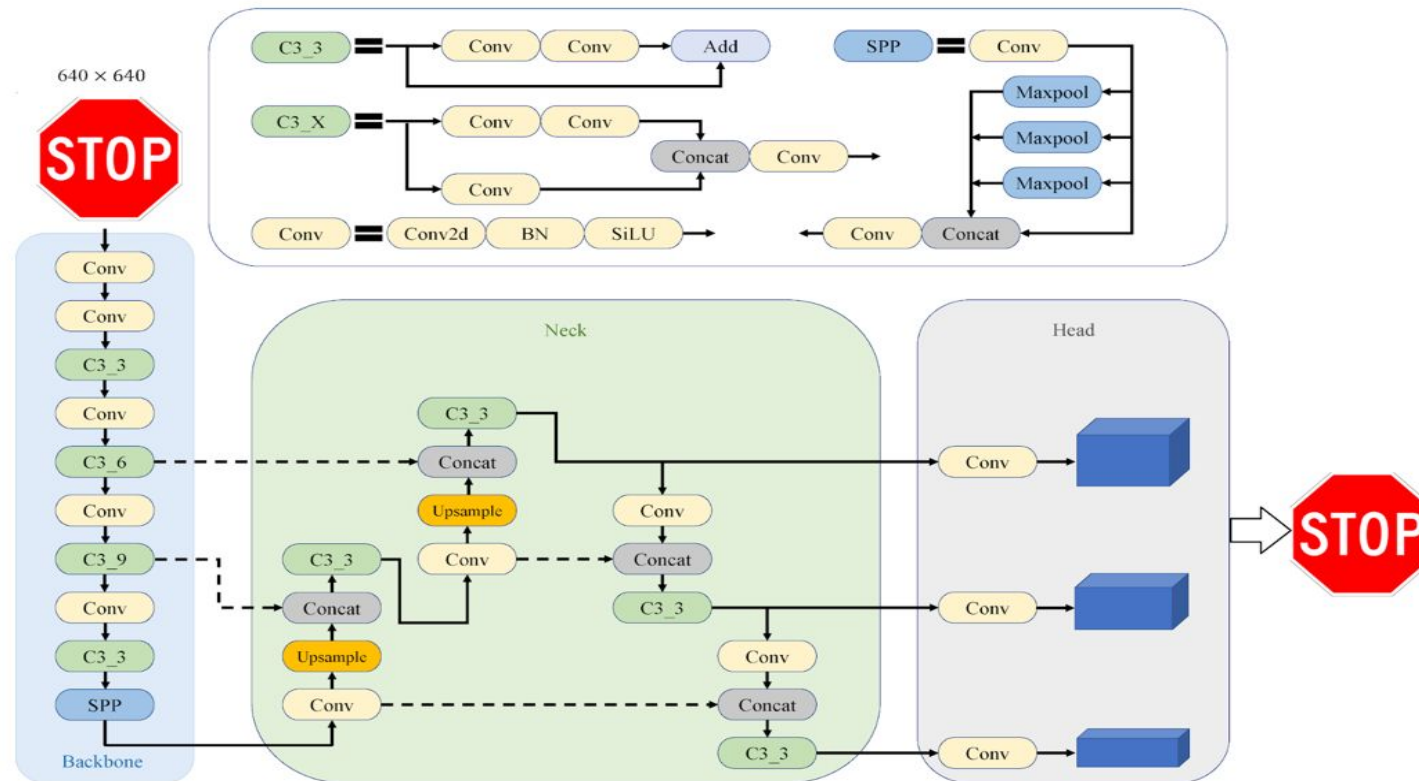


TrustworthyAI: Adversarial Attacks and Defensive Strategies in Self-Driving Systems using Computer Vision and Artificial Intelligence

Ethan Sun



Abstract

- Project focuses on enhancing autonomous vehicle safety against cyberattacks through precise stop sign detection using AI.
- Utilizes YOLO architecture variants including YOLOv5mu, YOLOv8 from Roboflow, and YOLOv8s from Ultralytics, with YOLOv8 from Roboflow proving most effective.
- Development involves calibrating AI system to reliably detect stop signs, crucial for safe autonomous vehicle operation.
- Creates six dataset versions with over 46,535 images, featuring manually crafted cyberattack simulations to comprehensively assess system's precision.
- Achieves up to 90% accuracy in stop sign classification confidence, showcasing effectiveness of YOLOv8 model and comprehensive datasets even in challenging conditions.
- Highlights critical role of computer vision techniques and extensive datasets in bolstering autonomous vehicle safety against cyber threats.
- Review underscores project's success in advancing Trustworthy AI and vehicle safety through innovative use of YOLOv8 model and targeted datasets.

Quick Facts

- 90% of car accidents are caused by human error
- Self-autonomous cars can reduce car accidents by 34%
- Market adoption of self-driving vehicles only projected to be 5% by 2030
- 56% of Americans remain skeptical about self-driving vehicles
- Cyberattacks have surged over 600% in 2023



Background Research

Problem:

- A critical challenge in the deployment of self-driving cars is ensuring their ability to accurately perceive and react to the environment.
- Many cyber attacks, although seemingly minor, can have significant repercussions, including the potential to cause accidents or disrupt traffic systems.
- This is especially pertinent in scenarios where external factors, such as vandalized road signs or subtle cyberattacks, can compromise the vehicle's sensors and decision-making algorithms.
- Such attacks aim to deceive the AI algorithms responsible for interpreting sensor data, potentially leading to catastrophic outcomes.

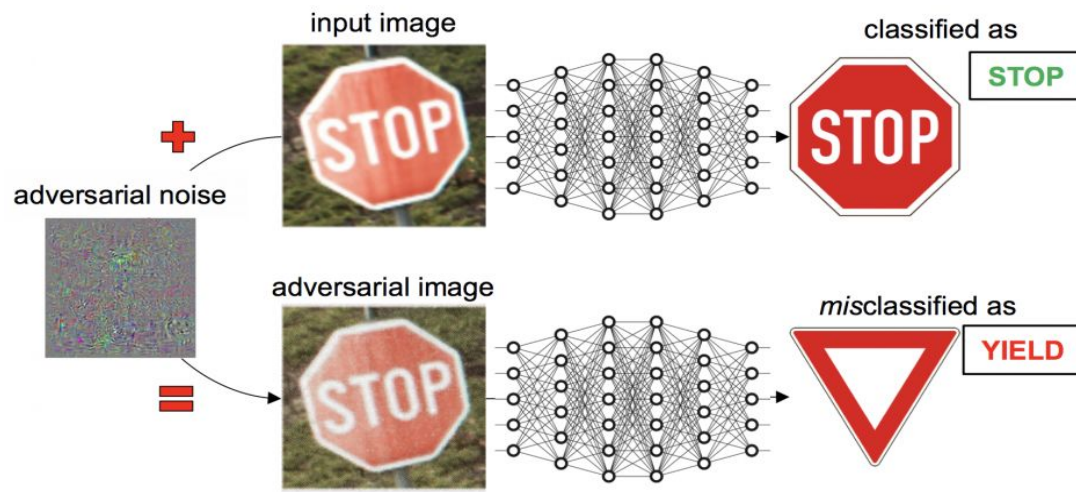


Common cyber attack scenarios to self-driving machine learning

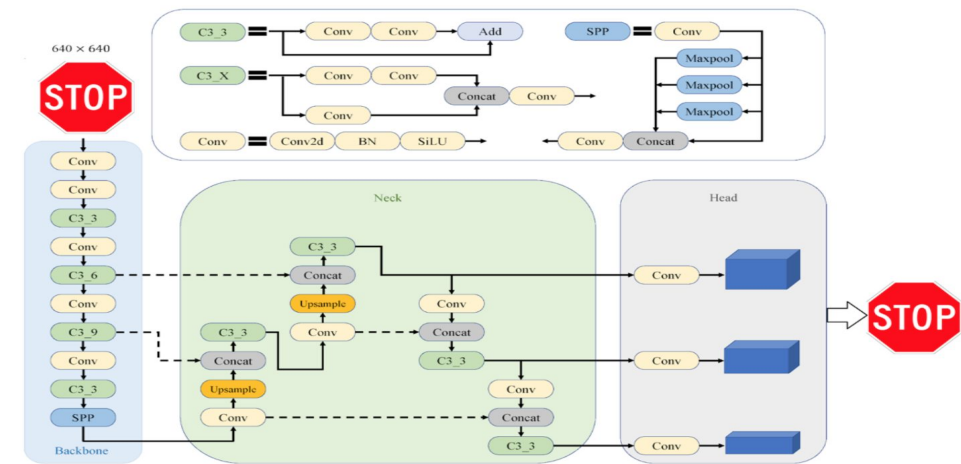
Background Research

Solution:

- In addressing the challenges posed by cyberattacks on autonomous vehicle sensors, the proposed solution involves a multi-layered approach leveraging advanced deep learning techniques, anomaly detection, and data augmentation.
- This solution aims to enhance the resilience of image detection systems, with a particular focus on improving traffic sign recognition in various adversarial scenarios.



Stop Sign recognition model based on computer vision algorithm



YOLO 5 architecture

Background Research

- The YOLOX model is an advancement in object detection, surpassing previous versions of YOLO in performance. It incorporates several key innovations. Firstly, it employs an anchor-free mechanism which simplifies the detection process by reducing the number of design parameters and predictions per image, enhancing efficiency. Additionally, YOLOX uses a concept called 'multi positives' which optimizes high-quality predictions to balance the training process. Finally, the SimOTA feature in YOLOX involves advanced label assignment, considering factors like loss/quality awareness and center prior, to enhance detection accuracy.

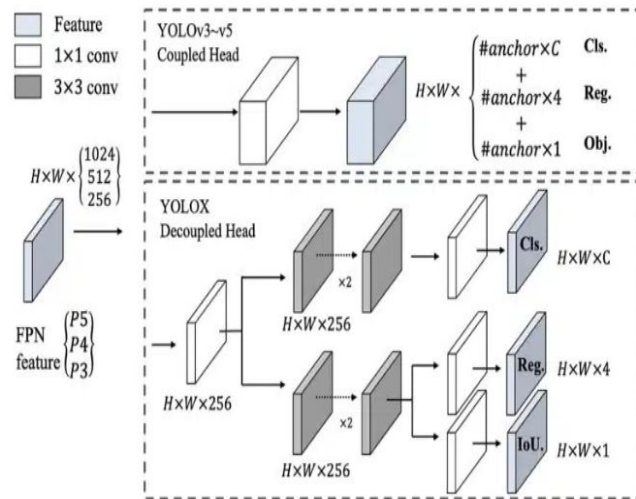
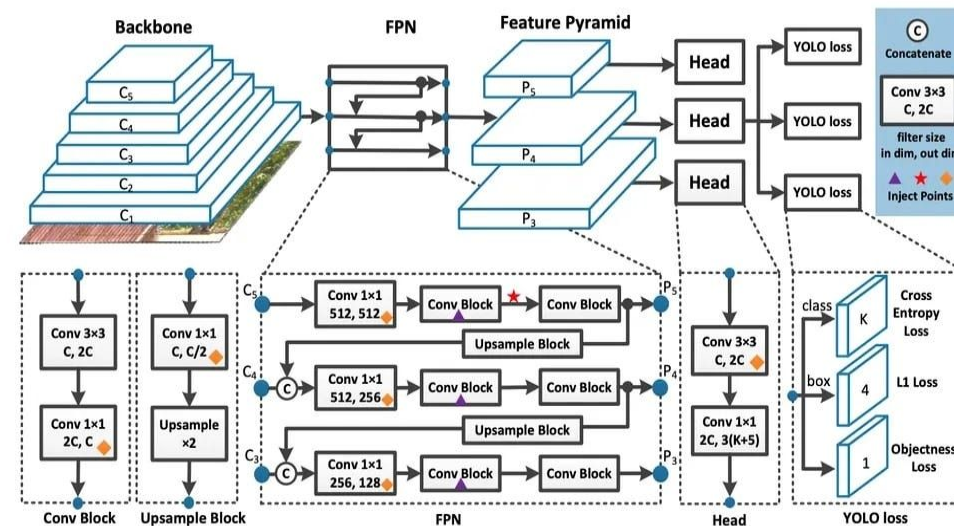


Illustration of the difference between YOLOv3 head and the proposed decoupled head



Computer vision neural network training architecture processing

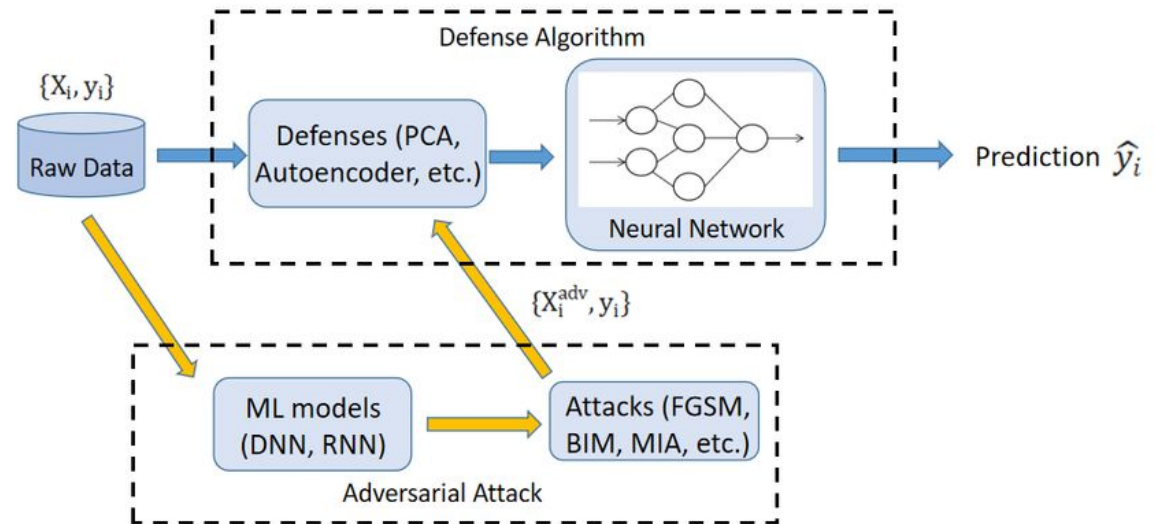
Statement of Purpose, Hypothesis

Statement of Purpose:

This project seeks to create an AI solution enhancing the cybersecurity of self-driving vehicles by improving image detection capabilities to more accurately identify and neutralize anomalies and adversarial attacks, such as traffic signs.

Hypothesis:

If our model is used on a self-driving car, then it should be able to detect if a traffic sign has undergone a cyber attack because with the given dataset, it recognizes the subtle differences between a regular traffic sign and one that has been cyber-attacked.



Procedure

1. Define “Cyber attack”

- Defining what constitutes a cyberattack in this context requires a clear understanding of the threshold at which a perturbation becomes significant enough to mislead the sensor.

$$T = \min \{ \delta : P(X + \delta) \neq P(X) \}$$

- The threshold T is the smallest amount by which the given sensor input X shifts such that the resulting perturbation is different from without perturbation.

2. Collect Data

- Collecting scenarios that will train the model to be able to recognize and respond to cyberattacks.
- $\Gamma(C)$ is the fraction that quantifies how comprehensively the set C represents the range of possible scenarios

$$\Gamma(C) = \frac{|\{\text{scenario covered by } C\}|}{|\{\text{total possible scenarios}\}|}$$

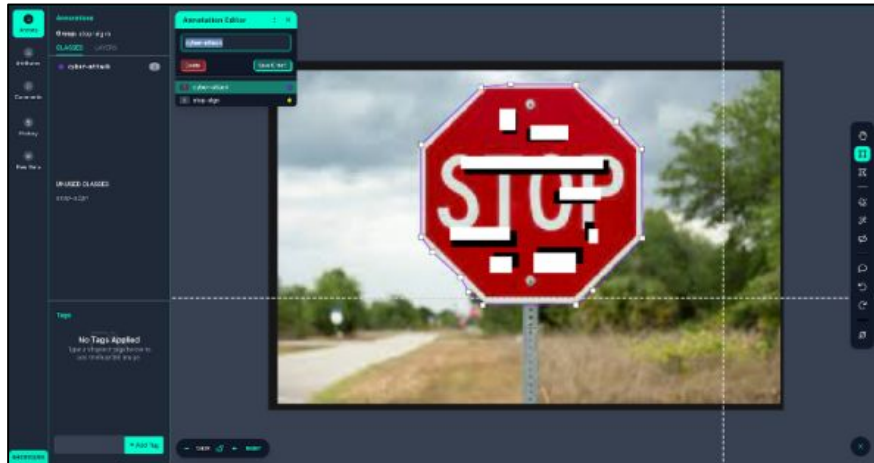
Procedure

3. Augment/label the data

- The data needs to be pre-processed so that the model can use it train

$$A(x) = \{x', x'', x''', \dots\}$$

- The augmentation of an image x contains some form of augmentation being rotation, scaling, noise represented by x', x'', x''' , etc.



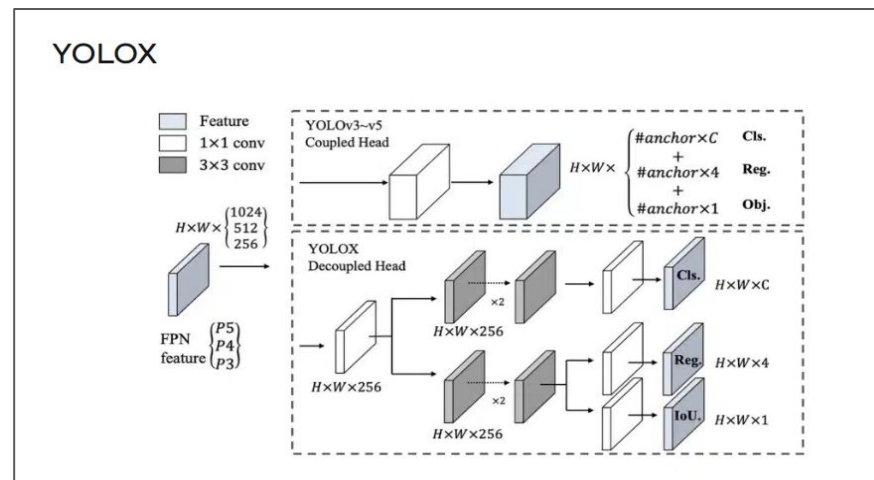
Dataset sample and dataset generation with Roboflow

Procedure

4. Train the Model

- The data is put under the YOLOv8 architecture to train and identify patterns/labels

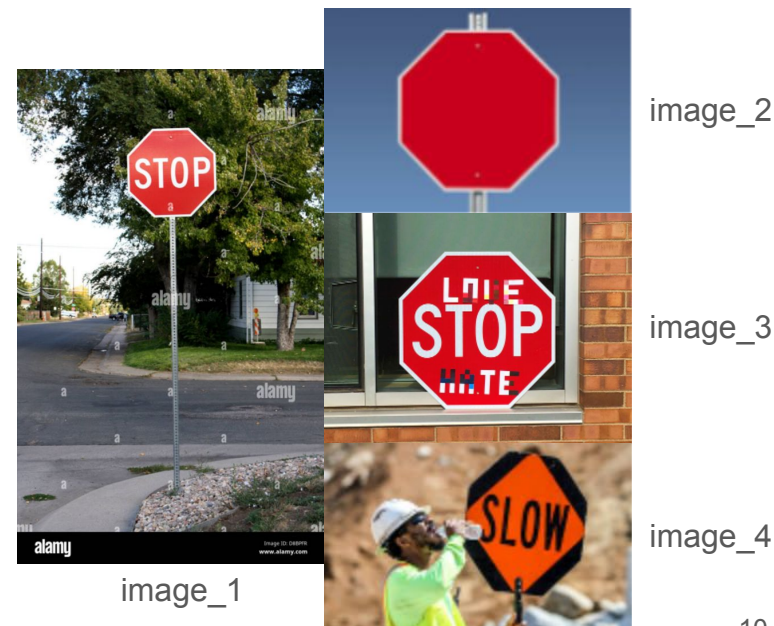
$$V(s) = E \sim E[R(s, a) + \gamma V(s!)]$$



5. Test the Model

$$mAP = \frac{1}{|classes|} \sum_{c \in classes} \frac{|TP_c|}{|FP_c| + |TP_c|}$$

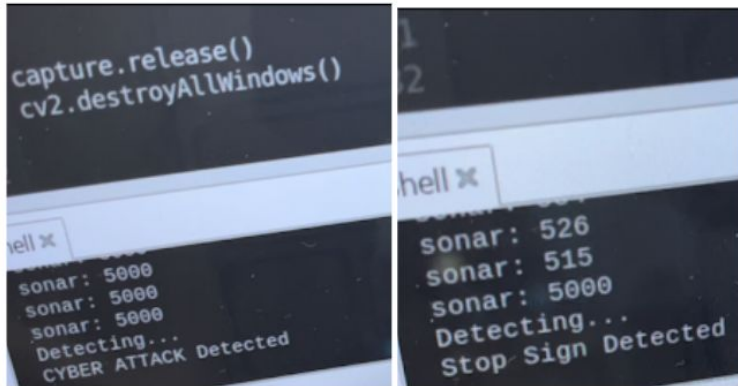
- Use base images to test accuracy and consistency of model
- Compare with other models



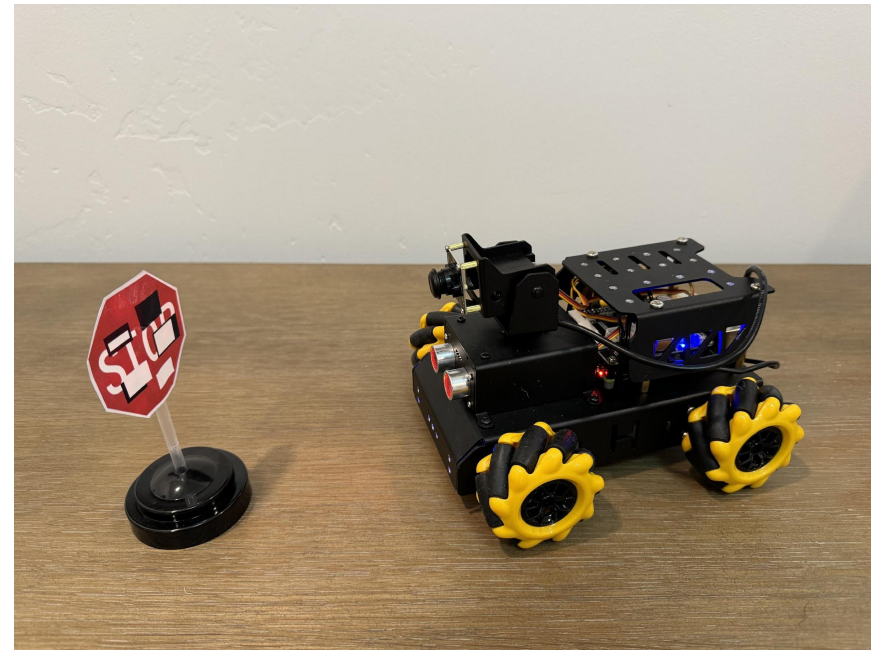
Procedure (Demo)

6. Model Accuracy Verification with Robot Self-Driving Car

- On-device computer vision algorithm
- Real-time detection and analysis
- Adaptive learning and data collection



```
capture.release()  
cv2.destroyAllWindows()  
hell x  
sonar: 5000  
sonar: 5000  
sonar: 5000  
Detecting...  
CYBER ATTACK Detected  
hell x  
sonar: 526  
sonar: 515  
sonar: 5000  
Detecting...  
Stop Sign Detected
```

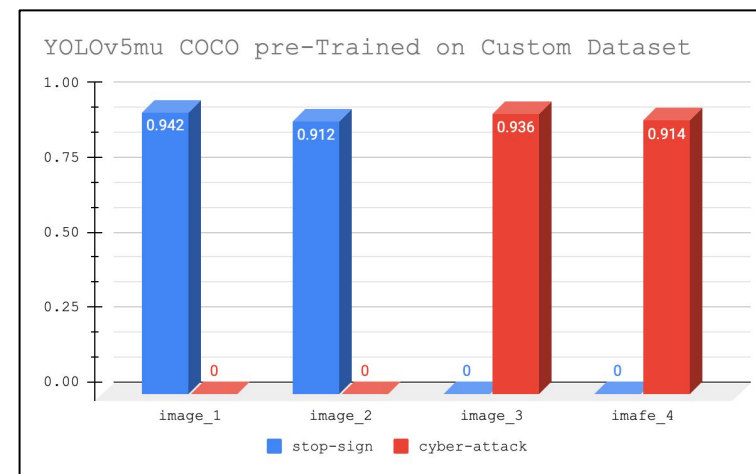
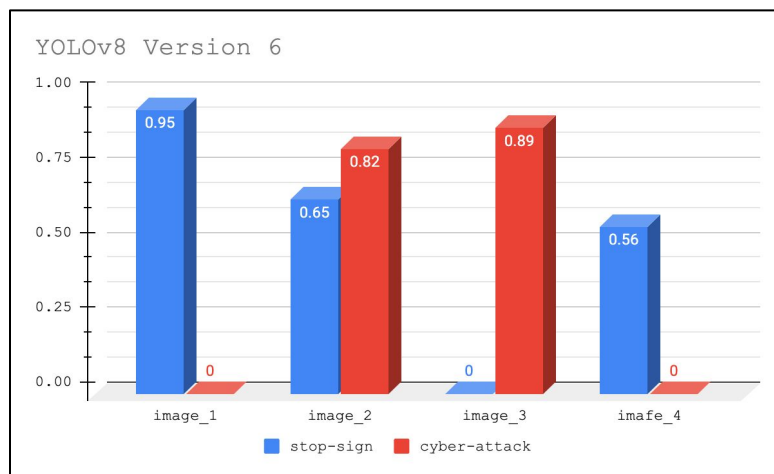


Results

A comparison between 2 different computer vision models in the same dataset. YOLOv8 Version 6 shows the algorithms ability to accurately identify stop signs as well as classify cyber-attacks, while also not mistaking the blank stop sign and slow sign as a stop sign. YOLOv5 mu performs just as well when identifying stop-signs but misclassified when identifying cyber-attacks

	YOLOv8 Version 6	
	stop-sign	cyber-attack
image_1	0.95	0
image_2	0.65	0.82
image_3	0	0.89
image_4	0.56	0

	YOLOv5mu COCO pre-trained	
	stop-sign	cyber-attack
image_1	0.942	0
image_2	0.912	0
image_3	0	0.936
image_4	0	0.914



Results

image	stop sign average	cyber attack average
image_1	0.857125	0.0375
image_2	0.316125	0.61075
image_3	0	0.838875
image_4	0.209875	0.284

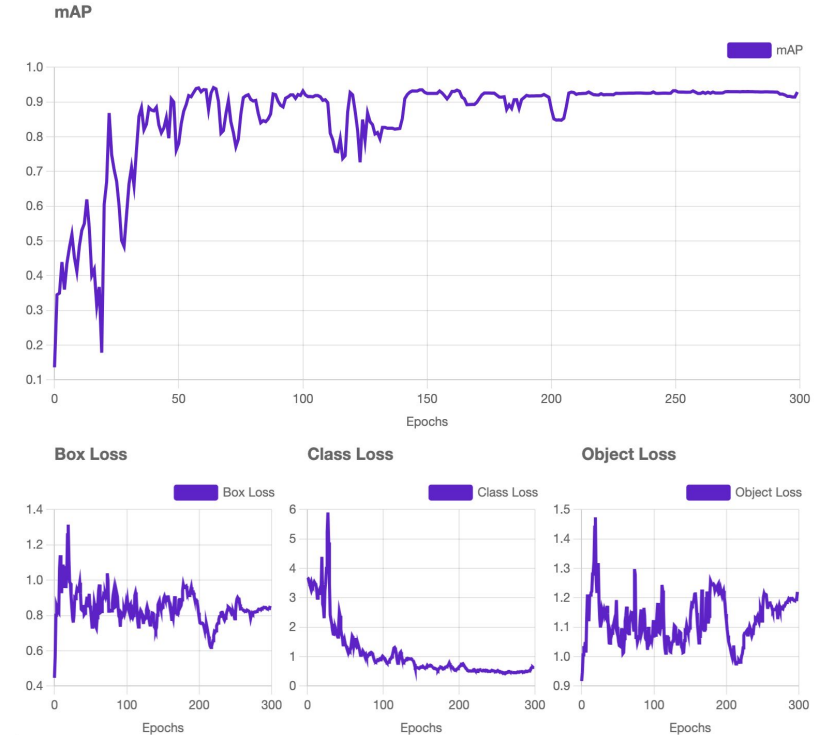
The average in accuracy of every model when classifying stop signs and cyber attacks

- Image 1 is normally accurately detected as a stop sign but in a few poor models is not
- Image 2, on average, can be accurately detected as not being a stop sign or a cyber attack
- Image 3 has a lower consistency than image 1 but normally is accurately detected as a cyber attack
- Image 4, on average, does better than image 2 on realizing the image is neither a stop sign or a cyber attack

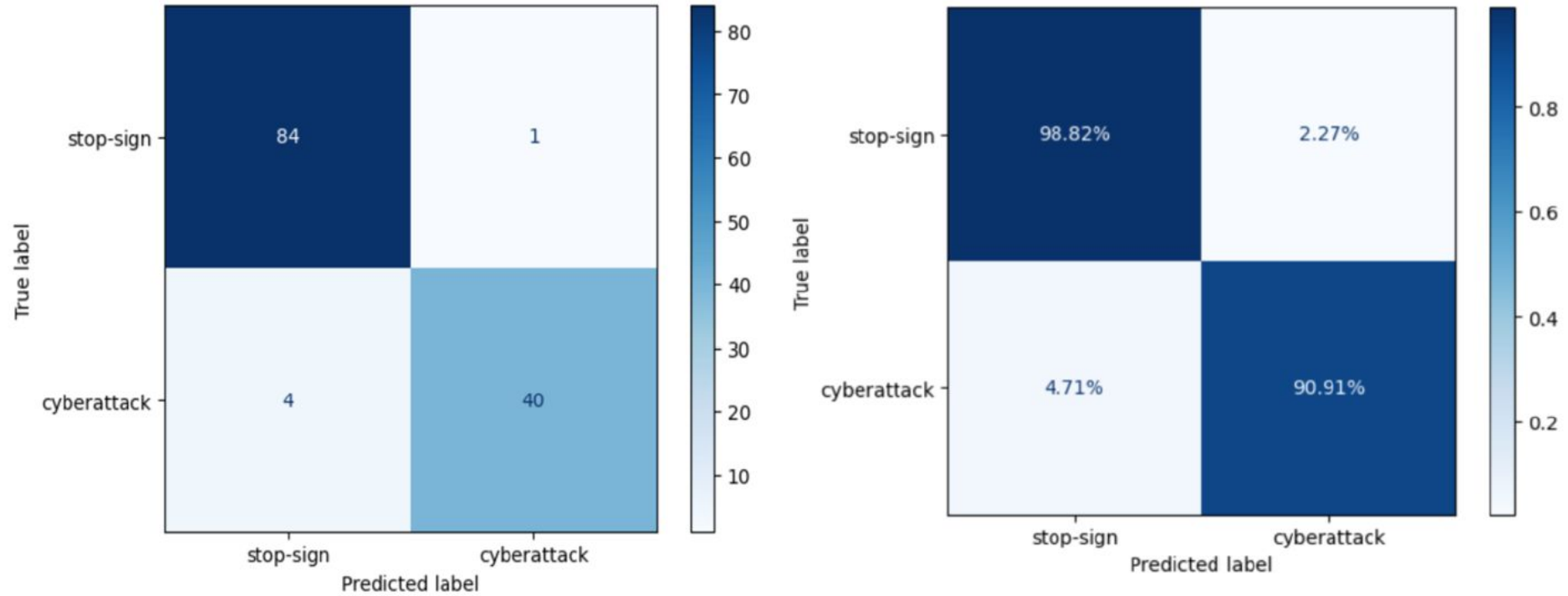
Results

(YOLOv8 Version 6 - Roboflow)

- Mean Average Precision = average precision across multiple levels of confidence thresholds
 - i. Provides a single value that summarizes the overall performance of a model
 - ii. Performance increases as more epochs
- Box Loss
 - i. A component of the total loss that penalizes the model for errors in predicting the coordinates of bounding boxes
- Class Loss
 - i. Another component of the total loss that indicates performance on predicting class
- Object Loss
 - i. Another component that penalizes the model for failing to detect and classify objects present in the grid cell
- More epochs -> better results



Confusion Matrix (YOLOv8)



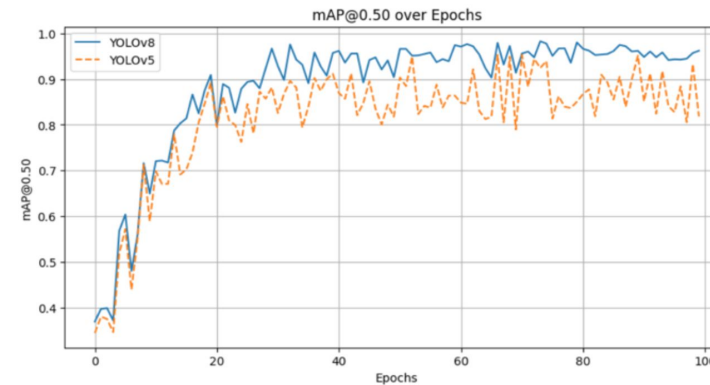
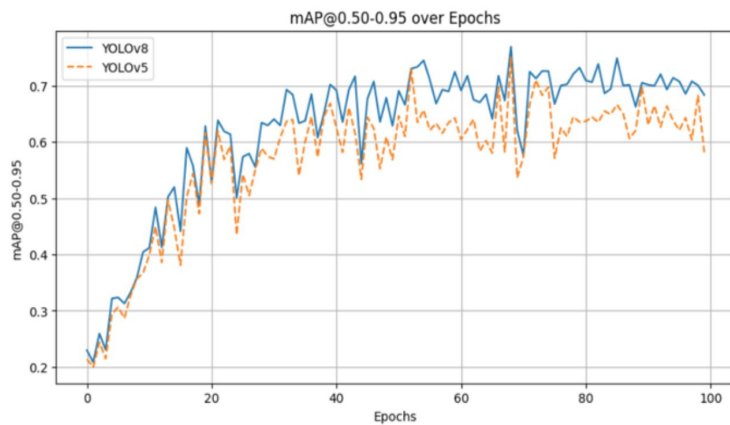
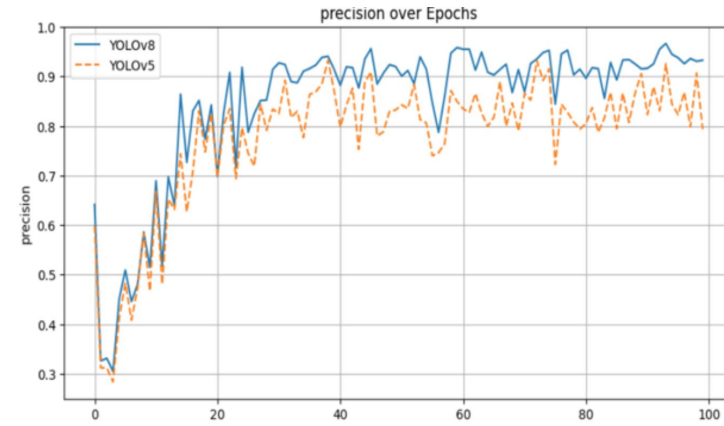
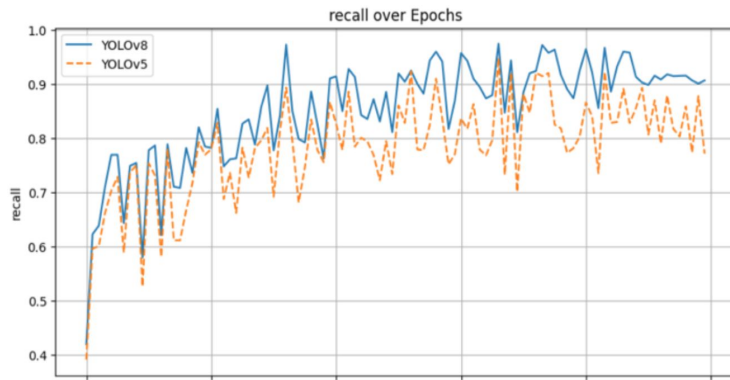
Confusion Matrix YOLOv8

Results (YOLOv8s - Ultralytics)



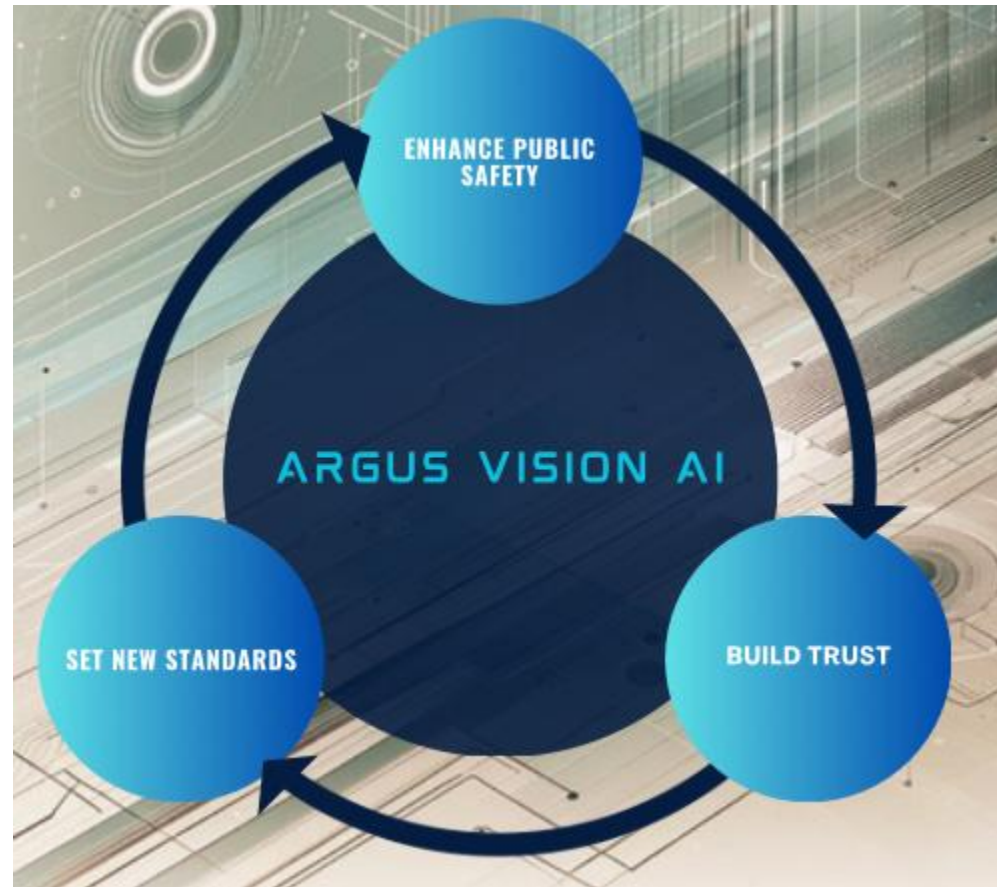
- mAP50(B) = Mean Average Precision at 50 IoU - for Boxes
 - i. 50 IoU = threshold used to determine whether a predicted bounding box is considered a correct detection (if it is above 50% it is a correct detection)
 - ii. (B) = bounding boxes
- mAP50-95(B) = same as above except at a higher predicted bounding box threshold
- precision(B) = accuracy of the predicted bounding box
- recall(B) = ration of true positive predictions to the total number of actual positives (how well the model captures all relevant objects in the dataset)

Results (YOLOv8 VS YOLOv5)



Comparison with Yolo V5 and Yolo V8 in Patch Training

Social Impact: The Flywheel Effect



Conclusion

- Various methodologies employed to address challenges of cyber attacks on autonomous vehicle sensors.
- Definition and classification of cyberattacks crucial in shaping diverse training sets for AI models.
- Datasets ranged from a few dozen to over 46,000 images, pivotal in training AI models.
- Version 6 dataset demonstrated remarkable ability to recognize normal and manipulated traffic signs with up to 90% accuracy.
- Selection and optimization of YOLOv8 model from Roboflow critical for superior performance in recognizing and classifying stop signs accurately.
- Project's uniqueness lies in its focus on recognizing cyber attacks on traffic signs in autonomous vehicle domain.
- Sets foundation for protective countermeasures in AI-assisted society.
- Represents significant advancement in Trustworthy AI for autonomous vehicles.
- Highlights potential of advanced computer vision models and comprehensive datasets in enhancing safety and reliability of autonomous vehicle systems.
- Provides insights for further advancements in protecting autonomous vehicles from evolving cyber threats.

References

1. Zhang, J., Lou, Y., Wang, J., Wu, K., Lu, K., & Jia, X. "Evaluating Adversarial Attacks on Driving Safety in Vision-Based Autonomous Vehicles." IEEE Internet of Things Journal. (2021).
2. D, D. (2023, June 19). Self-driving car accident statistics for 2023. Dordulian Law Group. <https://www.dlawgroup.com/self-driving-car-accident-statistics-2023/>
3. Self-driving car statistics. NST Law. (2023, June 8). <https://www.nstlaw.com/autonomous-vehicle-statistics/>
4. Giordani, J. (2022, April 21). Council post: Cyberattacks on vehicles pose a threat to drivers and manufacturers. Forbes. <https://www.forbes.com/sites/forbestechcouncil/2021/12/10/cyberattacks-on-vehicles-pose-a-threat-to-drivers-and-manufacturers/?sh=7e88bf954620>
5. Eliot, L. (2020, December 28). Largest ever Cyber Hack provides vital lessons for self-driving cars. Forbes. <https://www.forbes.com/sites/lanceeliot/2021/12/29/largest-ever-cyber-hack-provides-vital-lessons-for-self-driving-cars/>
6. 2024 must-know cyber attack statistics and Trends. Embroker. (2024b, January 4). <https://www.embroker.com/blog/cyber-attack-statistics/>
7. Richter, F. (2019, May 3). Infographic: Self-driving cars still cause for concern for pedestrians. Statista Daily Data. <https://www.statista.com/chart/17881/self-driving-car-safety/>
8. Kehtarnavaz, N., Griswold, N.C. & Kang, D.S. Stop-sign recognition based on color/shape processing. Machine Vis. Apps. 6, 206–208 (1993). <https://doi.org/10.1007/BF01212298>
9. T. P. Cao and G. Deng, "Real-Time Vision-Based Stop Sign Detection System on FPGA," 2008 Digital Image Computing: Techniques and Applications, Canberra, ACT, Australia, 2008, pp. 465-471, doi: 10.1109/DICTA.2008.37.
10. Wenpeng Wang, Yuxuan Su, and Ming Cheng "Real-time stop sign detection and distance estimation using a single camera", Proc. SPIE 10696, Tenth International Conference on Machine Vision (ICMV 2017), 106962A (13 April 2018); <https://doi.org/10.1117/12.2309793>