessment of the model's performance across different levels of overlap between predicted and true bounding boxes.

4.2 Traffic Sign Detection

To find the best model for supervising the active learning process in traffic sign detection, we used two datasets GTSDB [14] and TT100k [24] for training the supervisor model. The traffic sign labels from both datasets were combined into a single label representing any traffic sign. This allowed the model to focus on detecting traffic signs without recognizing their specific types. In the experiments conducted by [16], it was found that YOLOv8 achieved the highest score and speed. Therefore, we have chosen YOLOv8 for training our supervisor model and conducting our own experiments. The supervisor model was trained using YOLOv8 Nano, with 100 training epochs, a batch size of 128, and an image size of 640×640. The training process utilized 2 T4 GPUs. After training, we evaluated the two models using the validation datasets from GTSDB, TT100k, and VTSDB46 [16]. The evaluation results, displayed in Figure 2, include the mAP at the IoU thresholds of 0.5 and 0.5 to 0.9.



(a) mAP@50 (b) mAP@50:95

Fig. 2: Performance Comparison of YOLOv8 Nano on Different Datasets for Single Class Object Detection

Since the VTSDB46 dataset will be used for the active learning phase, we selected the model that performed best on its validation dataset. As shown in Figure 2, the model performs best when evaluated on the same dataset it was trained on. However, the GTSDB model's performance on other datasets is lower, making it unsuitable as a supervisor model. This is because the traffic signs in the GTSDB dataset are clear, easy to locate and detect, and the dataset itself is small, lacking enough diverse examples. On the other hand, the TT100k dataset is more diverse than GTSDB, resulting in better performance on different datasets. When evaluating the model trained on the TT100k dataset using the

VTSDB46 validation dataset, it achieves the highest performance. Therefore, the TT100k-trained model is the most suitable choice as the supervisor model for the active learning process.

4.3 Active Learning for Traffic Sign Recognition

Table 1: Performance comparison of different active learning settings on the VTSDB46 dataset

Method	Initialization samples	Samples per round	Total round	Train		Test	
				mAP@50	mAP@50:95	mAP@50	mAP@50:95
Human annotation	87,545	0	1	0.995	0.935	0.994	0.931
Human and machine annotation	10,908	5,000	17	0.962	0.844	0.962	0.845
Human and machine annotation	10,908	10,000	9	0.958	0.837	0.958	0.837
Human and machine annotation	10,908	15,000	7	0.958	0.837	0.958	0.838
Human and machine annotation	10,908	20,000	5	0.953	0.829	0.953	0.829

The active learning training process starts with a small dataset containing 10,908 labeled samples. To ensure a balanced representation, the dataset is carefully sampled so that each class appears at least 100 times. After each round of training, a certain number of unlabeled samples are randomly selected and added to the dataset. In this experiment, we conducted four different settings, where the number of unlabeled samples ranged from 5,000 to 20,000, with an interval of 5,000. Using a higher number of samples per round in the labeling process means that fewer labeling rounds are required to label all the samples in the unlabeled pool. The performance of the model trained with each setting was compared to the performance of a model trained with a fully manually labeled dataset, which is considered as the baseline case. The baseline model achieved the highest mAP@50 and mAP@50:95, both on the training set, validation set, and testing set. Table 1 presents the performance comparison between these experiments and the baseline model trained with the fully labeled dataset.

According to the table, active learning can potentially have a negative impact on the model's performance, specifically in terms of mAP@50:95. However, all the models achieved a mAP@50 score above 0.950, which is suitable for real-life applications. The model that used 5,000 samples per round achieved the highest mAP@50 value of 0.962 on both the training and testing datasets, representing a decrease of approximately 3.3% compared to the baseline model. Similarly, the other experiments also showed lower performance compared to the baseline, but they required fewer labeling rounds to label all the unlabeled samples. Notably, the approach using 10,000 samples per round only needed 9 rounds to label all the unlabeled samples, significantly fewer than the approach using 5,000 samples per round. This number could be further reduced to 5 rounds when using 20,000 samples per round to label the entire dataset. The experiments demonstrated that using a smaller number of samples per round generally resulted in better performance. However, t this also requires more training time since more rounds are needed to label all the unlabeled samples. This trade-off between performance

and training time provides users with the flexibility to choose between accuracy and the cost of labeling the dataset.

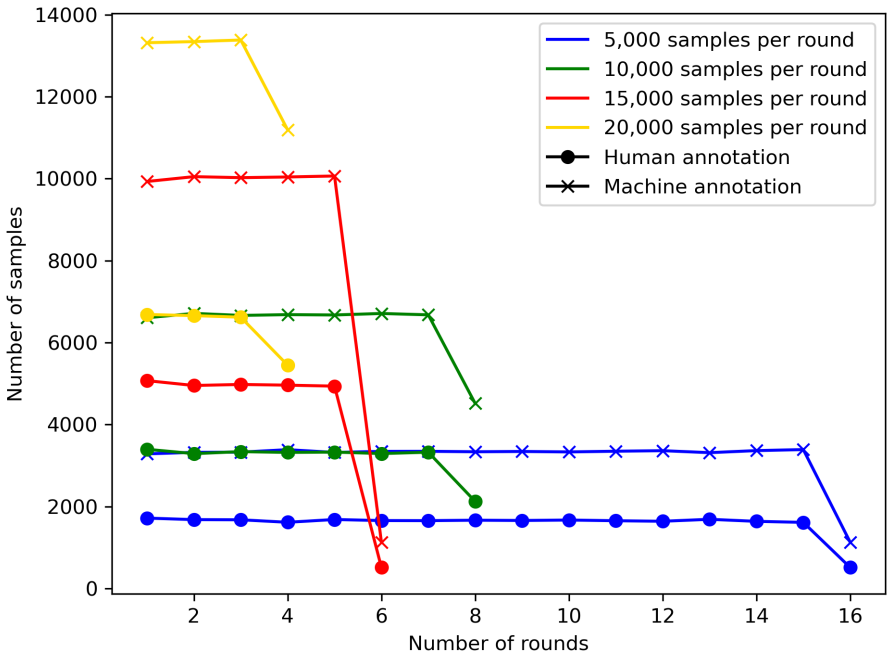


Fig. 3: Comparison of Samples Labeled by Humans and Machines

On the other hand, active learning reduces the amount of work that requires human labeling, as shown in Figure 3. When using 5,000 samples per round, the machine assists humans in labeling approximately 76% of the samples, leaving only 24% for humans to label in each round. This ratio remains consistent throughout all rounds and experiments. This reduction in human labeling workload is achieved by setting a fixed threshold $t = 0.8$ in Equation 2. This threshold serves as a clear guideline for the model to determine when to rely on machine annotation versus human annotation. However, this approach may result in additional work for humans because the model can become more accurate over time and classify certain samples with a lower threshold. By implementing active learning in conjunction with a supervisor model, the amount of human effort required for labeling datasets can be reduced by up to two-thirds, while still maintaining high performance compared to using a fully labeled dataset by humans.

5 Conclusion

In conclusion, our study introduces an active learning system for traffic sign recognition in object detection. An end-to-end system is developed that incorporates transfer active learning and supervisor model which supports human data labeling. Implementing this system in the future will reduce the amount of human work in the traffic sign recognition problem, which typically involves a large amount of data and an expensive labeling process. Our active learning system capitalizes on a pre-trained model acting as a supervisor to automatically verify machine-generated labels, thereby supporting the labeling process. The experiments conducted with various settings demonstrate the high performance of this approach, which achieves performance close to the traditional approach of using a fully manually labeled dataset for training the model. Starting with a small labeled dataset, the model progressively assists humans in labeling new data, making the most of the available data to reduce human workload by up to two-thirds of the time.

In future work, we aim to enhance our system by incorporating additional criteria related to the decision-making process of machine annotation. Furthermore, we will conduct experiments with various settings of active learning methods to determine optimal values and further improve the system's performance. The other aspect will be also consider for improving the performance of our system when apply in real life application.

References

1. Jingyuan Zhao, Wenyi Zhao, Bo Deng, Zhenghong Wang, Feng Zhang, Wenxiang Zheng, Wanke Cao, Jinrui Nan, Yubo Lian, and Andrew F. Burke. Autonomous driving system: A comprehensive survey. *Expert Systems with Applications*, 242:122836, 2024. 2
2. Vijay John, Ali Boyali, Simon Thompson, Annamalai Lakshmanan, and Seiichi Mita. Visible and thermal camera-based jaywalking estimation using a hierarchical deep learning framework. In *Proceedings of the Asian Conference on Computer Vision (ACCV) Workshops*, November 2020. 2
3. Khaldaa Alawaji, Ramdane Hedjar, and Mansour Zuair. Traffic sign recognition using multi-task deep learning for self-driving vehicles. *Sensors*, 24(11):3282, 2024. 2, 4
4. Rajesh Kannan Megalingam, Kondareddy Thanigundala, Sreevatsava Reddy Musani, Hemanth Nidamanuru, and Lokesh Gadde. Indian traffic sign detection and recognition using deep learning. *International Journal of Transportation Science and Technology*, 12(3):683–699, 2023. 2
5. Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Comput. Surv.*, 54(9), oct 2021. 2, 4, 5
6. Amílcar Soares Júnior, Chiara Renso, and Stan Matwin. Analytic: An active learning system for trajectory classification. *IEEE computer graphics and applications*, 37(5):28–39, 2017. 2, 5

7. Tim Fingscheidt, Hanno Gottschalk, and Sebastian Houben. *Deep neural networks and data for automated driving: Robustness, uncertainty quantification, and insights towards safety*. Springer Nature, 2022. 2, 5
8. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. 3
9. Rejin Varghese and Sambath M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6, 2024. 3
10. Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018. 3
11. Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. 3
12. Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. *SSD: Single Shot MultiBox Detector*, page 21–37. Springer International Publishing, 2016. 3
13. A. de la Escalera, L.E. Moreno, M.A. Salichs, and J.M. Armingol. Road traffic sign detection and classification. *IEEE Transactions on Industrial Electronics*, 44(6):848–859, 1997. 4
14. Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE, 2011. 4, 5, 10
15. J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012. Selected Papers from IJCNN 2011. 4
16. Nguyen Dinh Thuan, Phan Minh Khanh, Tran Phuong-Nam, and Dang Ngoc Minh Duc. Vietnamese traffic sign recognition using deep learning. In *Proceedings of the 2024 International Conference on Intelligent Information Technology*, New York, NY, USA, 2024. Association for Computing Machinery. 4, 10
17. Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, and Xinjing Cheng. Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8068–8078, June 2022. 4
18. Yue Zhao, Ciwen Xu, and Yongcun Cao. Research on query-by-committee method of active learning and application. In Xue Li, Osmar R. Zaïane, and Zhanhuai Li, editors, *Advanced Data Mining and Applications*, pages 985–991, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. 5
19. Wenbin Cai, Muhan Zhang, and Ya Zhang. Active learning for ranking with sample density. *Information Retrieval Journal*, 18:123–144, 2015. 5
20. Jürgen Bernard, Marco Hutter, Matthias Zeppelzauer, Dieter Fellner, and Michael Sedlmair. Comparing visual-interactive labeling with active learning: An experimental study. *IEEE transactions on visualization and computer graphics*, 24(1):298–308, 2017. 5
21. Seho Kee, Enrique del Castillo, and George Runger. Query-by-committee improvement with diversity and density in batch active learning. *Information Sciences*, 454-455:401–418, 2018. 5
22. Deniu He. Active learning for ordinal classification based on adaptive diversity-based uncertainty sampling. *IEEE Access*, 11:16396–16410, 2023. 5

23. Elmar Haussmann, Michele Fenzi, Kashyap Chitta, Jan Ivanecky, Hanson Xu, Donna Roy, Akshita Mittel, Nicolas Koumchatzky, Clement Farabet, and Jose M Alvarez. Scalable active learning for object detection. In *2020 IEEE intelligent vehicles symposium (iv)*, pages 1430–1435. IEEE, 2020. 5
24. Zhe Zhu, Dun Liang, Songhai Zhang, Xiaolei Huang, Baoli Li, and Shimin Hu. Traffic-sign detection and classification in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 10