Truck and Bus Detection Using YOLOv5

Kelvyn Lukito
School of Computing
Telkom University
Bandung, Indonesia
lukitokelvyn@gmail.com

***Abstract*— The rapid urbanization and industrial expansion in recent years have resulted in an unprecedented surge in vehicular traffic, particularly the movement of heavy trucks. While these trucks play a crucial role in transporting goods, their dense traffic patterns pose significant challenges to infrastructure, leading to accelerated wear and tear of road surfaces. The excessive load and frequency of heavy trucks contribute to increased road degradation, raising concerns about safety and escalating maintenance costs. Recognizing the urgent need to address these challenges, this research focuses on implementing the YOLOv5 (You Only Look Once version 5) model for efficient monitoring and control of heavy truck traffic. YOLOv5, a state-of-the-art object detection algorithm, has demonstrated remarkable accuracy and speed in real-time object detection tasks. Leveraging its capabilities, the study aims to accurately quantify the traffic volume of heavy trucks, facilitating a data-driven approach to road maintenance and traffic management. The research involves a tuning process to identify the optimal combination of parameters for the YOLOv5 model. The results indicate that the tuned model achieves precision values of 0.84 and 0.99 for bus and truck classes, respectively, and recall values of 0.84 and 1 for bus and truck classes, respectively. The mean Average Precision (mAP) is recorded at 0.97, and the F1-score reaches 0.92. These findings demonstrate the effectiveness of the YOLOv5 model in accurately detecting and classifying heavy trucks, providing a robust foundation for data-driven decision-making in road maintenance and traffic management strategies.**

Keywords—Object Detection, Vehicle Detection, Traffic Flow Monitoring, YOLOv5

I. INTRODUCTION

The rapid growth of urbanization and industrial activities has led to an unprecedented surge in vehicular traffic, particularly the movement of heavy trucks. While these trucks play a pivotal role in transporting goods and materials, their dense traffic patterns have begun to pose significant challenges to infrastructure, notably the wear and tear of road surfaces. The excessive load and frequency of heavy trucks contribute to accelerated road degradation, leading to increased maintenance costs and safety concerns.

Recognizing the urgent need to address these challenges, this research focuses on the implementation of the YOLOv5 (You Only Look Once version 5) model for efficient monitoring and control of heavy truck traffic. YOLOv5, a state-of-the-art object detection algorithm, has demonstrated remarkable accuracy and speed in real-time object detection tasks [1]. By leveraging its capabilities, we aim to accurately quantify the traffic volume of heavy trucks, allowing for a data-driven approach to road maintenance and traffic management.

In a study entitled Vehicle Detection and Classification using YOLOv5 on Fused Infrared and Visible Images, researchers used the YOLOv5 model that had been fused to reduce the impact of dataset imbalance. in the study it was found that the model produced performance with a sensitivity benchmark with a value of 0.97, accuracy with a value of 0.96, sensitivity with a value of 0.97 and a precision value with a value of 0.93[1].

In a study entitled Vehicle Classification Application on Video Using Yolov5 Architecture, researchers implemented the YOLOv5 model to overcome the problem of high levels of congestion and accidents. With a dataset in the form of a traffic vehicle video with 750 vehicle images collected using an outdoor surveillance camera on the road to train the model, the model performance is obtained with a benchmark accuracy value of 89% [2].

In a study entitled Vehicle Tracking Method Based on Attention-YOLOv5 and Optimized DeepSort Models, researchers compared two object detection models, namely YOLOv5 and Optimized DeepSort in tracking vehicles. By conducting experiments in the form of data collection of several highway traffic videos from various scenes. The researchers conducted visual analysis, then conducted ablation experiments, and evaluated using MOTA, MOTP and ID-Switch, concluding that the ECA-YOLOv5 model had the highest detection accuracy [3].

In a study entitled Real-Time Vehicle and Distance Detection Based on Improved Yolo v5 Network, researchers implemented the YOLOv5 model in a distance detection system. This research has the problem of many unsafe factors on the highway and it is the most crucial aspect in automatic driving technology. In this study, experiments were conducted in a virtual environment and the experimental results obtained Yolo v5s detection accuracy was 83.36% mAP (Average Precision), detection speed 28.57FPS (Frame Per Second), and Yolo v5-Ghost detection accuracy was 80.76% mAP, detection speed 47.62FPS [4].

In the research entitled Road Condition Detection Based on Deep Learning YOLOv5 Network, researchers implemented the YOLOv5 model to detect road conditions in automated vehicle systems. Based on the problems in detecting road cracks, vehicles and the imbalance of datasets, researchers modified the YOLOv5 model to be refined. Experimental results show that the enhanced model achieves a mean average precision (mAP) of 64.5%, compared to 62% for the original YOLOv5 model, while maintaining real-time performance at 150 FPS. Precision and recall also improved, from 69.3% to 71.4% and from 54.5% to 55.6%, respectively [5].

In a study entitled A Computer Vision based Vehicle Counting and Speed Detection System, researchers implemented the YOLOv5 model to perform vehicle detection and the StrongSORT algorithm which is a development algorithm from DeepSORT to perform vehicle detection simultaneously on each video frame. The experimental results gave an accuracy value of 85.27% for vehicle detection. The accuracy for vehicle speed was 87.9% with marginal room for error from the ground truth value. In addition, the model works well in terms of counting the number of vehicles [6].

With the justification explained using previous research, a solution that can be proposed is to use YOLOv5 to detect heavy trucks and bus on the highway in real-time. Through this detection process, we seek to obtain precise information on the number and movement patterns of these vehicles. Subsequently, the data collected will be used to formulate operational restrictions on the number of trucks allowed to ply a particular road segment. These strategic restrictions aim to reduce the adverse impact of heavy truck traffic on road conditions, as well as encourage sustainable and cost-effective road maintenance practices.

II. METHODOLOGY

A. Dataset

The dataset utilized in this research focus comprises a collection of vehicle images, totaling 153 pictures distributed across four classes. These classes include heavy vehicles such as trucks or buses, light four-wheeled vehicles such as cars or minibuses, and light two-wheeled vehicles such as motorcycles based on figure 1. The set of image data is divided into two subsets: a training set and a validation set, with proportions of 76% and 24% [7].

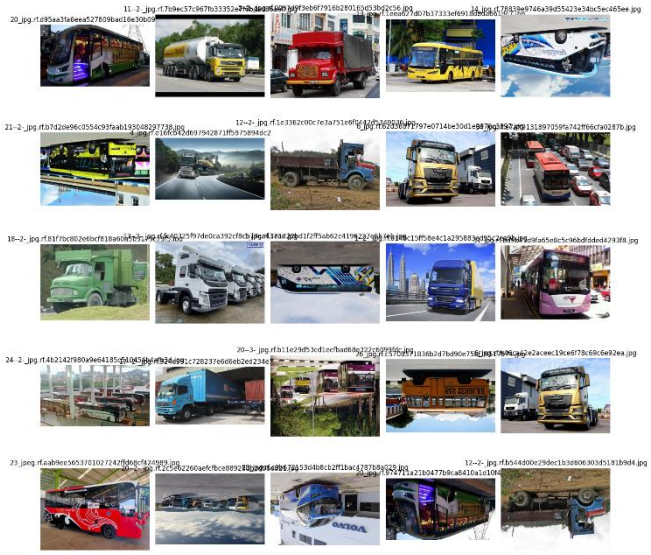

Fig. 1. Example of Dataset Truck and Bus Image

B. YOLOv5 (You Only Look One Version 5)

Ultralytics introduced YOLOv5 in May of this year, boasting superior detection speed and accuracy compared to both YOLOv3 and the earlier YOLOv4 algorithms on the COCO dataset. YOLOv5s has a data weight of 27 million, just 1/9th of the YOLOv4 version, enabling faster detection. According to official information, it can process images at an impressive speed of 0.007s, facilitating real-time detection. Moreover, it is well-suited for deployment in onboard devices with limited computing, memory, and energy consumption requirements, making it ideal for tasks such as implementing training templates for monitoring equipment at construction sites [8].

YOLOv5, a cutting-edge object detection algorithm, approaches its task by meticulously dividing the input image into a grid-like structure. Within this grid, each individual cell carries the responsibility of pinpointing objects by producing bounding boxes that encapsulate them and assigning appropriate class labels to those objects. To gauge the level of certainty associated with each detected object, YOLOv5 effectively employs confidence scores. It then strategically utilizes a process known as Non-Maximum Suppression to minimize any overlapping predictions and meticulously select the bounding box that exhibits the highest degree of confidence. The culmination of this process yields a comprehensive list of the detected objects, along with their precise locations within the image and their corresponding confidence scores. Prior to its deployment, YOLOv5 necessitates a thorough training phase that involves a dataset comprised of meticulously annotated images. A hallmark advantage of YOLOv5 lies in its remarkable capability to execute object detection in real-time, rendering it a highly sought-after choice for applications that necessitate swift responses. YOLOv5 perpetually undergoes continuous evolution and refinement, with the overarching objective of augmenting its performance in the realm of object detection across a vast spectrum of computer vision applications.

YOLOv5 builds upon the YOLO detection architecture and leverages a collection of recent advancements in convolutional neural networks to achieve optimal performance. These advancements include techniques such as auto learning bounding box anchors, mosaic data augmentation, and the cross-stage partial network, each contributing distinct functionalities within the YOLOv5 architecture.

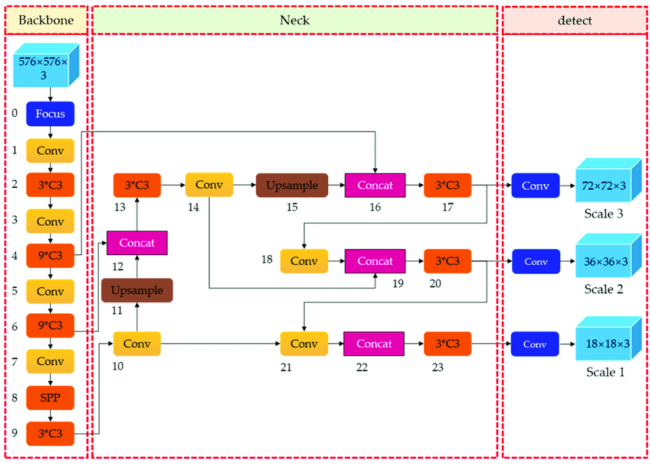

Fig. 2. Architecture of YOLOv5 [9]

YOLOv5's structural blueprint encompasses four primary components such as input, backbone, neck, and output. Input stage primarily handles data preprocessing, incorporating actions like mosaic data augmentation [10] and adaptive image filling. To ensure adaptability to diverse datasets, YOLOv5 embeds adaptive anchor frame calculation within the input stage, enabling automatic adjustment of initial anchor frame size based on dataset variations. backbone network predominantly employs a cross-stage partial network (CSP) [11] and spatial pyramid pooling (SPP) [12] to extract feature maps of varying sizes from the input image through a series of convolutional and pooling operations. BottleneckCSP acts to reduce computational overhead and accelerate inference speed, while the SPP structure facilitates feature extraction across multiple scales from a single feature map, generating three-scale feature maps that bolster detection accuracy. Neck network, feature pyramid structures known as FPN and PAN are utilized. The FPN structure [13] effectively propagates robust semantic features from higher-level feature maps to lower-level ones. Concurrently, the PAN structure

[14] transmits precise localization features from lower-level feature maps to higher-level ones. This collaborative interplay strengthens the features extracted from disparate network layers during Backbone fusion, further enhancing detection capabilities. The final stage, head output, primarily focuses on predicting targets of varying sizes upon the feature maps.

C. Model Evaluation

The evaluation model used in this study is to use the main benchmark, namely mean average precision value followed by several additional benchmarks such as sensitivity or recall, F1-Score and precision.

The sensitivity or recall value is one of the benchmark values which is the product of the confusion matrix diagram. The accuracy value can be calculated using a formula 1:

$$Recall\ /\ Sensitivity\ Score = \frac{TP}{TP+FN} \quad (1)$$

The F1-Score value is one of the benchmark values which is the product of the confusion matrix diagram. The accuracy value can be calculated using a formula 2:

$$F1\text{-}\ Score = 2\frac{Precision\ .\ Recall}{Precision + Recall} \quad (2)$$

The precision value is one of the benchmark values which is the product of the confusion matrix diagram. The accuracy value can be calculated using a formula 3:

$$Precision\ Score = \frac{TP}{TP+FP} \quad (3)$$

mAP stands for Mean Average Precision, and is a widely used metric to evaluate the performance of object detection models. Mean Average Precision is typically used in the context of tasks where the goal is to detect and localize objects within an image. Object detection involves predicting bounding boxes and associated class labels for objects present in the image. In generating the mAP value, it is necessary to calculate the average precision (AP) value in formula 4:

$$AP\ Value = \int_0^1 precision(R)\ dR \quad (4)$$

R represent recall.

After calculating the AP value, the next step is to calculate the mAP value using formula 5:

$$mAP = \frac{AP1+AP2+\ ...+APn}{n} \quad (5)$$

n is the number of classes.

III. Experiment and Analysis

A. Experiment Process

TABLE I. TUNING PARAMETER

YOLO	Model Parameters		
	Image Size	*Batch Size*	*Epochs*
YOLO #1	650	10	50
YOLO #2	750	15	50
YOLO #3	850	20	50
YOLO #4	950	25	50

Table 1 is an experiment that will be carried out by performing all parameters with the aim of getting a combination of parameters that have the best performance. The combination of parameters will be done by testing the combination of image size, batch size and epoch parameters. Image size refers to the size of the input image given to the algorithm for the object detection process. Image size is usually measured in pixels and can affect the performance and accuracy of object detection. Batch size refers to the number of images or datasets processed together in one iteration during model training. In the context of training neural network models, which includes YOLO, the training data is divided into batches to speed up the learning process and take advantage of the parallelism of calculations provided by GPU-enabled hardware and to provide learning variance from the dataset or images used. Epoch refers to one iteration through the entire training dataset. When an epoch is complete, the model has seen and processed the entire training data once. The training process consists of several epochs. At each epoch, the model parameters (weights) are updated based on the calculated gradient value of the loss function over the training data. The goal of training is to optimize the model so that it can make accurate predictions against new data.

B. Experiment Results and Analysis

Based on the tuning process that has been carried out with several parameter combinations, the YOLO 1 model with the parameters image size 650, batch size 10 and epoch 50 is the parameter combination with the best performance based on the presentation in table 2 and the visualization of the F1-Confidence Curve in figure 3.

TABLE II. TUNING RESULTS

YOLO	Model Parameters					
	Precision		*Recall*		*mAP*	*F1-Score*
	Bus	*Truck*	*Bus*	*Truck*		*All Classess*
1	0.84	0.99	0.84	1	0.97	0.92
2	0.97	0.83	0.78	0.91	0.94	0.87
3	0.97	0.79	0.84	0.94	0.92	0.89
4	0.89	0.77	0.85	0.75	0.84	0.82

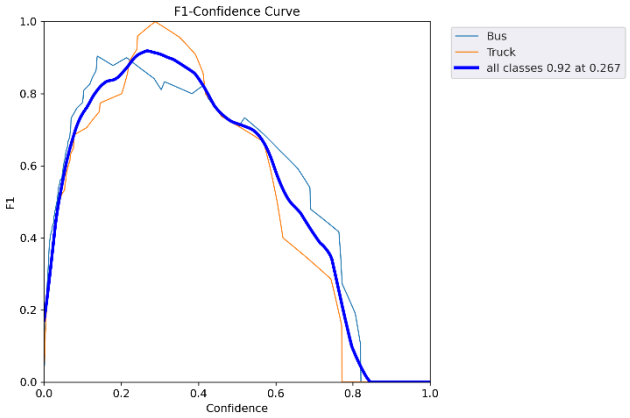

Fig. 3. F1-Score Curve YOLO 1

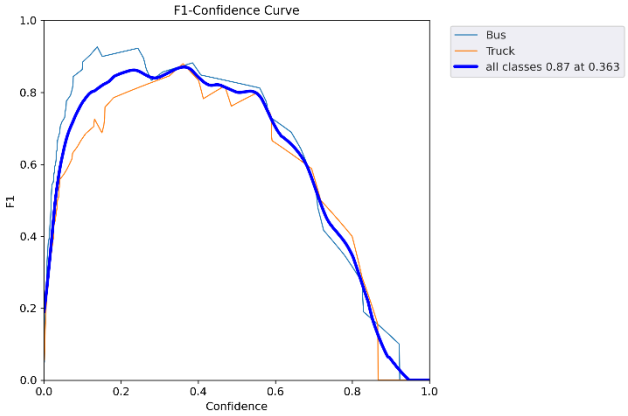

Fig. 4. F1-Score Curve YOLO 2

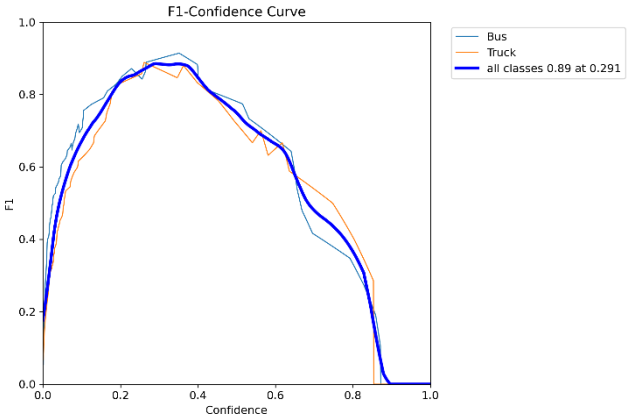

Fig. 5. F1-Score Curve YOLO 3

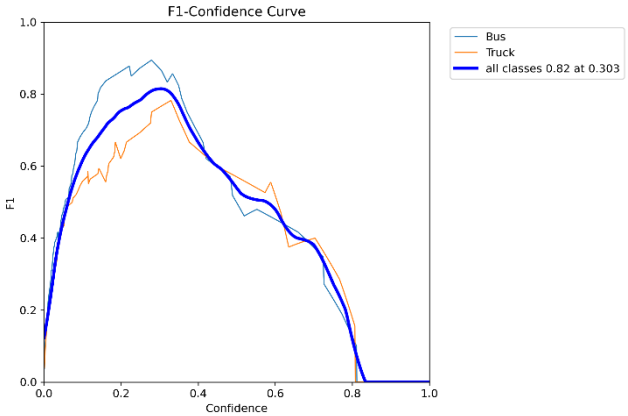

Fig. 6. F1-Score Curve YOLO 4

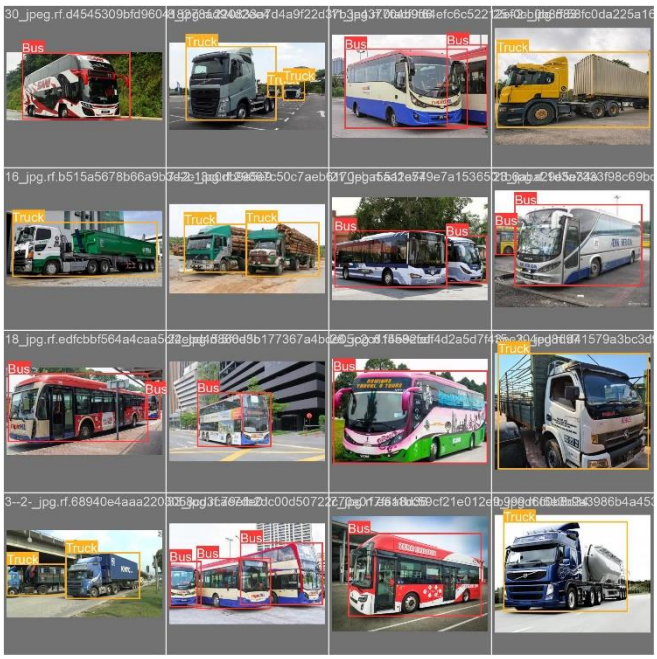

Fig. 7. Actual Object Detection Label

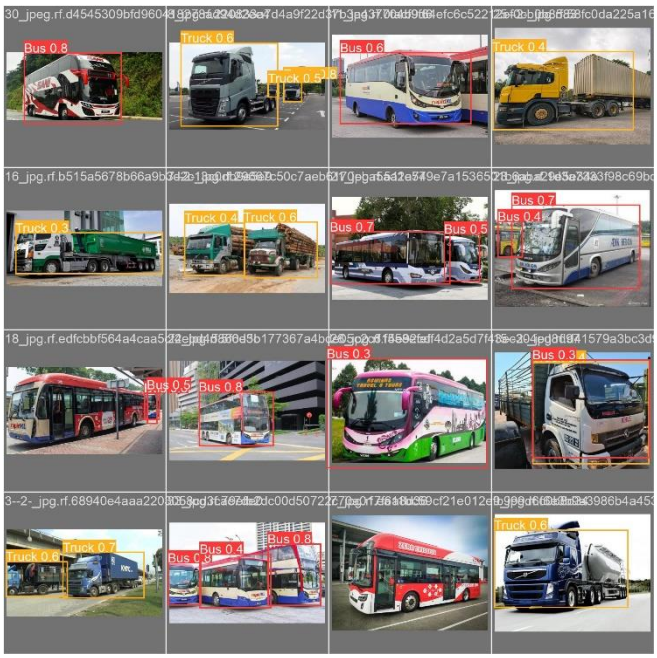

Fig. 8. Prediction Object Detection Label

Based on Figure 7 and Figure 8, it can be analyzed using human interpretation that the model is good enough to detect objects and perform classification processes based on bounding boxes and confidence levels to classify large vehicles such as buses. However, if viewed in more detail, there are several images where the bounding box results between the truck and bus classes appear simultaneously. This can be a shortcoming of the model caused by the type of dataset or image that is still lacking in terms of data variation.

IV. Conclusion

Through the experimental process by tuning the YOLOv5 model, the results were obtained that the YOLOv5 model with a combination of parameters image size 650,

batch size 10 and epoch 50 is the parameter combination with the best performance. This is proven by the mAP performance results with a value of 0.97, F1-Score 0.92, precision value in the bus class 0.84 and truck class 0.99, recall or sensitivity value in the bus class 0.84 and truck class 1. The results of this performance evaluation can be a justification for the YOLOv5 model The tuning results in this research can be used to overcome the problem of road damage by limiting the number of trucks operating on the highway. However, there are still shortcomings in this research, namely the lack of variation in the image dataset used in this research. It is highly recommended to generate models with more varied training data.

ACKNOWLEDGMENT

The dataset used in this research is a dataset collected and provided on the roboflow website. The dataset was published by roboflow in February 2023 [7].

The program code base used in this research is a code base provided by the Ultralytics website and developed by Glenn Jocher in 2020 [15].

REFERENCES

[1] S. H and R. J, "Vehicle Detection and Classification using YOLOv5 on Fused Infrared and Visible Images," *2023 International Conference on Inventive Computation Technologies (ICICT)*, Lalitpur, Nepal, 2023, pp. 1024-1030, doi: 10.1109/ICICT57646.2023.10134214.

[2] D. Snegireva and G. Kataev, "Vehicle Classification Application on Video Using Yolov5 Architecture," *2021 International Russian Automation Conference (RusAutoCon)*, Sochi, Russian Federation, 2021, pp. 1008-1013, doi: 10.1109/RusAutoCon52004.2021.9537439.

[3] Z. Li, X. Tian, Y. Liu and X. Shi, "Vehicle Tracking Method Based on Attention-YOLOv5 and Optimized DeepSort Models," 2022 IEEE 11th Data Driven Control and Learning Systems Conference (DDCLS), Chengdu, China, 2022, pp. 114-121, doi: 10.1109/DDCLS55054.2022.9858395.

[4] T. -H. Wu, T. -W. Wang and Y. -Q. Liu, "Real-Time Vehicle and Distance Detection Based on Improved Yolo v5 Network," 2021 3rd World Symposium on Artificial Intelligence (WSAI), Guangzhou, China, 2021, pp. 24-28, doi: 10.1109/WSAI51899.2021.9486316.

[5] Z. Lu, L. Ding, Z. Wang, L. Dong and Z. Guo, "Road Condition Detection Based on Deep Learning YOLOv5 Network," 2023 IEEE 3rd International Conference on Electronic Technology, Communication and Information (ICETCI), Changchun, China, 2023, pp. 497-501, doi: 10.1109/ICETCI57876.2023.10176545.

[6] F. Ahmad, M. Z. Ansari, S. Hamid and M. Saad, "A Computer Vision based Vehicle Counting and Speed Detection System," 2023 International Conference on Recent Advances in Electrical, Electronics & Digital Healthcare Technologies (REEDCON), New Delhi, India, 2023, pp. 487-492, doi: 10.1109/REEDCON57544.2023.10151423.

[7] yolov5, 'Bus and truck Dataset', *Roboflow Universe*. Roboflow, Feb-2023.

[8] S. Tan, G. Lu, Z. Jiang and L. Huang, "Improved YOLOv5 Network Model and Application in Safety Helmet Detection," 2021 IEEE International Conference on Intelligence and Safety for Robotics (ISR), Tokoname, Japan, 2021, pp. 330-333, doi: 10.1109/ISR50024.2021.9419561.

[9] Li, Zhuang & Tian, Xincheng & Liu, Xin & Liu, Yan & Shi, Xiaorui. (2022). A Two-Stage Industrial Defect Detection Framework Based on Improved-YOLOv5 and Optimized-Inception-ResnetV2 Models. Applied Sciences. 12. 834. 10.3390/app12020834.

[10] Wu, C.; Wen, W.; Afzal, T.; Zhang, Y.; Chen, Y. A compact DNN: Approaching GoogLeNet-Level accuracy of classification and domain adaptation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.

[11] Kim, D.; Park, S.; Kang, D.; Paik, J. Improved center and scale prediction-based pedestrian detection using convolutionalblock. In Proceedings of the 2019 IEEE 9th International Conference on Consumer Electronics (ICCE-Berlin), Berlin, Germany, 8–11 September 2019; pp. 418–419.

[12] He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 2015, 37, 1904–1916

[13] Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Amsterdam, The Netherlands; pp. 21–37.

[14] Wang, W.; Xie, E.; Song, X.; Zang, Y.; Wang, W.; Lu, T.; Yu, G.; Shen, C. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; IEEE: Seoul, Korea; pp. 8440–8449.

[15] G. Jocher, Ultralytics YOLOv5. 2020.