HAWKEYE

Yiqi Huang, Jiacheng Li, Xin Li, Xiang Zhao, Zhen Zhang

National University of Singapore, Singapore 119615

ABSTRACT

The escalating necessity for safe and efficient transportation systems for cyclists necessitates the development of innovative solutions. This study presents an Augmented Reality (AR) eyewear, a revolutionary system designed specifically to enhance cyclists' safety and navigation capabilities on the road. Unlike conventional cycling aids, which often lack real-time responsiveness and user-friendly interfaces, the AR eyewear integrates advanced detection and assistance technologies into a single, easy-to-use platform for cyclists, bikers, and personal mobility device users.

Specifically, we utilize an RGB-D camera for multi-object and multi-view detection, which provides critical information from the cyclists' blind spots and aids in real-time lane track segmentation. This solution also incorporates robust detection capabilities under varying light, weather, and traffic conditions.

Our tested various training strategy to ensure the efficiency and reliability of the proposed system. Initial results reveal substantial enhancements in situational awareness and decision-making in real-time traffic environments, thereby establishing the AR eyewear as a promising solution for safer and smarter cycling experiences.

Keywords: Object detection, Lane segmentation, Computer Vision, Helmet

Code:https://github.com/ValerioL29/Hawkeye

1. INTRODUCTION

Cycling in urban areas, while increasingly popular due to growing awareness of low-carbon travel, often poses significant safety risks. According to the Singapore Traffic Police, 23% of cycling accidents were caused by rear-end collisions, and 17% resulted from blind-spot incidents. This data illuminates the urgent necessity to address the safety issue of cyclists, especially in high-traffic areas.

This paper introduces an Augmented Reality (AR) eyewear, a groundbreaking system designed to enhance cyclists' safety and navigation capabilities while promoting sustainable transportation. This AR eyewear is designed to address the dire need for an intelligent, safety-enhancing solution that enhances cyclists' awareness and decision-making capabilities while minimizing potential road hazards.

The complex cycling conditions present various detection challenges ranging from object differentiation, real-time detection effectiveness, vibration, and limited computational capability. To this end, our approach integrates multiple sensors and advanced algorithms including environmental understanding, robust blind spot detection, drivable area keeping under different conditions. Specifically, We employ an RGB-D camera for multi-object detection and tracking, which providing vital information for real-time multi-object detection and lane keeping support. Our technology is designed to address common issues such as rear-end and blind spot collisions - the primary causes of cycling accidents.

We believe our system marks a revolutionary step towards transforming cyclists' safety and navigation experiences on the road. By integrating advanced detection technologies, our solution aims to significantly reduce the risks associated with cycling in traffic-heavy zones and improve overall public health and well-being.

2. LITERATURE REVIEW

2.1. Object Detection

Object detection is a critical task in computer vision, aimed at identifying and localizing objects within images or video frames. For our project, which focuses on enhancing cyclist safety using augmented reality eyewear, robust and real-time object detection is essential.

One of the most significant advancements in object detection is the development of the You Only Look Once (YOLO) framework by [1]. YOLO revolutionized object detection by framing it as a single regression problem, directly predicting bounding boxes and class probabilities from full images in one evaluation. This approach contrasts with earlier methods that required multiple stages of processing, such as R-CNN [2], Fast R-CNN[3], and Faster R-CNN[4].

Subsequent iterations of YOLO, including YOLOv3[5], and the latest YOLOv8, have continuously improved detection speed and accuracy. YOLOv8, in particular, offers a balance between speed and precision, making it ideal for real-time applications like ours, where detecting vehicles, pedestrians, and potential obstacles quickly is crucial for cyclist safety.

For our project, we fine-tune YOLOv8 based on the pro-

cessed BDD100K dataset[6], which provides a diverse set of driving videos covering various environmental and lighting conditions. This ensures that our object detection model is robust and performs well across different scenarios encountered by cyclists.

2.2. Multi-Object Tracking

Multi-object tracking (MOT) involves following multiple objects over time in video sequences, a crucial capability for monitoring dynamic environments such as traffic. Effective tracking ensures that once objects are detected, their movements can be consistently followed, providing continuous situational awareness.

Traditional tracking methods relied on techniques like the Kanade-Lucas-Tomasi (KLT) tracker and Discriminative Correlation Filter (DCF)[7]. However, the integration of deep learning with tracking has significantly enhanced performance. The Simple Online and Realtime Tracking (SORT) algorithm and its deep learning-enhanced version, DeepSORT, exemplify this integration, combining detection and tracking into a unified framework ([8]; [9])

In our project, we utilize YOLOv8 not only for object detection but also for multi-object tracking. This unified approach simplifies our workflow and enhances performance. Specifically, we fine-tune YOLOv8 for detection and employ advanced tracking algorithms to maintain the identities of multiple objects across video frames. This ensures robust tracking of vehicles and pedestrians, enhancing the safety and awareness features of our augmented reality eyewear.

2.3. Lane segmentation

Lane segmentation and drivable area prediction are crucial tasks in autonomous driving and advanced driver-assistance systems (ADAS). These tasks involve identifying and delineating lanes on the road and predicting the regions that are safe to drive on, respectively. For our project, enhancing cyclist safety with augmented reality eyewear, accurate and real-time lane segmentation and drivable area prediction are essential.

Lane segmentation involves detecting lane markings on the road to assist in vehicle positioning and navigation. We utilize the YOLOv8 model, fine-tuning it on a labeled dataset specifically for lane segmentation. The training data includes diverse road conditions, varying lane markings, and different lighting environments to ensure robustness and accuracy. We use the TuSimple dataset, which provides a comprehensive set of annotated lane images. The dataset is pre-processed to standardize image sizes and enhance lane markings using image augmentation techniques. YOLOv8 is fine-tuned using the pre-processed TuSimple dataset. The model's architecture is adapted to predict lane boundaries, utilizing its convolutional layers to capture intricate details of lane markings. During inference, the YOLOv8 model outputs lane boundary predictions. Post-processing techniques such as curve fitting and smoothing are applied to refine these predictions, ensuring continuous and accurate lane detection.

2.4. Drivable segmentation

Drivable area prediction identifies regions of the road that are safe for driving, which is critical for navigation and avoiding obstacles. Similar to lane segmentation, YOLOv8 is fine-tuned for this task, leveraging its fast and accurate detection capabilities. We use the BDD100K dataset, which includes annotated drivable areas under various environmental conditions. The dataset is curated to cover a wide range of driving scenarios, ensuring that the model performs well in different contexts. The YOLOv8 model is fine-tuned on the annotated BDD100K dataset. The model is trained to identify drivable areas, adapting its prediction layers to output segmentation maps that delineate safe driving zones. During inference, YOLOv8 generates segmentation maps indicating drivable areas. Post-processing techniques such as morphological operations and smoothing are applied to enhance the clarity and accuracy of these predictions. For our augmented reality eyewear, it is crucial that lane segmentation and drivable area prediction operate in real-time. The optimized YOLOv8 model ensures low latency and high accuracy, making it suitable for real-time applications. The integration of these tasks within a single YOLOv8 framework simplifies the workflow and improves overall system performance. By fine-tuning YOLOv8 for lane segmentation and drivable area prediction, we enhance the safety features of our augmented reality eyewear, providing cyclists with real-time, accurate information about their environment. This ensures better situational awareness and safer navigation.

3. PROPOSED APPROACH

3.1. Object detection

The first component of the system is the object detection model, which identifies and localizes objects such as vehicles, pedestrians, and obstacles in real-time, crucial for enhancing cyclist safety. We use YOLOv8 as the backbone for our object detection due to its efficiency and high accuracy. The BDD100K dataset is pre-processed for training, providing a diverse range of driving scenarios under various environmental and lighting conditions. During the training process, the model is initialized with pre-trained weights and fine-tuned on our specific dataset. The loss function used combines classification loss, localization loss, and confidence loss to ensure accurate object classification and localization. Evaluation metrics such as precision, recall, and mean Average Precision (mAP) are employed to assess the model's performance and make necessary adjustments.

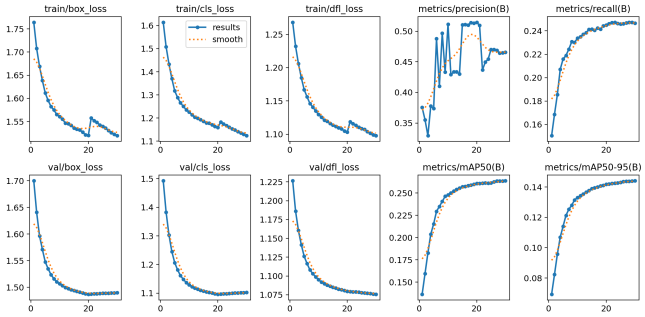

Fig. 1. Training Loss over Epochs

3.2. Multi-Object Tracking

Multi-object tracking model maintains the identities of detected objects across video frames, which is essential for continuous monitoring and situational awareness. We integrate the detections from YOLOv8 with BoT-SORT, a state-of-the-art tracking algorithm that combines appearance features and motion information for robust tracking.

The tracking process begins with the detections backbone from YOLOv8, where each object is assigned a unique identifier (ID). BoT-SORT uses a Kalman filter to predict the position of each object in the next frame based on its current state (position and velocity). The algorithm extracts appearance features for each detected object using a convolutional neural network (CNN) and employs the Hungarian algorithm to associate detected objects in the current frame with tracked objects from the previous frame. This association metric combines appearance similarity and spatial proximity to find the best matches, ensuring continuous tracking even when objects are temporarily occluded. BoT-SORT also incorporates re-identification (ReID) models to handle long-term occlusions and recover identities when objects reappear after being hidden. This improves the robustness and accuracy of tracking in complex environments.

By combining YOLOv8 for robust object detection and BoT-SORT for accurate multi-object tracking, our system ensures reliable real-time detection and tracking of vehicles, pedestrians, and other potential hazards. Fine-tuning the models on the BDD100K dataset ensures robustness across diverse environmental conditions. This integrated approach enhances the safety features of our augmented reality eyewear for cyclists, providing real-time awareness and improving situational awareness, which is crucial for preventing accidents and ensuring cyclist safety.

4. EXPERIMENTAL RESULTS

4.1. Implementation details

Object detection We use YOLOv8 as the backbone for our object detection model due to its high efficiency and accuracy. YOLOv8 processes entire images in a single pass, predicting

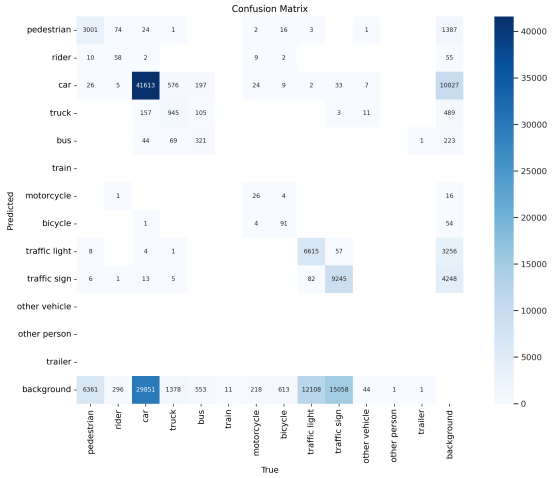

Fig. 2. Confusion matrix for different classes object

bounding boxes and class probabilities simultaneously, which is crucial for real-time applications. The training process begins by initializing YOLOv8 with pre-trained weights from the Coco dataset. These weights serve as a strong foundation, leveraging extensive training on a large and diverse dataset. We then fine-tune the model on our pre-processed BDD100K dataset, which is specifically tailored to our use case involving varied driving scenarios and environmental conditions.

To optimize the fine-tuning process, we experimented with different strategies. These include: (1) full training of all layers, where every parameter in the network is updated during training; (2) fine-tuning only the object detection head, which involves updating the parameters of the final layers responsible for the detection tasks while keeping the earlier layers fixed; (3) fine-tuning only the last layer of the object detection head with the parameters of all preceding layers frozen; and (4) multi-task training with other tasks share the same backbone.These approaches allow us to identify the most effective strategy for transferring learned features to our specific task while maintaining computational efficiency.

We also conducted experiments to determine the optimal number of training epochs, ensuring the model achieves the best performance without overfitting. The loss function used in training combines classification loss, localization loss, and confidence loss. This comprehensive loss function ensures that the model not only accurately classifies objects but also precisely localizes them within the image, and remains confident in its predictions.

Multi-Object Tracking We use BoT-SORT as the tracker for our multi-object tracking due to its robustness and accuracy. The process begins with the object detections from YOLOv8, which include bounding boxes and class probabilities for each detected object. The BoT-SORT tracker initializes with these detections, assigning unique identifiers (IDs) to maintain object identities across frames. We use a Kalman filter to predict the positions of these tracked objects in the

next frame, considering their velocities and previous states. Appearance features for each detected object are extracted using a convolutional neural network (CNN), which helps in accurate data association. We employ the Hungarian algorithm to match detections in the current frame with existing tracks by combining appearance similarity and motion prediction. This ensures that each detected object is correctly matched with its corresponding track. BoT-SORT handles occlusions by using the predicted positions from the Kalman filter and reidentifies objects when they reappear. The final output consists of tracked objects with their IDs, bounding boxes, and class probabilities.

Test Time Augmentation Test Time Augmentation improves model accuracy by applying transformations to input images during inference. In our implementation, TTA involves:

- Left-Right Flipping: Each image is horizontally flipped.
- Multi-Resolution Processing: Images are processed at three different resolutions (e.g., 640, 832, and an intermediate size). The model generates predictions for each augmented version and merge them using Non-Maximum Suppression (NMS) for the final output.

Enabling TTA increases inference time by 2-3 timesdue to larger image sizes (832 vs. Original 640 pixels) and additional processing steps. Despite the longer inference time, TTA significantly enhances prediction accuracy and robustness during real-world demo testing, so we still involve it and plan to optimize in the future.

4.2. Dataset

Table 1. Dataset Overview

Type	Train	Validation
Image (1280x720)	70,000	30,000
Object Detection	69,853	10,000
Drivable Segmentation	66,921	9,546
Lane Segmentation	66,640	9,526

The dataset we used in this project is BDD100K (Berkeley DeepDrive 100,000), which is a comprehensive and diverse collection of driving videos and images that is widely used in the fields of computer vision and autonomous driving research. It comprises 100,000 video clips and images, capturing a broad array of driving scenarios across different weather conditions, times of day, and locations, thus providing a rich and varied dataset for researchers.

One of the standout features of BDD100K is its extensive annotations. These include bounding boxes for object detection, pixel-wise annotations for semantic segmentation, lane marking annotations, and drivable area annotations. Additionally, the dataset includes tracking information, with objects annotated with consistent IDs across frames, facilitating object tracking tasks. This level of detailed annotation supports a wide range of applications, from autonomous driving and advanced driver assistance systems (ADAS) to general computer vision research.

For our Hawkeye project, which aims to enhance cyclist safety through functions like blind spot detection, multi-object detection, and lane area segmentation, BDD100K offers an excellent foundation. The rich annotations and diverse driving scenarios in the dataset can help in training and testing the algorithms required for these safety functions. Furthermore, the dataset's focus on safety-critical elements aligns well with the goals of our project, making it a particularly suitable choice for our research and development efforts.

4.3. Performance metrics

For evaluation of object detection module, the fine-tuned model is rigorously assessed on a validation set derived from the BDD100K dataset. We utilize several metrics to measure the model's performance, including precision, recall, and mean Average Precision (mAP). Precision measures the accuracy of the detected objects, recall assesses the ability of the model to identify all relevant objects, and mAP provides a comprehensive evaluation of both precision and recall across different threshold levels. These metrics collectively help in understanding the overall performance of the model and guide further refinements to enhance its accuracy and robustness in real-world scenarios.

4.4. Experimental results

Object detection We trained the entire detection head of the YOLOv8 model over multiple epochs to assess its performance on the BDD100K dataset. The loss function combined classification loss, localization loss, and confidence loss. The training loss steadily decreased over the epochs, indicating effective learning and optimization. The plot of training loss against the number of epochs is shown in Figure 1. As the training loss presents, the performance improves steadily and achieved its optimum at the 18th epoch. The overall mAP is 23%, which is almost 10 times the zero-shot benchmark. To evaluate the model's performance in distinguishing between different classes, we analyze the confusion matrix based on the validation set of the best performance. The confusion matrix, presented in Figure **??**, illustrates the true positive rates and highlights areas where the model made incorrect predictions. As shown in the confusion matrix, different classes of object have distinctive performances, with high precision on objects such as cars and pedestrians and low precision on bicycles and trains. The main results may lie in that the training

Table 2. Object Detection Performance with Different Fine-Tuning Strategies.

Ex.No	Epochs	mAP50	MAP50-95	Precision	Recall
Zero-shot	-	2.7	0.52	3.4	3
Freeze head	1	4	1.7	14.5	5.16
Freeze head with 40 epochs	Best 19th epoch	23	12.5	65.7	22.6
Freeze backbone	1	11	5.8	26.4	13.1
Freeze backbone with 40 epochs	Best 18th epoch	23	12	65.8	22.5

Table 3. Lane Segmentation Performance with Different Fine-Tuning Strategies

Ex.No	Epochs	mAP50	MAP50-95	Precision	Recall
Zero-shot	-	0.00042	0.00014	0.00078	0.0356
Freeze head	1	0.02977	3E-05	0.21	0.05215
Freeze head with 40 epochs	40	0.06243	0.02824	0.24853	0.09381
Freeze backbone	1	0.11062	0.05982	0.02579	0.02313
Freeze backbone with 40 epochs	40	0.24188	0.15668	0.23857	0.24123

data has very unbalanced distribution, some classes have very limited amount of samples.

Lane Segmentation In the zero-shot experiment, the model showed minimal performance with an mAP50 of 0.00042 and mAP50-95 of 0.00014. When the head was frozen and trained for 1 epoch, the mAP50 improved to 0.02977, while mAP50-95 remained low at 3E-05. Training the frozen head for 40 epochs further enhanced the mAP50 to 0.06243 and mAP50-95 to 0.02824, indicating better detection performance.

Freezing the backbone and training for 1 epoch resulted in a significant improvement, with an mAP50 of 0.11062 and mAP50-95 of 0.05982. Finally, freezing the backbone and training for 40 epochs achieved the highest performance, with an mAP50 of 0.24188 and mAP50-95 of 0.15668. Precision and recall metrics also showed considerable improvements with increased training epochs.

Drivable Prediction In the zero-shot experiment, the model performed poorly with a negative mAP50 of -0.24377 and mAP50-95 of 0.08043. Freezing the head and training for 1 epoch significantly improved the performance, achieving an mAP50 of 0.3683 and mAP50-95 of 0.13298. However, training the frozen head for 40 epochs resulted in an mAP50 of 0.06243 and mAP50-95 of 0.02824, showing moderate improvements.

Freezing the backbone and training for 1 epoch showed an mAP50 of 0.23646 and mAP50-95 of 0.07031. The best performance was achieved by freezing the backbone and training for 40 epochs, with an mAP50 of 0.08068 and mAP50-95 of 0.61002. Precision and recall metrics also showed substantial improvements, with precision reaching 0.80406 and recall 0.78299.

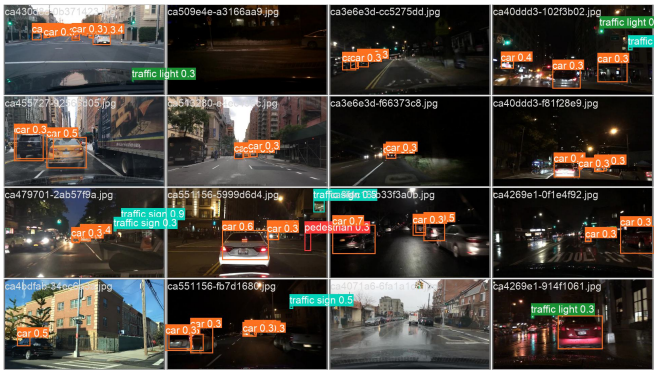

Fig. 3. Train detection head with 1 epochs

4.5. Ablation study

Object detection For object detection based on YOLOv8, we explored various fine-tuning strategies to optimize model performance. Our training experiment results are shown in 4. We utilize zero-shot evaluation as a baseline, observing minimal effectiveness. We then proceeded with fine-tuning the model by freezing the detection head and training for one epoch, which showed modest improvements. Extending this approach, we trained the detection head for 40 epochs, achieving significant performance gains. Notably, the best results were obtained around the 19th epoch, demonstrating the value of extended training. Similarly, we experimented with freezing the backbone and training for one epoch, which yielded better results compared to freezing the head alone for a single epoch. Further extending this strategy to 40 epochs, we found that the best performance was achieved around the 18th epoch, indicating that extensive training with a frozen backbone also leads to substantial improvements. Overall, our results highlight the importance of fine-tuning strategies in enhancing the YOLOv8 model's accuracy and precision, with extended training of specific components proving to be

Table 4. Drivable Prediction Performance with Different Fine-Tuning Strategies

Ex.No	Epochs	mAP50	MAP50-95	Precision	Recall
Zero-shot	-	-.24377	0.08043	0.27541	0.28942
Freeze head	1	0.3683	0.13298	0.34091	0.4456
Freeze head with 40 epochs	40	0.06243	0.02824	0.24854	0.09381
Freeze backbone	1	0.23646	0.07031	0.25263	0.35462
Freeze backbone with 40 epochs	40	0.08068	0.61002	0.80406	0.78299

Fig. 4. Train detection head with30 epochs

particularly effective. We also compared detection samples from the model at different training epochs to visually assess the improvement in detection accuracy and localization, which is shown in 3 4. As training progresses, detection performances keeps increasing with many undetected objects being detected successfully. Meanwhile, the performance improvement can be seen in different weather conditions and lighting conditions, which makes our system more robust.

4.6. Discussions and limitations

Despite the promising results, our system encounters several limitations that need addressing in future work:

- First, the availability of around 100K samples constrained our model fine-tuning. During training, we observed that the models tend to overfit after approximately ten epochs, struggling to achieve higher mean Average Precision (mAP) during evaluation.
- Secondly, the system's inference logic follows a naive sequential approach, processing each of the three models individually. This method is less efficient and harms real-time performance, as it does not leverage parallel processing capabilities.
- Even though we applied TTA to mitigate the effects of poor illumination and inclement weather conditions, the model can still become unstable and even fail under such circumstances. It highlights the need for further improvements to ensure consistent performance in adverse conditions.

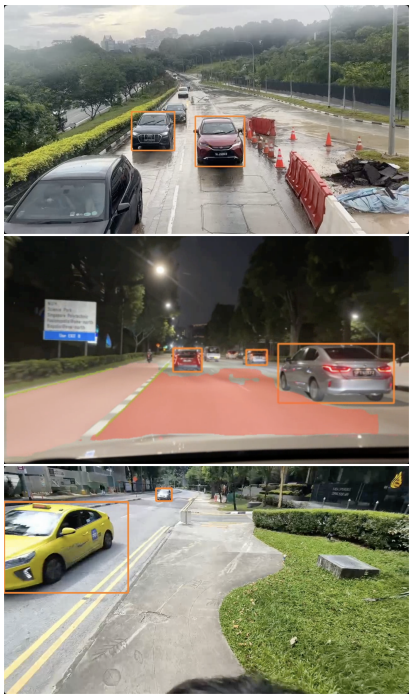

Fig. 5. Demo showcases

5. CONCLUSIONS AND FUTURE WORK

Our AR eyewear represents a significant advancement in enhancing cyclists' safety and navigation capabilities. By integrating multiple sensors and advanced algorithms, we have developed a system that effectively addresses the complex detection challenges faced by cyclists. The use of an RGB-D camera for multi-object detection and tracking, along with robust blind spot detection and drivable area keeping, significantly reduces the risk of rear-end and blind spot collisions, which are primary causes of cycling accidents.

We utilized the robust pre-trained YOLOv8 backbone, which we fine-tuned on the BDD100K dataset for each specific task. By strategically employing transfer learning, we adapted our models to handle the comprehensive panoptic segmentation task, ensuring that our system can flexibly meet a wide range of needs and achieve commendable performance in testing.

Looking ahead, our future work will focus on further refining our algorithms to improve accuracy and responsiveness

under diverse environmental conditions. We plan to explore the integration of additional sensors to enhance object differentiation and detection capabilities. Moreover, we aim to optimize the system's computational efficiency to ensure seamless real-time processing.

To enhance the capabilities and performance of the "Hawkeye" system, we plan the following improvements and developments:

- **Multitask Learning Implementation** We aim to implement a comprehensive multitask learning system. Although we have made significant progress, we face some internal conflicts with the Ultralytics framework. Resolving these conflicts will be a priority to achieve seamless multitask learning, releasing the system's full potential efficiency and accuracy.
- **Parallel Processing for Inference** We will redesign the inference logic to enable parallel processing of frames. By leveraging parallelism, we can significantly reduce latency and improve the system's real-time capabilities.
- **Rich Features Exploration** We will also try to expand the scope of our technology to include more advanced features, such as predictive analytics for collision avoidance and enhanced navigation assistance.

By addressing these limitations and pursuing these future enhancements, we aim to build a more robust and efficient "Hawkeye" system capable of delivering superior performance and reliability in real-world applications. Ultimately, our goal is to continue innovating and pushing the boundaries of AR technology to create a safer and more sustainable future for cyclists.

6. AUTHOR CONTRIBUTIONS

Yiqi Huang
Jiacheng Li
Xin Li
Xiang Zhao
Zhen Zhang

7. REFERENCES

[1] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[2] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[3] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[5] Joseph Redmon and Ali Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[6] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.

[7] David S. Bolme, J. Ross Beveridge, Bruce A. Draper, and Yui Man Lui, "Visual object tracking using adaptive correlation filters," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2544–2550.

[8] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3464–3468.

[9] Nicolai Wojke, Alex Bewley, and Dietrich Paulus, "Simple online and realtime tracking with a deep association metric," 2017.