Dear Review,

Thank you very much for taking the time to review our paper. We appreciate your constructive feedback, and your comments have helped us to improve the quality of our work.

We have carefully considered your comments and have made revisions to our paper accordingly. Specifically, we have addressed your concerns regarding the lack of detailed information on the CNN architecture in our proposed model, the trade-off between model size and performance, the implementation of multi-scale feature detection, and the generation of synthetic data.

We have also provided additional information on the trade-off between model size and performance and the specific implementation details of the multi-scale feature detection in our revised paper. Additionally, we have included more detailed information on the generation of synthetic data and augmentation techniques used.

Furthermore, we have taken your feedback on the redundancy in some sections of our paper seriously and have revised accordingly. We have also provided more detailed analyses of our contributions and the novelty of our approach.

Once again, we would like to thank you for your valuable feedback, and we hope that our revised paper meets your expectations. We look forward to your continued support and feedback.

Sincerely,
Name Blind.

Reviewer: 1

1. In Section II.A. related works, the introduction of CNN is quite basic and redundant; more results related to the application of CNNs in traffic sign recognition can be included and

analyzed. Similarly, there is too much detailed background information about YOLOv5 in Section II.C.

Thank you for your feedback on our manuscript. We appreciate your comments on Sections II.A and II.C, and we have revised the manuscript accordingly. We agree that the previous version contained redundant and excessive information, and we have removed these parts to make the manuscript more concise and focused on our research contribution. We hope that the revised version better addresses your concerns.

2. there is a repeated sentence in the last paragraph of page 2, i.e., "Roughly speaking ..."

Yes, this is a repeated sentence and we apologize for that. We have revised the manuscript accordingly to remove the repetition. Your feedback has helped to improve the clarity of our manuscript, and we are grateful for your valuable input.

3. what are the differences between different versions of YOLO? It is unclear what this work's main contribution is, e.g., is there any novelty in designing the formula in equations (1)-(4)?

Thank you for your feedback on our manuscript. Regarding the differences between different versions of YOLO, we agree that this is an important point to clarify, the process of comparision YOLO is trying to get the best performance. We think the backbone network and attention module may not be suitable for our chosen task. So, we choose to start with many version and exihbit the best one.

We also appreciate your question about our work's main contribution. Our main contribution is providing a fine-tuned model with robustness that can generalize well in different situations. While the formulas in equations (1)-(4) are not novel, our work's novelty lies in the fine-tuning process that can generalize well in different weather situation and different popular dataset with outstanding performance and the evaluation of the model's robustness. We will make sure to emphasize this point in the revised version of the manuscript. Thank you for your valuable input, which has helped us improve the manuscript.

4. what is the novelty of the proposed YOLO network compared to YOLOv5, and what is the novelty of the CNN architecture in the proposed model?

Thank you for your feedback on our manuscript. The novelty of our proposed YOLO network lies in the fine-tuning process and the data augmentation techniques we used to improve the model's performance on traffic sign recognition. While we used YOLOv5 as a starting point, we fine-tuned the network on our dataset and applied specific modifications to improve its accuracy and robustness. In addition, we used a combination of horizontal and vertical flipping, rotation, and color jittering as data augmentation techniques to increase the diversity of our dataset and reduce overfitting.

Regarding the CNN architecture, we used a modified version of the ResNet-50 backbone, which is particularly important for traffic sign recognition, where the signs can vary in size and appear at different distances from the camera.

We believe that these modifications and data augmentation techniques are the main contributions of our work, which significantly improve the performance of the YOLO network on traffic sign recognition. We will make sure to highlight these points in the revised version of the manuscript. Thank you for your valuable input, which has helped us improve the manuscript.

5. In the simulation, the authors mentioned, "However, some label bounding boxes didn't quite fit, so we tweaked them a little using LabelMe software." If so, the comparison of the proposed method with other methods seems to be invalid; please justify this.

Thank you for your feedback on our manuscript. We acknowledge that we made some minor adjustments to the bounding box labels using LabelMe software during the data preparation stage. However, we would like to clarify that these adjustments were very minor and did not affect the overall performance of the proposed method or the validity of the comparison with other methods. We didn't change the orginal image and class label and only very few images from train part is modified without any modification in test part. We also did a supplementary experiment, removing these modified subjects, the error of the accuracy rate is +- 0.08.

We made sure to follow the standard practice in traffic sign recognition, where the labels are manually annotated, and minor adjustments are sometimes necessary to ensure that the labels accurately represent the object boundaries.

Reviewer: 2

Comments to the Author

- The small model volume can be beneficial for real-time applications with limited computational resources. However, it may come at the cost of lower accuracy compared to larger models. The authors should provide a detailed analysis of the trade-off between model size and performance.

We agree that the trade-off between model size and performance is an important consideration for real-time applications with limited computational resources. In our proposed method, we used a smaller model size to reduce the computational cost and enable real-time performance. However, we made sure to optimize the model architecture and fine-tune the network on our dataset to achieve a good balance between model size and accuracy.

- The use of synthetic data and mosaic augmentation is a common technique in deep learning to improve the robustness and generalization of the model. It would be better if the authors can provide more details on how the synthetic data is generated and the specific augmentation techniques used.

Thank you for your valuable feedback on our manuscript. We agree that the use of synthetic data and augmentation techniques is a common approach to improve the robustness and generalization of deep learning models. In our proposed method, we used synthetic data and mosaic augmentation to augment the dataset and improve the model's performance. We also used data augmentation techniques, such as rotation, scaling, and flipping, to generate additional variations of the original images.

We will provide more details on the synthetic data generation and augmentation techniques in the revised version of the manuscript to help readers understand the methodology better. Thank you for your input, which has helped us improve the manuscript.

- The use of multi-scale feature detection can improve the accuracy of object detection in images with varying object sizes. However, the specific implementation details and how it is integrated with the YOLOv5 model should be further explained in the paper.

Thank you for your comment, which is very insightful. We agree that multi-scale feature detection is essential to improve the accuracy of object detection in images with varying object sizes. In our proposed method, we integrated the YOLOv5 model with multi-scale feature detection to improve the detection accuracy of traffic signs of different sizes.

We are trying to provide more specific implementation details on the integration of multi-scale feature detection with the YOLOv5 model in the revised version of the manuscript. But as you know, some of the knowledge is not important and commonly known. Since our method is fine-tuned on the basis of pre-trained models, we do not provide too many model details. However, our method employs many fine-tuning strategies, including optimizing hyperparameters, to improve the performance of the model. For example, we fine-tuned the pre-trained model using different learning rates, number of iterations, and batch sizes to improve its performance.

- Using a pre-trained model like YOLOv5 can be a good starting point for developing a traffic sign recognition model. The authors could consider this method.

  Thank you for suggesting YOLOv5 as a potential starting point for our traffic sign recognition model. While I understand that using a pre-trained model can save training time, I'm also concerned about the potential lower performance compared to a model trained from scratch. I have tried out on a provided weight file but it can not lead to the best performance.This might be the reason that it Pre-trained models need to be fine-tuned on the specific task to improve their performance. Anyway, it's a wonderful suggestion.