Evaluating Robustness of Object Detection in Varying Image Complexity

Vatsal Piyush Shah

Received July 27, 2023
Accepted September 15, 2023
Electronic access October 31, 2023

The development of object detection algorithms has increased in the context of autonomous systems, but the prevalence of algorithmic biases, however, continues to be a substantial barrier, especially when meeting objects in surroundings that are different from their training environments. This study aims to explore these biases, assessing the robustness of object detection algorithms across varying conditions, including unfamiliar scenarios and modified image sets. We adopted a method to synthetically generate images and automatically annotate them. By using a dictionary of coordinates for image superimposition, along with YOLOv8 segmentation and OpenCV's masking features, we effectively mapped new foreground images on existing background images from a standardized dataset of images in order to add to the complexity of the images. This methodology provides a versatile platform for testing the performance of various detection algorithms as it can be replicated to test other industry object detection algorithms apart from YOLOv8, aiding the path of understanding the flaws in these algorithms. To evaluate its effectiveness, we created three distinct datasets, namely L1, L2, and L3, each designed with varying degrees of image modification. Our extensive ablation study rigorously examines the influence of synthetic perturbations on the performance of object identification models against real-world distribution shifts, revealing key algorithmic biases in certain unfamiliar scenarios such as the inability to detect a motorbike or misclassifying a truck for a train. This is due to factors such as image lighting, visibility, object size, and colour. Our findings give significant insights for academics and practitioners interested in designing more robust and accurate object identification models for real-world applications even in unconventional situations to ensure a greater reliance on such autonomous systems. In summary, this study aims to not only propel the trajectory of object detection technologies but also to anchor their assimilation within our evolving world.

Keywords: object detection, algorithmic biases, autonomous systems, image modification

Introduction

Deep learning, empowered by sophisticated sensors and GPUs, has revolutionized diverse fields, from virtual assistants and medical imaging to content recommendation and fake news detection. A paramount application is autonomous vehicles, which is making strides in an effortless driving experience where all tasks can be done with the speed of a single click[1]. Nonetheless, these technologies are not without imperfections. Autonomous systems, be they vehicles or others, depend on accurate environmental perception and prompt decision-making. While tools like cameras, Lidars and radars enable data collection, there remain obstacles in ensuring reliable detection under variable real-world conditions, such as fluctuating weather or lighting. This underscores that we have yet to enter an era in which complete reliance on these technological systems is feasible, thereby requiring a prudent approach to their utilization and raising concerns for exercising caution when operating such vehicles.

With this backdrop, this paper aims to investigate the following research question: "How do machine vision algorithms, employed in autonomous driving, respond when exposed to modified existing data, and what potential algorithmic biases arise in their capacity to precisely detect objects?"

Overview of Object Detecting Models

Exploring one-stage object detection algorithms like You Only Look Once (YOLO) is pivotal in this research. These algorithms merge object classification and bounding box regression into a single step, prioritizing real-time applicability and swift detection speeds. Despite historically exhibiting slightly less accuracy than their two-stage counterparts, ongoing optimizations, as observed in recent benchmark studies[2], are effectively narrowing this performance gap. This trend towards refinement isillustrated in Figure 1[3], underscoring the algorithm's evolving efficacy.

However, even with these strengths, deep learning algorithms often grapple with the intricacies of object detection, stemming from variations in object color, background settings,

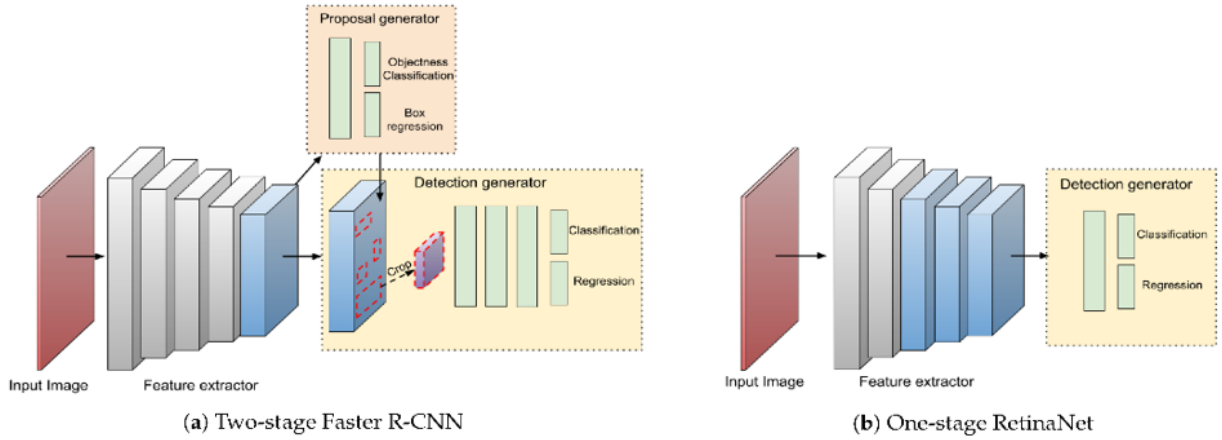

Fig. 1 Diagram illustrating difference between one-stage and two-stage detector

and size. This persistent challenge underscores the imperative for perpetual research and development within this realm.

These limitations have been exemplified by unfortunate incidents involving Tesla owners, Joshua Brown and Jeremy Banner, who tragically lost their lives while using Tesla's Autopilot system[4]. Brown's Model S collided with a tractor-trailer as the camera failed to recognize the white truck against a bright sky, which emphasizes the importance of comprehending how algorithms respond to modified data and unfamiliar scenarios. Considering these incidents, this study gains substantial relevance and urgency. By shedding light on the intricacies behind object detecting algorithms, this study not only contributes to advancing the understanding of algorithmic behavior but also holds the potential to drive improvements in object detection systems, ultimately leading to safer and more reliable autonomous driving. Additonally, describing the Faster R-CNN object detection model has two primary phases to object identification architecture: the Region Proposal Network (RPN) and Region of Interest (RoI) Pooling. Region Proposal Network produces many candidate regions that are expected to contain items. These suggestions are implemented to draw information from the feature map produced by the backbone network. RPN creates anchors at various sizes and aspect ratios, then predicts whether or not each anchor includes an item (classification) and then refines the anchor's coordinates (bounding box regression) These steps can be described by the equations below:

$$LR_{PN}(p_i,t_i) = \frac{1}{N_{cls}} \sum_{i}^{N_{cls}} L_{cls}^{(p_i,p_i^*)} + \lambda \frac{1}{N_{reg}} \sum_{i} p_i^* L_{reg}^{(t_i,t_i^*)}$$

Where:

p_i^* If the anchor is a positive sample, the ground truth label is 1, otherwise it is 0.

t_i^* For positive samples, the bounding box regression parameters are denoted by t_i^*.

N_{cls} The total number of classifiable positive and negative samples is denoted by N_{cls}.

N_{reg} The number of valid samples used for regression is denoted by N_{reg}

λ A balancing parameter.

Region of Interest (RoI) Pooling takes the feature map produced by the backbone network and extracts fixed-size feature maps based on the region suggestions provided by RPN. Equations may be written to explain the RoI pooling process, which involves segmenting the RoI into a predetermined number of bins and then using bilinear interpolation to combine the characteristics from each bin.

$$P_{i,j} = \frac{1}{A_{i,j}} \sum_{(x,y)\in A_{i,j}} x(x,y)$$

Coordinates (x, y) denote a point within the RoI's $i-th$ bin of $j-th$ channel.

Bin's area is denoted by the notation $A_i j$.

Overview of Performance Matrices

This study focuses on key metrics for evaluating object detection algorithms, including recall, precision, F1 score, and Mean Average Precision (mAP). Recall measures the correct identification of true positives, while the F1 score balances precision and recall. Figure 2 visually illustrates the relationship between true and false positives and negatives[5]. mAP, a common benchmark, summarizes precision and recall at different thresholds, providing a comprehensive assessment of an algorithm's accuracy in detecting objects. These metrics offer valuable insights into the algorithm's performance, accounting for false positives and false negatives.

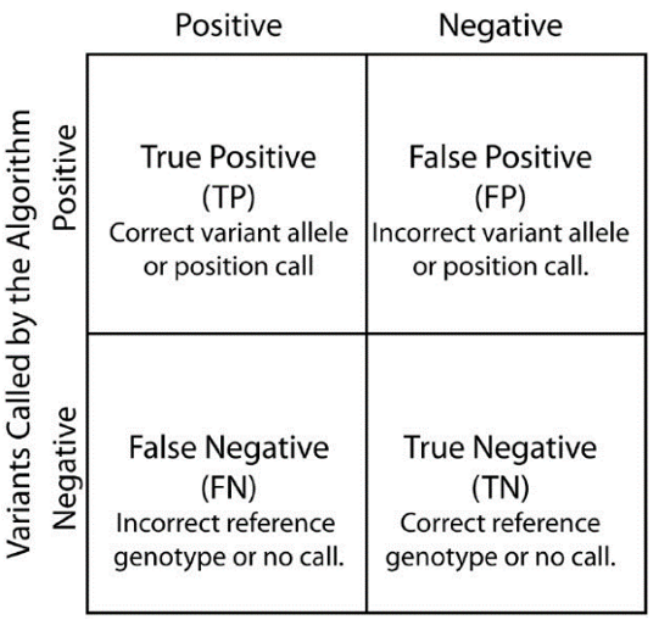

Fig. 2 Relationship between true and false positives and negatives

YOLOv8 Architecture

The architecture in Figure 3[6] is based on an adapted CSP-Darknet53 foundation, transitioning to the utilization of a C2f module instead of the previously employed CSPLayer witnessed in YOLOv5. In order to hasten the computational process, a swift spatial pyramid pooling (SPPF) layer is integrated, which consolidates features into a predetermined dimension map. During each convolution process, batch normalization paired with SiLU activation is implemented. Moreover, the nucleus of the structure is designed to individually manage tasks related to object recognition, classification, and regression, enhancing the focus and efficiency in each sector.

Research Hypothesis

Deep learning algorithms have undeniably advanced the field of object detection. However, when such algorithms are subjected to artificially alter visual data, their robustness and adaptability become questionable. Specifically, it can be posited that their performance degrades when processing images that deviate considerably from the training dataset, especially concerning the target object's size, placement, and its interplay with the background.

As illustrated in Figure 4, a basic background image is taken, and onto this image, a portion of a different foreground image is added using OpenCV. This creates a new and potentially complex connection between the images for object detection algorithms. In such altered scenarios, there is a potential that the algorithms might misidentify or mis-categorize objects. This concern is heightened for smaller objects or those with distinctive colors. The motive of this study is to uncover the impact of these artificial alterations on the performance of detection algorithms, using visual aids to demonstrate these concepts and biases.

Related Work

This study extends the foundation laid by previous research efforts. Its primary objective is to comprehend how object detection algorithms function when confronted with altered data. While earlier investigations have focused on dissecting the inner workings of Convolution Neural Networks (CNNs) and achieving real-time pedestrian detection, this current research takes a distinctive approach by exploring the subtle impacts of data that, although appearing ordinary to humans, could potentially pose challenges for machine learning algorithms operating within the domain of autonomous driving.

Gomez[7] has highlighted the critical role of CNNs in ensuring safe autonomous driving systems. Their work on object detection and synthetic dataset generation offers an imperative backdrop. It raises questions about the reliability of such systems, especially when confronted with altered datasets—a core concern of our research.

Peng et al.[8], by developing a framework using the YOLO algorithm and the mAP method, have reinforced the importance of optimizing object detection in autonomous scenarios. Their use of the Monte Carlo dropout method for uncertainty quantification provides a benchmark for evaluating the reliability of object detection algorithms and assessing the consequences of altered data.

Similarly, Iftikhar et al.[9], while evaluating pedestrian detection and its challenges, underlined the significance of refining real-world object detection—essentially echoing our research's focal point. Their emphasis on various confidence sensor technologies and YOLOv3 reiterates the scope for improvements when algorithms encounter modified data.

Concrete assessment criteria are provided by accuracy metrics such as precision, recall, F1 score, and detection speed[10]. However, it's the ability to withstand real-world challenges, including occlusions, fluctuations in lighting, and variations in object size and color[11] that truly emphasizes the intricacies involved in accomplishing real-time object detection under modified conditions.

Research has aptly highlighted the biases in deep learning object detection algorithms due to factors such as imbalanced training data, representational limitations, or environmental conditions[12]. This emphasizes the crucial link between algorithmic biases and the accurate detection of objects, making it even more pertinent in the context of autonomous driving.

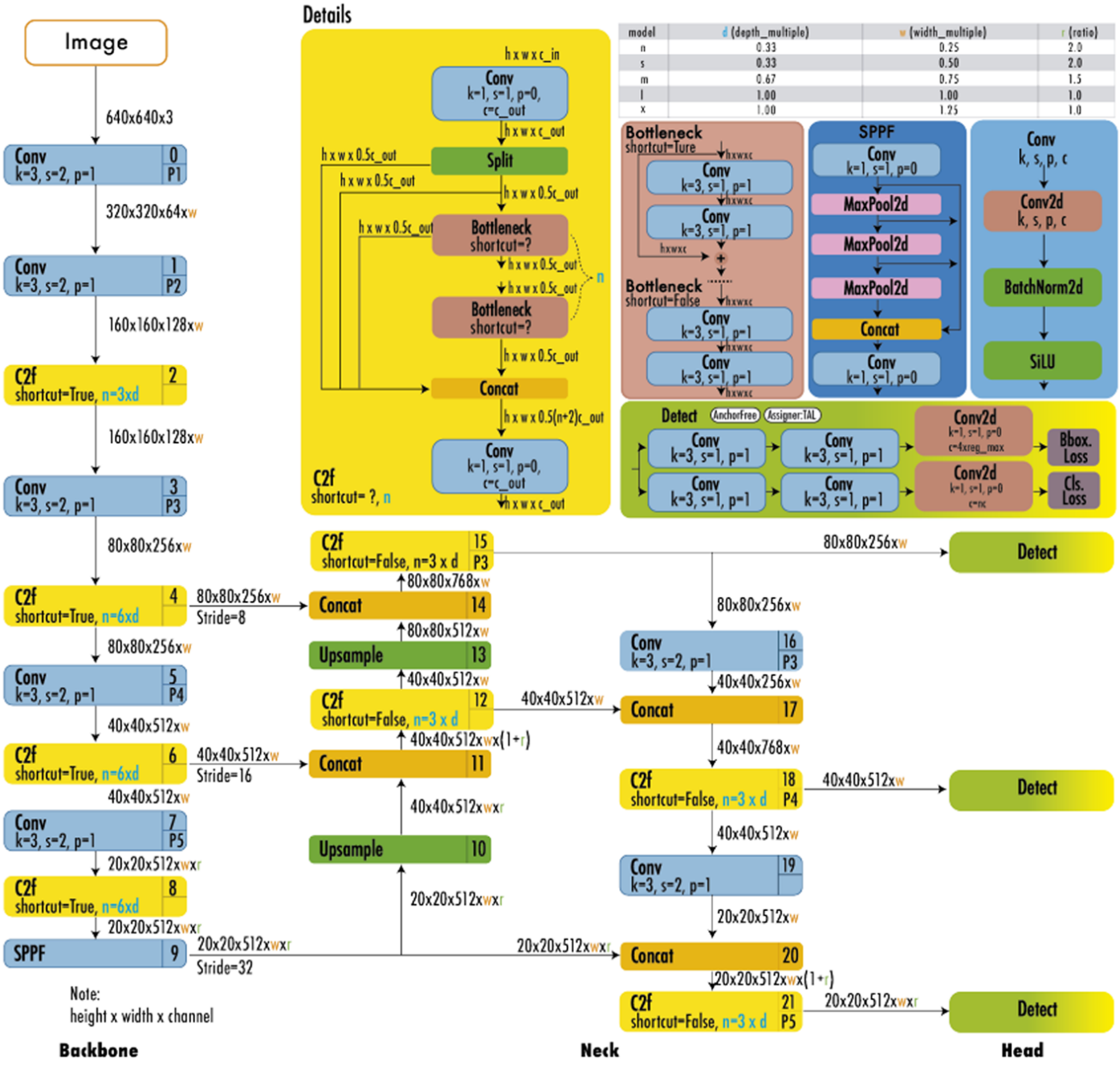

Fig. 3 YOLOv8 Architecture

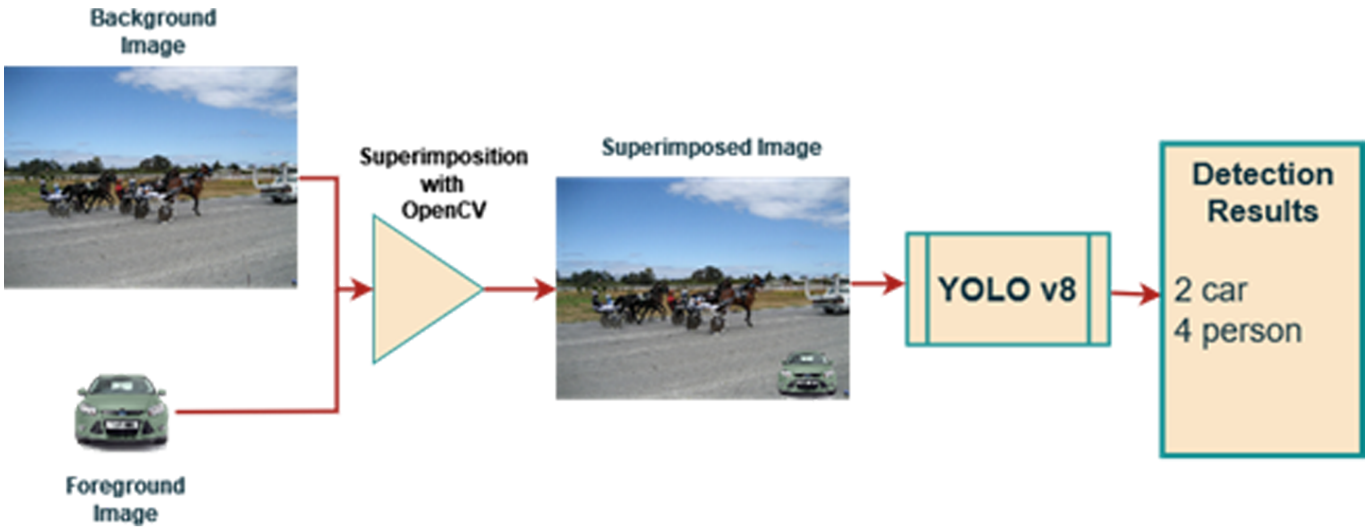

Fig. 4 Diagram illustrating image superimposition and object detection

Methods to enhance object detection algorithms vary, with some proposing two-stage methodologies[13] and others suggesting one-stage algorithms[14]. The use of intermediate layers and receptive field blocks for better detection indicates the ongoing innovations in this field, setting the stage for our exploration into the influence of modified data.

Thus, the literature underscores the gaps and questions that remain in the realm of object detection, particularly concerning altered data. This research aims to address these gaps, offering a meticulous analysis of algorithmic responses to challenges such as lighting variations, occlusions, and object manipulations, thereby contributing a fresh perspective to the broader discourse.

Results

Performance Comparison

Our study evaluated an object detection algorithm on three complexity levels of datasets: L1, L2, and L3. L1 sourced directly from the COCO (Common Objects in Context) dataset, provided standard scenarios, and resulted in an average confidence level of 0.864 across all classes. Confidence, in this context, refers to the model's certainty or level of confidence in its predictions. It is calculated as an average of the confidence scores assigned to each predicted object. Additionally, the preprocessing time (time for operations like image normalization and scaling) was 1.59 seconds, inference time (time for the model to make predictions) was 7.08 seconds, and post-processing time (time for class assignment, non-max suppression, and bounding box regression) was 1.47 seconds. The complexity of the L2 dataset, composed of 150 original and 300 modified images, increased the model's inference time to 9.95 seconds, demonstrating added computational demand. Nevertheless, the model maintained a high average confidence of 0.877, indicating resilience against moderate alterations. The L3 dataset, the most complex, reduced preprocessing time to 1.36 seconds due to fewer detected features, yet increased inference time to 7.26 seconds, suggesting challenges in feature extraction and pattern recognition. The average confidence level dropped to 0.823, signalling the model's difficulty in managing the unfamiliarity of automated modifications, thereby revealing a need for improved adaptability.

Several variables can be linked to the observed differences in Table 1. The shared nature of theplatform may result in varied preprocessing, inference, and post-processing times due to potential server congestion and resource allocation fluctuations when using Google Colab with GPU (T4) accelerators. The difference in confidence intervals between L1 and L2 can be attributed to the presence of 10 outlier photos in L1 that were not adequately eliminated. These outliers can have an impact on the model's overall confidence estimation.

Evaluation Metrics

The datasets L1, L2, and L3, each presenting varying degrees of complexity, have allowed for a meticulous study of the algorithm's proficiency. For dataset L1, which represents baseline, unmodified data, the model displayed commendable performance, evidenced by a mAP score of 79%, precision of 79.5%, and a recall of 75.3%. This indicates a high true positive rate and reliable object detection capacity un-

Dataset	Preprocess (seconds)	Inference (seconds)	Postprocess (seconds)	Confidence
L1	1.59	7.08	1.46	0.864
L2	1.31	9.95	1.38	0.877
L3	1.35	7.26	1.57	0.823

Table 1 Performance indicators across all datasets

Dataset	**Evaluation Metrics**		
	mAP(50-95)	Precision	Recall
L1	77.3%	79.5%	75.3%
L2	59.7%	79.3%	70.2%
L3	38.5%	71.4%	52.4%

Table 2 Scores across all Datasets

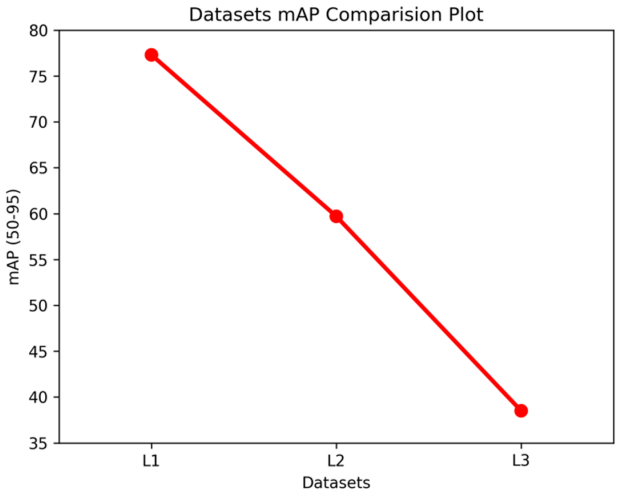

Fig. 5 mAP(50-95) visualization of datasets

der standardized conditions, serving as a benchmark for the subsequent datasets. Conversely, datasets L2 and L3 introduced incremental complexity and modifications, thereby testing the model's adaptability. With dataset L2, which possesses moderate complexity, the model's performance somewhat faltered but remained commendable. The mean average precision (mAP) was 59.7%, precision was 79.3%, and recall was 70.2% as illustrated in Figure 5. Despite a reduction from the baseline L1 dataset, these metrics demonstrate the model's resilience in the face of increased complexity.

However, with the escalation of complexity in dataset L3, a significant dip in performance was observed. The mAP fell to 38.5%, precision reduced to 71.4%, and recall dropped to a low of 52.4% (Refer to Table 2). These figures illustrate that as the dataset modifications became increasingly complex and unfamiliar, the model's performance suffered, indicating difficulties in generalizing to unseen or highly varied scenarios. Comparatively, while the model's performance in a standardized setting (L1) aligns reasonably well with YOLOv8 benchmarks from the literature [15], which typically achieves an AP score of 54%, the model slightly underperforms for more complex datasets (L3). This finding underscores the need for further optimization to improve the model's capacity to handle increased complexity and unfamiliarity, ensuring broader real-world applicability. The Precision scores in the L3 dataset exhibit an interesting trend as confidence levels increase illustrated in Figure 6. On average, Precision improves and reaches its peak at a confidence level of 0.925, with an average Precision score of 1. However, an anomaly is observed in the train class, whereprecision decreases with increasing confidence. This discrepancy is prominent in the L3 dataset, which contains primarily roadside images, making the presence of trains unlikely.In contrast, the L1 dataset shows a more consistent pattern, with Precision increasing proportionally as confidence levels rise without any noticeable outliers. In addition, L2 and L3 datasets also saw a decrease in precision at certain confidence intervals suggesting incorrect predictions made by the model, possibly influenced by ambiguous images, limited generalization, or algorithmic biases.For L2, the average precision curve was slightly smoother than L3 in the beginning, however, toward the end, it became more variable, though not as smooth as L1, suggesting that the model faced difficulties in object detection. Overall, the comparison between L1 and L3 datasets highlights the influence of dataset composition and modification on classifier performance. The balanced and less complex L1 dataset facilitates smoother learning and more consistent predictions, while the variability and complexity in the L3 dataset led to anomalies. These observations underscore the importance of thoughtful dataset composition and manipulation to optimize performance metrics like Precision.

Figure 7 portrays an intriguing trend concerning the F1 score as confidence levels increase within the L3 dataset. The initial growth in confidence levels sees an improvement in the F1 score across all classes, peaking around a score of 0.55. Surprisingly, subsequent boosts in confidence initiate a decline in the F1 score. As the model's confidence increases, it grows increasingly sure about its predictions, which enhances both precision and recall, reflected in the initial rise of the F1 score. However, the latter drop in the F1 score, despite mounting confidence levels, can be attributed to a decline in recall,

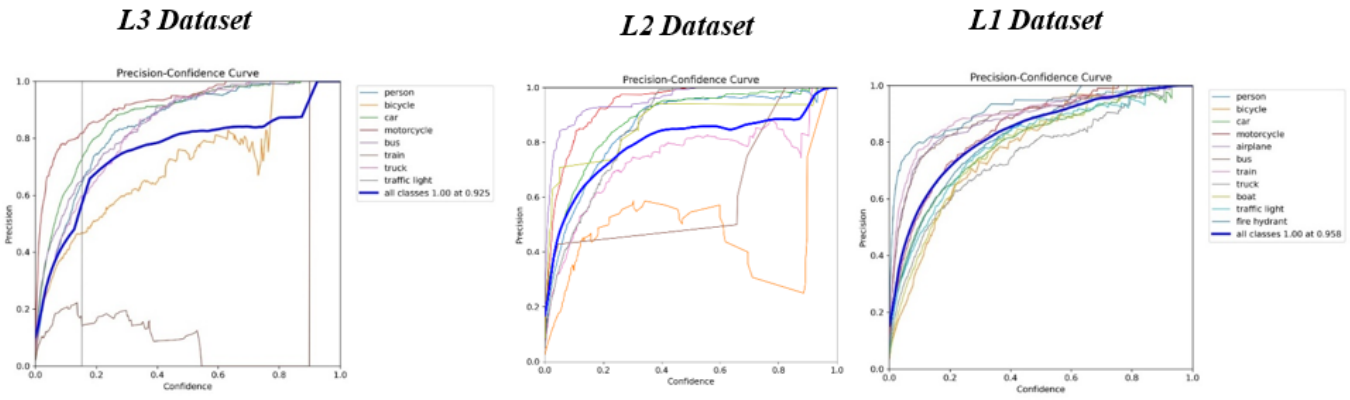

Fig. 6 Precision vs Confidence Curves

even though precision might still be improving. This occurs because, with heightened confidence, the model tends to predict fewer instances to minimize errors. Consequently, while it notes fewer false positives, improving precision, it may miss correctly identifying all positive instances, leading to a fall in recall and consequently the F1 score.

In comparing the L3 dataset to the L1 dataset, the L1 dataset shows a more consistent growth in the F1 score with increasing confidence levels, reaching a higher average F1 score of 0.74. Similarly, when comparing L2 to L3, although L2 has a higher average F1 score, it lacks the consistency found in L1, indicating that while it may perform better on average, it has difficulties in maintaining consistent performance across all classes, unlike the stable behavior observed in L1.

Moreover, an important aspect to address is the anomaly observed in the confidence scores for bicycle detection in the L2 dataset, along with an unexpected pattern in the motorcycle detection curve. The bicycle curve is notably separated from the other curves, showcasing a steady downward trend. This could potentially indicate that the model struggles to accurately detect bicycles as the confidence level increases, possibly due to overlapping features with other classes or insufficient training data. This trend suggests that increased confidence levels may not always result in better detection, as illustrated by the bicycle curve, affecting the overall reliability of the model.

Furthermore, the motorcycle curve displays a unique pattern, initially presenting a linear increase till an F1 score of about 0.45, followed by a sudden jump to 0.7, and an abrupt fall to zero thereafter. This erratic behavior might be due to a variety of reasons such as overfitting to specific features in the training data, or it might be experiencing difficulties in distinguishing between closely related classes at higher confidence levels. The sudden drop to zero indicates that at a certain point, the model fails entirely to detect motorcycles, pointing to a significant limitation in the model's ability to generalize well across different confidence thresholds. This anomaly warrants a deeper investigation into the model's learning patterns to improve its performance and reliability further.

Visualizations and Examples

Upon examining the plot on the right, a distinctive pattern is evident in the performance of the model across different datasets as illustrated in Figure 7. In the context of the L1 dataset, consisting of 500 images, the model exhibits remarkable precision, with the bulk of confidence levels for predictions situated within the range of 0.75 to 1. Anomalies at 0.0, albeit sparse, are disregarded when computing the mean confidence level, which stands at an impressive 0.864.Contrastingly, as the model navigates through the L2 dataset, there is an observable decline in confidence interval, descending to 0.843. The descent is even more prominent when dealing with the most intricate L3 dataset, where the confidence interval plummets to a mere 0. 823.These observations underline the nuanced interaction between the model's performance and the complexity of the dataset. This analysis offers pivotal insights into the robustness of the model and its adaptability to variations in dataset complexities, pivotal elements to consider in the broader discourse of computer science research.

Discussion

Discussion of Findings

Our study reveals the responsiveness of YOLOv8 to varied data. On the L1 dataset, the model achieved an mAP of 77.3%. However, this declined to 59.7% on L2 and further dropped to 38.5% on the more intricate L3 dataset. These results confirm that complex modifications challenge real-time object de-

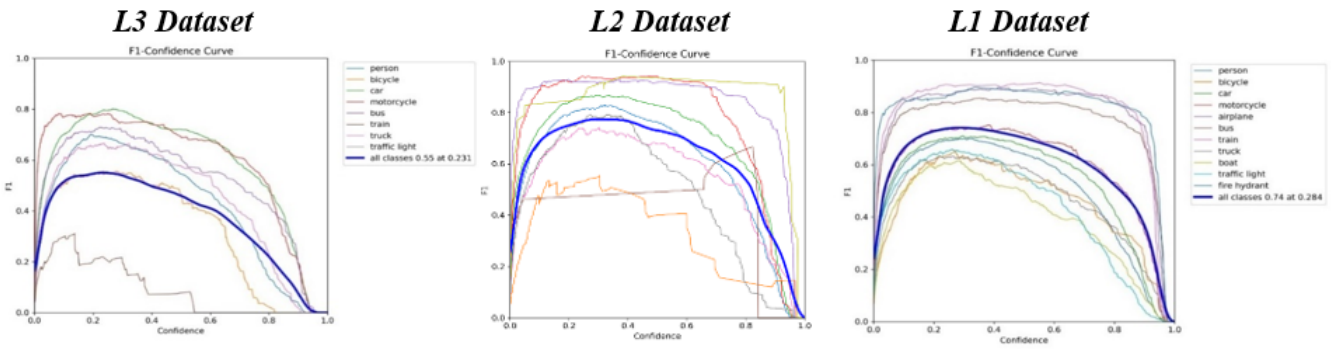

Fig. 7 F1 vs Confidence Curve

Fig. 8 Confidence plots for all

tection. These findings support our hypothesis that complex modifications can disrupt real-time object detection performance. The detection bias in became apparent with smaller objects or those with unique colours, indicating a preference for dominant image features. Figure 9 exposes a bias in the object detection algorithm, resulting in the omission of smaller or less common objects and misclassification of a flipped car as an airplane, highlighting poor performance in unfamiliar scenarios.

Considerations for Real-World Applications

The study's findings emphasize the need for careful consideration of biases in real-world autonomous driving applications using YOLOv8 or similar models. Training should account for biases by including diverse objects and backgrounds. Regular testing and updates can address emerging scenarios, and sensor fusion can improve detection accuracy. These insights contribute to ongoing efforts to enhance the safety and effectiveness of autonomous driving technologies.

Ethical and Safety Implications

Throughout the research, adherence to ethical standards was paramount. To ensure privacy, we exclusively used publicly available image sets, such as the COCO dataset, generated by recognized authorities in the field for the explicit purpose of training machine learning models. Consequently, there was no privacy concerns associated with data usage.

Comparison of Results With Other Studies

To comprehensively assess the contributions of our study, it is valuable to compare our findings with existing research reports. For instance, we encountered a study[16] documented by Wang et al., which aimed to enhance the average precision (AP) scores of YOLOv4 through specific model training techniques. The outcome of this effort was a notable improvement, with the highest achieved AP score reaching 62.7%. This finding underlines the fact that even YOLOv4, a preceding version of the YOLO model, demonstrated superior accuracy compared to the mAP score of 38.5% we recorded in our study for the L3 dataset, underscoring our conclusion that the mAP score attained in our research falls below established standards when certain modifications are made to the image set. Additionally, we observed another research report[17] that introduced a novel approach incorporating a multi-scale MobileNeck module and an algorithm tailored to enhance object detection model performance by generating a series of Gaussian parameters. This exploration yielded a mAP score of 80.5% for YOLOv3. This high-performance benchmark serves as a reminder of the rigorous standards that several research initiatives have managed to uphold within the field. It's noteworthy that, despite the optimization efforts undertaken by these studies, our evaluation of the YOLOv8 model revealed subpar performance, particularly with the L2 and L3 datasets, reinforcing the scope of further improvements to be made in order to maintain the high mAP scores across varied conditions.

Limitations of the Study and Solutions

While our study offers valuable insights into the performance of YOLOv8 across varied datasets and its implications for real-world applications, a few limitations should be acknowledged. These limitations, while not diminishing the significance of our findings, provide avenues for further research and improvement.

1. Data Diversity and Volume

One notable limitation of our study lies in the diversity and volume of the datasets used for evaluation. While we utilized publicly available datasets, such as COCO, which are widely recognized in the field, the representation of certain scenarios and object types might still be limited. This could potentially affect the generalizability of our conclusions to a broader spectrum of real-world situations. To overcome this obstacle, future studies could benefit from incorporating more diverse datasets that encompass a wider range of scenarios, lighting conditions, and object scales to ensure a more comprehensive assessment of model performance.

2. Computational Capacity and Image Set Constraints The computational limitations of our system played a role in determining the scope of our image sets and the size of the datasets used for experimentation. This could potentially lead to the omission of certain challenging scenarios that require larger computational resources. A solution to this could be the inclusion of more complex scenarios and a larger variety of objects might offer a more nuanced understanding of the model's behavior and its real-world applicability. Collaborative efforts that pool computational resources could aid in addressing this limitation.

3. Model Hyperparameters and Optimization The present study primarily focused on the default configuration and hyperparameters of YOLOv8. While this choice provides a baseline assessment, it is important to acknowledge that model performance could be further enhanced through hyperparameter tuning and optimization. To bridge this limitation, it would be advisable to Explore a wider range of hyperparameters, consider different anchor box sizes, or experiment with various training techniques to improve detection accuracy.

Future Directions and Research Opportunities Future studies should focus on larger datasets and diverse detection models for comprehensive comparisons. Algorithm im-

Fig. 9 Comparative Illustration of Object Segmentation and Detection: Before and After Modification Implementation

provements through techniques like adversarial training, data augmentation, and transfer learning can address identified biases. Utilizing real-world data is going to provide insights into model performance in unpredictable scenarios. Integration of different sensors and considerations for AI ethics and fairness are also crucial. Furthermore, Table 2 presents a notable trend of high precision accompanied by relatively lower recall scores across the evaluated datasets. In scenarios where safety is crucial, such as self-driving cars, high precision is desirable to prevent unnecessary interventions or abrupt reactions caused by false alarms. However, the challenge lies in maintaining a balance between precision and recall. While high precision signifies a low rate of false positive predictions, the trade-off with lower recall implies that a considerable number of actual objects are not detected, possibly leading to hazardous situations. For instance, failing to identify pedestrians or obstacles could result in accidents or unsafe driving conditions. Several factors contribute to this observed trend, however fine-tuning the model through confidence threshold adjustments and ensemble methods can mitigate the precision-recall trade-off. Augmenting the model's exposure to varied scenarios using domain-specific knowledge and data augmentation can enhance generalization without compromising precision. Future research should focus on strategies for harmonizing high precision and recall. Techniques like class-specific optimization, multi-task learning, and transfer learning hold promise. Additionally, adapting algorithmic decisions to specific application contexts, while acknowledging dynamic real-world environments, can foster balanced and effective object detection systems for critical applications like autonomous driving.

Methodology

This section investigates the effect of data modifications on autonomous driving object detection, using the COCO dataset and YOLOv8 architecture. The adaptable methodology includes three datasets (L1, L2, and L3) for varying alteration levels. Images were annotated with Pixlr for consistent bounding box definition and resolution. Evaluation relied on metrics like mean average precision (mAP) to gauge model accuracy and efficacy.

Dataset Collection and PreparationThe research study involves applying a pretrained model to 3 modified datasets. The baseline dataset used was the COCO dataset which involves more than 330,000 pictures, and each one is labelled with one of 80 item categories and five descriptions. The COCO dataset is widely employed in the field of computer vision research, serving as a valuable resource for training and assessing advanced algorithms related to object detection and segmentation. This dataset has played a crucial role in the development and evaluation of state-of-the-art techniques. Given its significance, we have chosen the COCO dataset as a foundational baseline for our study. We intend to compare its performance against two additional datasets, namely L2 and L3. In order to enhance the comprehensiveness of these datasets, we have introduced modifications. This involves augmenting the existing COCO dataset images with new images and incorporating external data sources to form entirely new images. These alterations involve superimposing distinct foreground images onto the original background images. To generate a comprehensive dataset, we incorporate multiple source images for each object, captured from various perspectives. This selection encompasses images of bicycles, pedestrians, cars, and cycles, which are subsequently superimposed onto the entire dataset, thereby generating a fresh set of images. This methodology follows an automatic annotation process that relies on the information contained in the source image filenames, hence, every source image is uniquely named following a combination of its categorical data and a distinct numeric identifier. This process creates a flow of data from the source images to the synthetic image, providing an efficient and automated method for dataset preparation and annotation.

The initial phase of the study focused on the selection and preparation of distinct image sets to provide varying levels of complexity:

1. **L1**: 500 images were curated from the COCO dataset on which the model was trained. This dataset encompassed a broad variety of scenarios the model would typically encounter, effectively representing 'normal' conditions. Images in this set were unaltered, serving as a control group a baseline for the model's performance.

2. **L2**: This level introduced manual modifications to increase complexity. Starting with 150 images from the original dataset, each image was manually transformed by integrating various foreground objects. The selected objects included regular items such as cars, bikes, and trucks, as well as unexpected elements, resulting in approximately 450 modified images.

3. **L3**: The highest complexity level incorporated 100 unique images, not included in the model's training dataset. These selected images underwent considerable modifications using automation, including the addition of unexpected foreground elements and novel object transformations such as scaling and rotation. With 5 such automated modifications per image, approximately 500 distinct images were generated.

Level / Complexity	Original Image Set	Total Image Set Size (including altered images)
Level 1	500	500
Level 2	120	450
Level 3	100	500

Table 3 The three-tier approach

Research Motivation and Design

The study aims to evaluate how a machine vision algorithm reacts to manipulated data and if any inherent biases are persistent in its object detection. To this end, a three-tier category of dataset, labelled as L1, L2, and L3, was devised based on image generation complexity (Refer to Table 3).

Image Annotation and Management

Image annotations were performed using Pixlr to define bounding boxes around intended areas in the background image. The corresponding foreground images were mapped to these areas, and their coordinates noted. The annotations were maintained to track specific locations and notes for each image. In the image augmentation stage, each foreground object is segmented, resized, and overlaid onto a background image, creating an augmented image. The images are then stored in the Augmented Images set. In the object detection stage, the chosen detection algorithm is applied to each image in Augmented Images, generating a set of detection results, Detection Results. This two-stage process iteratively augments image data and applies object detection to the augmented images, allowing for a versatile examination of machine vision algorithms in varying image contexts.

Modified Images

Accurate object masking is essential for creating convincing synthetic images. In our study, we utilized YOLOv8's latest segmentation capabilities for precise object masking and accurate bounding boxes. This, coupled with YOLOv8's feature of assigning confidence levels to each detection (as shown in Figure 11), allows a reliable assessment of the model's performance across diverse datasets.

In Figure 11, it is evident that distinct detection boxes are generated, segmented, and classified according to their identified object types, facilitated by the YOLOv8m.pt model.

Evaluation Metrics

To assess the performance of the object detection model, several evaluation metrics were utilized. The model's performance assessment hinges on four core metrics: mean average precision (mAP), precision, F1 score, and recall. The mAP score considers both detection accuracy and quality by measuring the overlap between predicted and actual object locations using Intersection over Union (IoU). It provides a nuanced view, accounting for partial detections and assigning lower scores for incomplete detections. Precision represents the proportion of correct predictions for a specific object class, capturing the model's ability to correctly identify objects. Recall, on the other hand, measures the proportion of correctly detected objects out of all actual objects, highlighting the model's sensitivity in identifying all instances of the object class. To calculate precision and recall, a confusion matrix is generated by comparing the model's predictions against the ground truth labels. This matrix provides a comprehensive view of true positives, false positives, and false negatives, enabling the calculation of precision and recall for each object class. It's important to note that while this discussion emphasizes the example of cars, these metrics are applicable to multiple object classes, providing a comprehensive evaluation of the model's performance across different categories.

Conclusion

In conclusion, this investigation examined the performance characteristics of the YOLOv8 object detection algorithm, unravelling its efficacy in conventional tasks and shedding light on its constraints in intricate scenarios. Our study offers a comprehensive overview of the algorithm's capabilities and limitations, serving as a compass for its practical application. It is imperative to acknowledge that every technological advancement is driven by the pursuit of overcoming challenges. Yet, as with any technology, the broadening adoption of object detection necessitates a thorough understanding of its limitations and the corresponding strategies for improvement. To this end, our study identifies several key limitations that merit attention. The restricted diversity and volume of training data, constrained by computational capacity, pose a potential limitation to the model's adaptability. Addressing this could involve collaborating with multiple data sources and leveraging distributed computing to enable access to more comprehensive datasets. Another critical challenge lies in maintaining a balance between precision and recall, particularly in safety-critical scenarios like autonomous driving. While high precision minimizes false alarms, lower recall can result in missed detections of crucial objects, introducing potential hazards.

Fig. 10 Sample from modified and baseline dataset of images

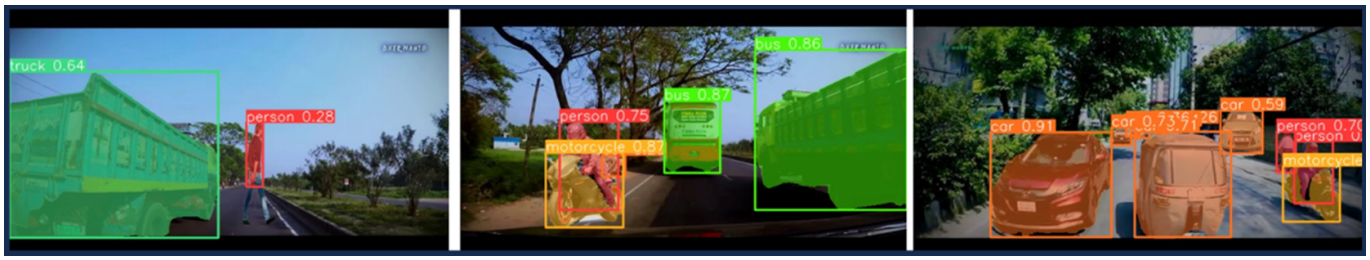

Fig. 11 Sample from modified and baseline dataset of images

Tackling this challenge involves fine-tuning the model, exploring ensemble methods, and incorporating domain-specific knowledge and data augmentation techniques. The results underscore the significance of refining the algorithm to achieve universal applicability. As the technology advances towards real-world deployment, the spotlight on issues of bias, generalizability, and robustness becomes more pronounced. This study has highlighted the imperative of not only refining the algorithm's technical aspects but also ensuring its ethical and practical facets are aligned. Beyond its immediate implications, our findings hold wider relevance for the field of object detection and autonomous systems. As these technologies evolve and integrate into various sectors, including transportation, surveillance, and manufacturing, the insights gained from this study can drive further innovation. The journey to enhance these models is ongoing, and our research opens the gateway to promising avenues for future investigation. In essence, the YOLOv8 algorithm encapsulates both potential and challenges. By acknowledging the need for addressing challenges alongside highlighting advantages, we hope to contribute not only to the advancements of object detection technologies but also to the establishment of a solid foundation for their responsible and effective integration into our rapidly evolving world.

Acknowledgments

I wish to express my profound gratitude to my mentor, Dr. Joe Isaacs. The invaluable insights and guidance he provided have been pivotal in steering this research project to fruition.

References

1 S. Lahange, P. Nalawade, P. Bide, D. Nayak and A. Mohite, *Computer Vision Techniques in Autonomous Vehicles: A Survey*.

2 P. Soviany and R. T. Ionescu, *Optimizing the Trade-off between Single-Stage and Two-Stage Object Detectors using Image Difficulty Prediction*, arXiv: 1803.08707 (2018).

3 M. Carranza-García, J. Torres-Mateo, P. Lara-Benítez and J. García-Gutiérrez, *Remote Sens*, **13**, 89.

4 A. Hawkins, *The Verge*.

5 N. Olson, S. Lund, R. Colman, J. Foster, J. Sahl, J. Schupp, P. Keim, J. Morrow, M. Salit and J. Zook, *Frontiers in genetics*, **6**, 235.

6 J. Terven and D. Cordova-Esparaza, *FROM YOLOV1 TO YOLOV8 AND BEYOND UNDER REVIEW IN ACM COMPUTING SURVEYS*.

7 L. Peng, H. Wang and J. Li, *Automot. Innov*, **4**, 241–252.

8 S. Iftikhar, Z. Zhang, M. Asim, A. Muthanna, A. Koucheryavy and A. El-Latif, *Electronics*, **11**, 3551.

9 X. Li, W. Wang, L. Bo, T. Wu, Y. Yuan, Y. Lu and Y. Chen, *Dynamic R-CNN: Towards High Quality Object Detection via Dynamic Training*, arxiv:.

10 C. Peng, T. Xiao, Z. Li, Y. Jiang, X. Zhang, K. Jia, G. Yu and J. Sun, Proceedings of the IEEE conference on computer vision and pattern recognition, p. 6181–6189.

11 M. Ahmed, K. Hashmi, A. Pagani, M. Liwicki, D. Stricker and M. Afzal, *Sensors*, **21**, year.

12 S. Ren, K. He, R. Girshick and J. Sun, *Advances in neural information processing systems*.

13 J. Redmon, S. Divvala, R. Girshick and A. Farhadi, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR, p. 779–788.

14 Y. Zhou, Q. Ye, Q. Qiu and J. Jiao, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

15 J. Terven and D. Cordova-Esparza, *A Comprehensive Review of YOLO: From YOLOv1 and Beyond*, arXiv:.

16 X. Wang, L. Kong, Z. Zhang, H. Wang and X. Lu, *ACP-YOLO: Asymmetric Center Point Bounding Box Regression Strategy and Angle Loss-based YOLO for Object Detection*.

17 W. Li and K. Liu, *Sensors*, **21**, year.