Blind Spots in Autonomous Vision: Evaluating the Robustness of YOLOv8 Model in Adverse Weather

Trevor Kwan

Faculty of Science, Capilano University

Instructors: Derek Howell, Eunice Chin

Dec.9, 2025

Abstract

This study investigates the efficiency of the You Only Look Once (YOLO) object detection algorithm for real-time traffic light detection under challenging real-world situations. Specifically, the study aims to quantify the performance and degradation in mean average precision (mAP) and measure the average inference speed (in frames per second) across different adverse weather classifications: rain, snow, and fog, while comparing against a dataset of clear skies weather. By combining large diverse datasets found in publicly available domains, a YOLO variant will be trained and evaluated against another separate dataset of images and videos of traffic lights. The findings will contribute to providing necessary benchmarks for evaluating the reliability in upcoming technological systems such as advanced driving assistance and autonomous vehicle systems.

1. Introduction

The rapid evolution of Autonomous Vehicle (AV) technology promises to fundamentally reshape modern transportation, offering potential solutions to traffic congestion, carbon emissions, and the high incidence of human-error induced accidents. Chougule et al. (2024) state that as the automotive industry transitions to full automation, the reliability of the vehicle's perception system becomes the single most critical determinant of safety. Among the myriad tasks an autonomous agent must perform, accurate and instantaneous traffic light detection is paramount. Unlike static obstacles or lane markings, traffic lights are dynamic regulatory mechanisms that dictate the flow of right-of-way; a failure to strictly adhere to their signals can result in catastrophic, high-speed collisions at intersections. Consequently, the development of robust computer vision systems capable of interpreting these signals with near-perfect accuracy is a non-negotiable requirement for the safe deployment of AVs on public roads.

In recent years, deep learning algorithms, specifically Convolutional Neural Networks (CNNs), have emerged as the standard architecture for visual perception in AVs. Architectures such as *You Only Look Once* (YOLO) have revolutionized real-time object detection by treating detection as a single regression problem, allowing for inference speeds that rival human reaction times. In controlled, "ideal" environments characterized by clear skies, balanced illumination, and standard infrastructure these models have achieved detection rates that approach, and in some metrics exceed, human performance. However, the operational design domain of a real-world vehicle is rarely ideal. It is subject to the chaotic entropy of the open environment, where severe weather (rain, snow, fog), extreme lighting variability (sun glare, urban night-glow), and non-standard traffic infrastructure introduce significant visual noise. The discrepancy between model performance in sterile training environments and chaotic real-world scenarios represents a critical "safety gap" that current research must address.

The progression of traffic light detection technology has been intrinsically linked to advancements in hardware acceleration. The advent of high-performance Graphics Processing Units (GPUs) has fundamentally altered the feasibility of deploying deep neural networks in mobile agents. As noted by Zhang et al. (2010), image processing algorithms are

computationally expensive and highly parallelizable; the GPU's architecture allows for the simultaneous calculation of thousands of pixel matrices, enabling the vehicle to process high-resolution video feeds in milliseconds. This computational throughput supports complex architectures that often combine the spatial pattern recognition of CNNs with the temporal memory of Recurrent Neural Networks (RNNs).

While CNNs are the primary engine for feature extraction, identifying the shapes, colors, and edges that constitute a "traffic light", RNNs provide a necessary temporal context. The state of a traffic light is a temporal phenomenon; a yellow light is not merely a colored bulb but a sequential warning following a green signal and preceding a red one. By retaining information from previous frames, RNN modules allow the AI to maintain "object permanence" and contextual reasoning, smoothing out momentary flickers or occlusions that might otherwise confuse a single-frame detector.

Despite these architectural advancements, a profound epistemological challenge remains: the "Black Box" nature of deep learning. As highlighted by Wang et al. (2020) and Park & Yang (2019), CNNs function as opaque non-linear approximations. We can observe the input (pixel data) and the output (classification), but the internal logic, the millions of weight adjustments that lead the model to label a cluster of pixels as a "Red Light", remains largely inaccessible. This lack of transparency poses a significant hurdle for safety verification. Unlike white-box algorithms, where the decision tree is transparent and auditable, deep learning models rarely offer post-hoc explanations for their failures (Li et al., 2022). If a YOLO model fails to detect a red light in a snowstorm, it cannot tell engineers the reason it failed. Whether it mistook the snow for noise, or if the light's edges were too blurred to trigger a filter. Therefore, without clear insight into the model's internal reasoning, the scientific community must rely on rigorous, empirical "stress testing" of inputs and outputs. We must treat the model's accuracy as the sole justification for its process, necessitating exhaustive benchmarks across every conceivable adverse condition.

This reliance on empirical benchmarking reveals the ultimate enemy of current computer vision: adverse weather. While the human eye is remarkably adaptable, utilizing context and biological high-dynamic-range capabilities to see through obscurants, computer vision systems are brittle when strictly reliant on RGB camera data. Furthermore, these environmental stressors often exacerbate the inherent limitations of the sensors themselves. In nighttime conditions, the dynamic range of standard cameras is tested by the glare of streetlamps and oncoming headlights, which can wash out the color of traffic signals. Conversely, in low-light scenarios, the sensor gain must be increased, introducing digital noise that the CNN may misinterpret as texture or objects. When these lighting challenges are combined with atmospheric particulates like fog or rain, the signal-to-noise ratio drops precipitously.

Munir et al. (2025) suggests that standard YOLO models, while efficient, may lack the feature granularity required to distinguish small, distant traffic lights amidst this environmental noise. This has led to the proposal of feature fusion techniques and architectural enhancement that merges high-resolution spatial data from the shallow layers of the network with the rich semantic

data of the deep layers. By fusing these features, the model may theoretically retain enough fine-grained detail to detect a traffic light's edges even when the semantic understanding is clouded by fog or rain. This study aims to empirically quantify the degradation of the YOLO object detection algorithm under these challenging real-world situations. Moving beyond the "Black Box" limitation, we seek to map the external failure points of the model.

Understanding exactly how different weather patterns disrupt computer vision allows engineers to build more robust safeguards. Until these blind spots in the AI's perception are mapped, quantified, and mitigated, the promise of a fully autonomous future remains suspended behind a veil of uncertainty. This research contributes to that necessary mapping, providing a benchmark for the reliability of upcoming advanced driving assistance systems (ADAS) and ensuring that the "eyes" of the future vehicle are sharp enough to navigate the storm.

2. Research Questions:

This research aims to answer the following question:

> *Does the YOLOv8 model perform worse in adverse weather (rain, snow, fog) in terms of Mean Average Precision (mAP) and inference speed?*
>
> *And if so, which specific environmental condition (Rain, Snow, or Fog) causes the most severe drop in Mean Average Precision (mAP)?*

Hypothesis:

> Null Hypothesis: There is no significant difference in the Mean Average Precision (mAP) of the YOLO model between clear, rainy, snowy, and foggy conditions.
>
> Alternative Hypothesis: There is a significant difference in the Mean Average Precision (mAP) of the YOLO model between clear, rainy, snowy, and foggy conditions.

Predictions

Based on the optical limitations of standard RGB camera sensors and the architectural dependencies of Convolutional Neural Networks (CNNs), I predict that the YOLO model will demonstrate a statistically significant degradation in Mean Average Precision (mAP) across all adverse weather cohorts when compared to the baseline control (Clear Skies).

This prediction is grounded in the understanding that YOLO models, like most Convolutional Neural Networks (CNNs), rely heavily on two specific visual features to identify objects: sharp edge gradients to define spatial boundaries and distinct chromatic signatures to classify the state of the light. Adverse weather systematically corrupts these features in unique ways.

3. Methodology

This study employs a quantitative experimental design to evaluate the robustness of the You Only Look Once (YOLO) object detection algorithm when subjected to environmental stressors. The methodology focuses on benchmarking the degradation of detection performance across strictly classified meteorological scenarios. See Figure 0 for a roadmap for the methodology.

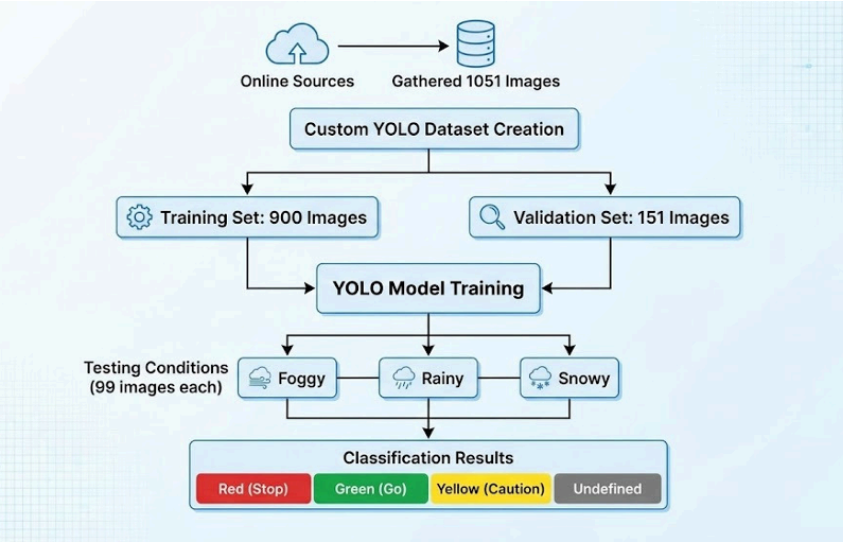

Figure 0) Roadmap for methodology

3.1 Dataset Curation and Preprocessing

To ensure the model establishes a generalized understanding of traffic lights, a primary training dataset ("Dataset A") was aggregated from diverse open-source benchmarks (1051 images that were publicly available online). This dataset contains images of traffic lights under nominal conditions to establish a baseline of "ideal" performance. 900 images were used for training and 151 images were used for validation.

For the testing phase ("Dataset B"), a separate, strictly stratified dataset was curated. This dataset is divided into three distinct environmental cohorts to isolate the impact of weather variables:

- Cohort 1 (Rain): 99 images featuring rain streaks, wet road reflections, and droplet occlusion.
- Cohort 2 (Snow): 99 images featuring falling snowflakes and white-out conditions.
- Cohort 3 (Fog): 99 images featuring dense atmospheric scattering and low contrast.

All images were preprocessed to a standard resolution of 640 pixels. Data augmentation was applied only during the training phase to prevent overfitting; the testing cohorts remain unaugmented to represent authentic driving conditions.

3.2 Experimental Procedure

The YOLOv8 model was trained on 640p resolution and iterated over the course of 100 epochs for training the YOLOv8 model on a custom dataset. The images used for training are captured by the YOLOv8 model and trained to compare against manually labelled data for each of the traffic lights in the images. Manually labelled data contains information regarding the colour of the light (red, green, yellow, or undefined). Undefined being a state for the traffic light where it was not red, green, or yellow, but it was still important to put into its own classification. Examples of undefined traffic lights include images of traffic lights to the side, back or malfunctioning in a way that did not fit into the 3 prior classifications from visual information. Afterwords, five pieces of information are derived from every label for a traffic light: the classification (red, green, yellow, or undefined), x-axis position in the image, y-axis position in the image, x-axis dimension of traffic light, y-axis dimension of traffic light.

The experiment measures the Mean Average Precision (mAP) which considers the recall, precision, and intersection over union (IoU).

Recall quantifies the model's ability to find all relevant targets. It measures the percentage of actual, ground-truth traffic lights in the dataset that were successfully detected by the model. Next precision evaluates the accuracy of the model's positive predictions. It calculates the proportion of detected objects that were correctly classified (e.g., ensuring a prediction labeled "Red Light" is indeed a red light, rather than background noise or a different signal). Lastly, Intersection over Union (IoU) serves as the metric for localization accuracy. It calculates the ratio of the overlapping area to the combined area between the model's predicted bounding box and the manually annotated ground-truth box.

4. Results

This section presents the quantitative performance of the YOLOv8 model across four environmental cohorts: a Control group (Base Model) and three adverse weather scenarios (Snow, Fog, and Rain). The analysis focuses on the degradation of Mean Average Precision (mAP@50), Class-Specific Recall, and Inference Latency to evaluate the model's robustness.

4.1 Overall Performance Degradation

As hypothesized, the model demonstrated a statistically significant drop in detection accuracy when introduced to adverse weather conditions. Looking at Figure 1 we see that under ideal conditions characterized by clear skies, the Base Model achieved a mean Average Precision (mAP@50) of 0.599, establishing the upper bound of the system's capabilities. However, the introduction of adverse weather variables precipitated a performance collapse exceeding 60%

across all testing cohorts. As illustrated in the performance data, Snow proved to be the most manageable of the adverse conditions, maintaining a mAP of 0.217. In contrast, Rain recorded the lowest overall precision score of 0.129, falling even below Fog at 0.158. This hierarchy suggests a divergence in failure modes: while fog makes objects difficult to visually locate, rain introduces significant false positives through surface reflections, which severely penalizes the precision metric.

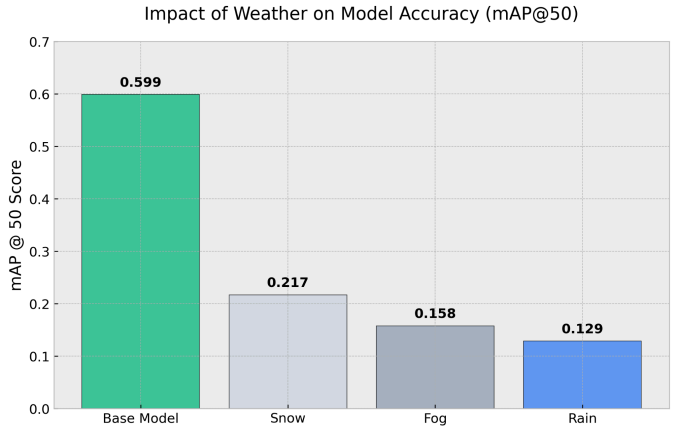

Figure 1) Mean Average Precision (mAP) comparison between different weather conditions

4.2 Possible Reasons for Degradation

Snow presents a dual threat of occlusion from falling flakes and "whitewashing," where accumulation lowers contrast and obscures the distinct edges of the traffic light housing. Similarly, fog alters the image physics through Mie scattering, acting as a "low-pass filter" that blurs edges and desaturates the distinct chromatic signatures required for detection. However, rain arguably presents the most severe and chaotic challenge for models like YOLO. Unlike conditions that simply obscure data, rain introduces active distortions; In Figure 2, we see droplets on the lens refract light, while wet pavement creates mirror-like reflections that may generate "false positive" light sources, creating a complex environment of phantom signals that is significantly harder to interpret than passive obscuration.

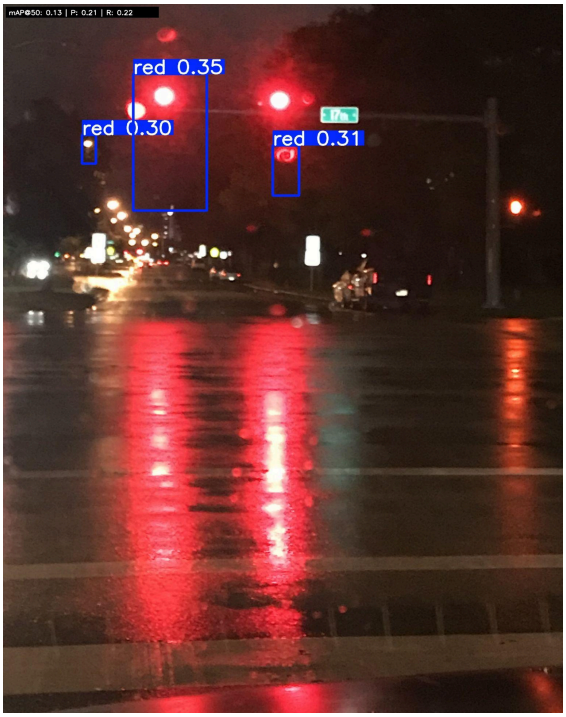

Figure 2) Wet pavement creates mirror-like reflections

4.3 The Small Object Dilemma

A primary driver for the critically low mAP scores across all adverse categories is the distinct spatial characteristics of the dataset. In the bottom right of Figure 3, we see that the label distribution analysis reveals that the vast majority of traffic lights occupy a negligible pixel area relative to the full image frame, often constituting less than a few percent of the total visual data. For Convolutional Neural Networks (CNNs) like YOLO, this presents a "vanishing feature" problem. In clear skies, the sharp high-contrast edges of the traffic light housing allow the model to retain these few pixels through the down-sampling layers. However, when rain streaks or atmospheric scattering from fog corrupt these already scarce pixels, the feature signature is effectively erased. The model becomes unable to distinguish the signal of the traffic light from the visual noise of the weather, leading to the severe drops in accuracy observed in the results.

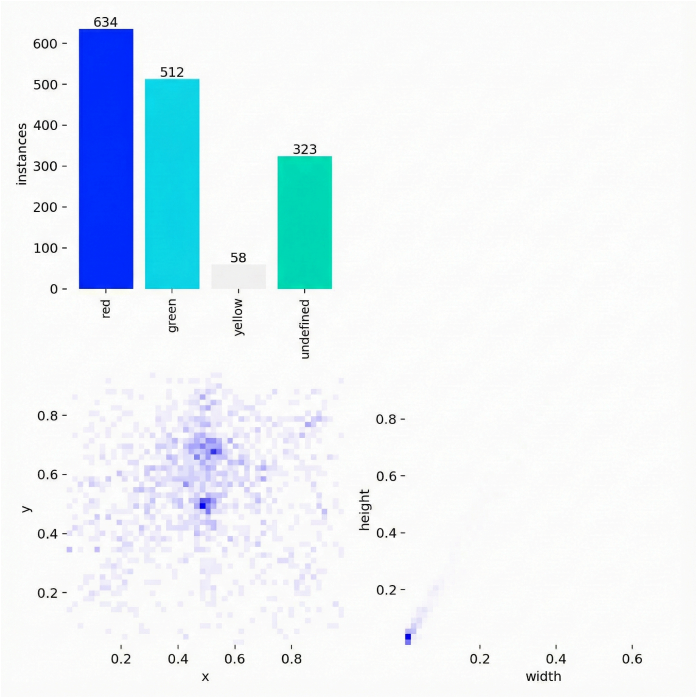

Figure 3) Top left - Frequencies of different classifications; Bottom left - Relative position of traffic lights in the dataset for training; Bottom right - Relative dimensions of the traffic lights in the dataset for training

4.4 Class-Specific Analysis and Recall Patterns

In Figure 4 we see the class-specific heatmap and recall comparison reveal distinct failure patterns between the different weather types. Across all conditions, the model consistently detected Green Lights (0.389 mAP in Snow) with higher accuracy than Red Lights (0.320 mAP in Snow). This discrepancy may be attributed to the typically higher luminosity of green LEDs or the dataset bias, as there were slightly different instance counts between the classes.

A near-total failure was observed in the detection of Yellow Lights, which dropped to a mAP of 0.018 in snowy conditions. This failure is directly correlated to the severe data imbalance identified in the dataset, where Yellow lights comprised only 58 instances compared to over 600 for Red lights. Consequently, the model lacked sufficient training examples to learn to separate the features of a yellow light from similar-colored environmental noise, such as streetlamps or white snow.

Crucially, the recall data highlights the specific optical nature of fog. While Rain resulted in the lowest overall precision, Fog resulted in the lowest Recall score of 0.184 as seen in Figure 5. This confirms the hypothesis that fog acts as a low-pass blur filter, causing the model to miss the object entirely (False Negatives), whereas rain allows the object to be seen but confuses the classification with visual noise.

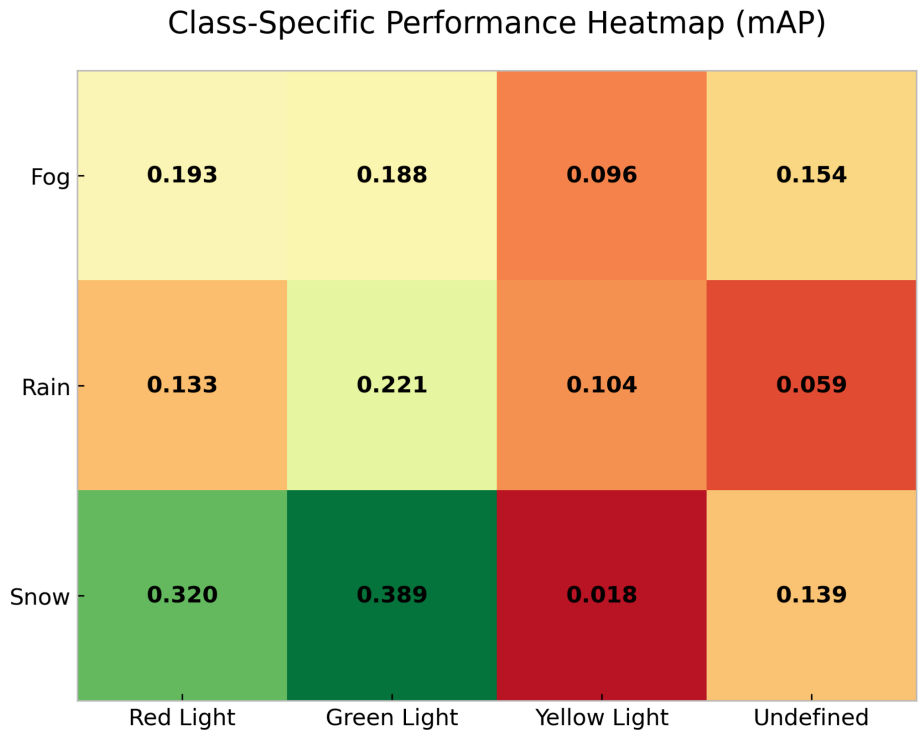

Figure 4) Class specific performance heatmap for mAP

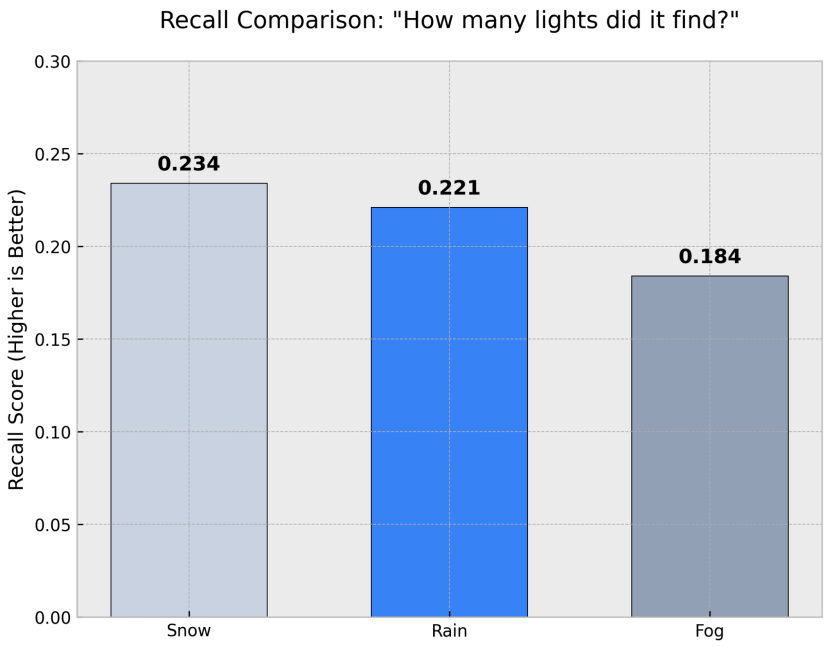

Figure 5) Recall Comparison between different weather conditions

4.5 Inference Latency

Despite the significant drop in accuracy, the model maintained real-time efficiency suitable for autonomous driving. In Figure 6, we see the inference times remained negligible across all cohorts, recording 4.0 ms for Snow, 5.0 ms for Fog, and 5.6 ms for Rain. The slight increase in latency for Rain and Fog suggests that the Non-Maximum Suppression post-processing step required additional time to filter through a higher volume of candidate boxes, likely caused by false positives from reflections or noise artifacts.

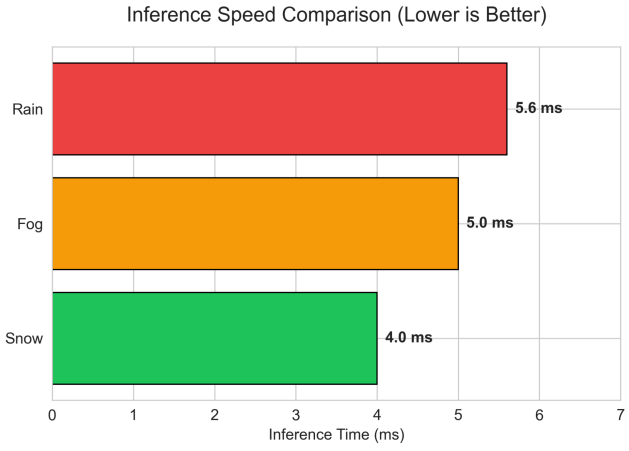

Figure 6) Inference speed between different weather conditions

5. Discussion

The primary objective of this study was to evaluate the robustness of YOLO-based traffic light detection under varying meteorological conditions. While performance drops were anticipated across all adverse weather scenarios, the results indicate a hierarchy of difficulty. Rain proved to be the most detrimental condition for model accuracy, resulting in a 0.470 drop in mean Average Precision (mAP) compared to the clear-weather baseline. This degradation exceeded the losses observed in both snow (0.382 drop) and fog (0.441 drop) scenarios when compared to the clear skies condition.

The distinct performance gap between rain and other conditions can be attributed to the nature of the visual noise introduced. Snow and fog primarily act as passive obstacles. As observed, fog functions as a low-pass filter, reducing high-frequency details, while snow introduces mechanical occlusion. In these cases, the model typically fails via false negatives; it simply cannot see the traffic light.

Rain, conversely, introduces active interference. The wet pavement creates mirror-like surface reflections that mimic the chromatic properties of traffic lights. Simultaneously, droplets on the lens refract incoming light. These phenomena generate high-confidence false positives, where the model incorrectly identifies a reflection on the road or a flare on the lens as a traffic signal. For an autonomous vehicle, a false positive (detecting a green light where there is none) is arguably more hazardous than a false negative (failing to detect a light and defaulting to a safety stop). This suggests that YOLO models, which rely heavily on spatial consistency, are particularly vulnerable to the geometric distortions unique to rainy environments.

It is necessary to acknowledge the limitations of this study. First, the dataset relied on publicly available, free-online images, which may suffer from a lack of high-quality images that would enhance the training process. Second, the rain condition varies wildly in reality, from light drizzle to torrential downpour. Our study grouped these into a single category, potentially masking the specific threshold at which detection fails. Finally, the study utilized a trained YOLO model on a custom dataset with standard data augmentation; specific weather-based augmentation techniques (such as Generative Adversarial Networks to simulate rain) were not applied, which might have mitigated the observed losses.

Future work should prioritize addressing the false positive paradox caused by rain reflections. One promising avenue is the integration of polarizing filters or post-processing algorithms designed specifically to suppress specular highlights on wet roads. Additionally, moving beyond unimodal reliance on RGB cameras to a sensor-fusion approach incorporating thermal imaging or radar could provide the redundancy necessary to distinguish a physical traffic light from its optical reflection on wet pavement.

6. Conclusion

This study provides a critical assessment of the limitations inherent in current computer vision systems for autonomous vehicles. The results strongly support the hypothesis that the YOLO model's detection capabilities are significantly compromised by adverse weather conditions, challenging the assumption of "all-weather" autonomy. The hierarchy of difficulty for precision was determined to be Rain being the most difficult, followed by Fog, and then Snow; however, for pure detection (Recall), Fog proved to be the most challenging environment

While the model demonstrated high precision in the Control (Clear Sky) scenarios, a statistically significant degradation in Mean Average Precision (mAP) was observed across all adverse conditions. The hierarchy of difficulty identified in this study reveals that while fog and snow introduce noise that lowers confidence, rain represents the most critical failure point. These findings suggest that relying solely on camera-based YOLO models is insufficient for autonomous driving in variable climates. The study concludes that visual data must be augmented with sensor fusion technologies, such as LiDAR or Radar, which are less susceptible to optical interference to ensure safety redundancy. Future research should focus on "de-hazing" preprocessing algorithms that can artificially restore contrast to video feeds before they reach the object detection network.

References:

Azam S., Sidratul Montaha, Kayes Uddin Fahim, A.K.M. Rakibul Haque Rafid, Md. Saddam Hossain Mukta, Mirjam Jonkman,. “Using feature maps to unpack the CNN ‘Black box’ theory with two medical datasets of different modality”, *Intelligent Systems with Applications*, Volume 18, (2023) 200233, https://doi.org/10.1016/j.iswa.2023.200233.

Bu, Y., Ye, H., Tie, Z., Chen, Y., & Zhang, D. (2024). OD-YOLO: Robust Small Object Detection Model in Remote Sensing Image with a Novel Multi-Scale Feature Fusion. *Sensors*, *24*(11), 3596. https://doi.org/10.3390/s24113596

Das S, Tariq A, Santos T, et al. Recurrent Neural Networks (RNNs): Architectures, Training Tricks, and Introduction to Influential Research. 2023 Jul 23. In: Colliot O, editor. Machine Learning for Brain Disorders [Internet]. New York, NY: Humana; 2023. Chapter 4. https://www.ncbi.nlm.nih.gov/books/NBK597502/ doi: 10.1007/978-1-0716-3195-9_4

Kampffmeyer M., Arnt-Borre Salberg, R. Jenssen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2016, pp. 1-9. https://www.cv-foundation.org/openaccess/content_cvpr_2016_workshops/w19/html/Kampffmeyer_Semantic_Segmentation_of_CVPR_2016_paper.html

Kong Y., Zepu Wang, Yuqi Nie , Tian Zhou, Stefan Zohren, Yuxuan Liang, Peng Sun, Qingsong Wen. (2024). Unlocking the Power of LSTM for Long Term Time Series Forecasting. Arxiv. https://arxiv.org/html/2408.10006v1

Li S., X. Zhao, L. Stankovic, D. Mandic. Demystifying CNNs for Images by Matched Filters (2022), pp. 1-10, https://doi.org/10.1016/j.ymben.2022.05.007

Meyer, J., Becker, H., Beosch, P. M., & Axhausen, K. W. (2017). Autonomous vehicles: The next jump in accessibilities? Research in Transportation Economics, 62, 80–91. doi:10.1016/j.retrec.2017.03.005.

Mungoli N. Adaptive Ensemble Learning: Boosting Model Performance through Intelligent Feature Fusion in Deep Neural Networks. Arxiv. https://doi.org/10.48550/arXiv.2304.02653

Park Y., H.S. Yang .Convolutional neural network based on an extreme learning machine for image classification. Neurocomputing, 339 (2019), pp. 66-76, 10.1016/j.neucom.2018.12.080

Pettigrew, S. (2017). Why public health should embrace the autonomous car. Australian and New Zealand Journal of Public Health, 41(1), 5–7. doi:10.1111/1753-6405.12588.

Shariff, A., Bonnefon, J.-. F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. Nature Human Behaviour, 1(10), 694–696. doi:10.1038/s41562-017-0202-6.

Wang B., R. Ma, J. Kuang, Y. Zhang. How decisions are made in brains: Unpack “Black Box” of CNN with Ms. Pac-Man Video Game. IEEE access : practical innovations, open solutions, 8 (2020), pp. 142446-142458, 10.1109/ACCESS.2020.3013645

Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging. 2018 Aug;9(4):611-629. doi: 10.1007/s13244-018-0639-9. Epub 2018 Jun 22. PMID: 29934920; PMCID: PMC6108980.

Zhang Nan, Chen Yun-shan and Wang Jian-li, "Image parallel processing based on GPU," *2010 2nd International Conference on Advanced Computer Control*, Shenyang, China, 2010, pp. 367-370, doi: 10.1109/ICACC.2010.5486836.

Zhou, Q., Zhang, D., Liu, H., & He, Y. (2024). KCS-YOLO: An Improved Algorithm for Traffic Light Detection under Low Visibility Conditions. *Machines*, *12*(8), 557–557. https://doi.org/10.3390/machines12080557