AN EFFICIENT TEXT DETECTION IN ADVERTISEMENT IMAGES BASED ON YOLOV5 ALGORITHM

Aanuoluwa Oyebola ADIO[1], Caleb Olufisioye AKANBI[2]
Adepeju Abeke ADIGUN[3]
[1]Redeemer's University, Ede, Osun State, Nigeria
[2]Osun State University, Osogbo, Osun State, Nigeria
[3]Osun State University, Osogbo, Osun State, Nigeria
[1]adioa@run.edu.ng, [2]akanbico@uniosun.edu.ng, [3]adepeju.adigun@uniosun.edu.ng

Keywords: Text Detection, Advertisement Images, YOLO, Computer Vision, Deep learning

Abstract: *The use of digital media for advertisement is on the increase due to advancements in technology and mobile devices and in their applications. Prominent of the digital media is images. Digital Advertisement images are images of products sold and/or services rendered by the advertiser. In addition, advertisement images also contain textual information that helps connect the potential customer to the advertiser, such as telephone numbers. Potential customers tend to copy or cram the contact information which could lead to the loss of this information. The incorrect information. will hinder the aim of the advertisement from being fulfilled. The Automatic detection of these texts using computer vision algorithms can foster the establishment of communication between the advertiser and the potential customer. In this study, the You Only Look Once (YOLO) model is utilized for text detection from advertisement images. YOLO is a widely used object detection algorithm in particular YOLOv5. It is known for its balance between speed and accuracy of detections and has been used in various sectors. However, its performance for text detection from advertisement images is yet to be explored. Therefore, this paper investigates and compares the performance of various variants of YOLOv5 for the accurate detection of texts from digital advertisement images. Results of experiments showed that Yolov5x achieved a precision, recall, and mAP50 of 98.9, 97.7, and 99.4.*

1. INTRODUCTION

An advertisement image is a visual image created to promote a product, service, brand, or idea with the aim of capturing the attention and persuading potential customers to take a specific action, such as purchasing, subscribing, or engaging with the promoted content [1]. The advertisement images are basically used in advertisement and marketing campaigns across various digital platforms, such as print media, websites, social media, and so on [2]. Advertisement images contain attention-grabbing as well [3] and clear messages prompting the target viewers to take action towards patronizing the brand [4]. These advertisement images also contain a call to action, which could be in the form similar to "Buy Now," "Call," "Subscribe," "E-mail," "Visit our website for more information," etc. The texts on advertisement images vary in size, fonts, and shapes and can be located on any part of the image [5]. The automatic detection of these texts on the advertisement images can be used for various purposes such as product label identification [6], intelligent grocery list identification [7], [8], food package label identification [9] in various shopping malls, sales outlets, television adverts and so on.

The research on automatic detection of objects and, in particular, texts [10] from images is vast and has shown tremendous results in the area of Computer Vision [11]. Various techniques have been used to detect texts from images automatically [12] – [15]. In the case of advertisement images, extracting real-time and precise contact information concerning how the potential customer will access the goods and services rendered by the advertiser is key to the success of advertising [16]. One of the challenges of automatic detection of texts from advertisement images is that the texts on advertisement images are clustered among other pictures of products

advertised in the image [5], [17]. Several works have been done in the area of text detection in clustered images, however, limited works have been done on text detection from advertisement images. Several algorithms have been presented in literature to address the problem of text detection, e.g. Single shot detector (SSD) [18], Retinanet algorithm [19], Faster RCNN [20], and YOLO algorithm [21]. The YOLO algorithm has gained a lot of attraction for its efficient and accurate detection of objects in images [22]. The YOLO algorithms has evolved over the years resulting in different versions. This study is particularly interested in investigating the performance of YOLOv5 for accurate text detection from advertisement images.

2. RELATED WORKS

Quite a number of techniques have been designed for the text detection in natural images; however, specialized techniques are needed for advertisement images due to their particular issues, such as different fonts types, color variants, complicated backdrops, and random placements. Recent developments in deep learning, especially in frameworks for object detection, have demonstrated encouraging outcomes in tackling these difficulties. These include Faster – RCNN [20], EfficientDet [23], Retinanet [19], YOLO[21], Single shot multi-box detection [18], etc. YOLO (You Only Look Once) is one of the most popular techniques for real-time object detection; in recent years, its potential for text detection has been investigated.

Since its introduction by [24], YOLO has been altered for several use cases, such as text detection, noting that its original purpose and target was object detection. It was desirable for high-speed applications because of its real-time detection capacity, which treats the detection of objects as a single regression problem.

The performance of the YOLOv4 algorithm for real-time text detection and recognition was investigated by [25]. The study proposed the YOLO algorithm as a potential alternative for a fast and accurate real-time text direction and recognition system in videos and images. The video frames and images were first preprocessed and enhanced. The text detection was performed next using the improved YOLO model to handle various issues of photometric distortion and geometric distortion. This process is immediately followed by the text recognition process using the Tesseract algorithm. The result of the experiment shows that it is suitable for the real-time detection of texts from images and videos. The study on natural scene text detection was carried out by [26]. In this study, YOLOv2 was used to detect texts in natural scene images. Results from the experiment showed that the model is accurate as well as the detection slip is encouraging.

YOLOv5, a member of the YOLO family, has demonstrated advancements in accuracy as well as speed compared to its predecessors by utilizing multi-scale feature maps, sophisticated loss functions, and anchor boxes. The capacity of YOLOv5 to recognize objects at numerous scales and in varying orientations has been investigated by researchers as a potential text identification tool. This capability is particularly useful for advertisement photos, where text frequently appears in unusual layouts. In [27], the adaptation of YOLOv5x for scene text detection and recognition, for example, showed its resilience in both structured and unstructured contexts after fine-tuning it on datasets like COCO-Text and ICDAR.

Because of the various qualities of texts, which include a variety of styles, shapes, text effects, and crowded backgrounds, advertising graphics present particular obstacles. The challenge of identifying text in advertising is not well-trodden, although there are many methods that can be applied to this domain that were created for text detection in natural scenes.

In recent research, [28] designed a study that provides an alternative scanning system with a camera instead of Radio-Frequency Identification (RFID) and barcode scanners was used to develop a YOLOv5-Based Smart Item Recognition System for Grocery Shopping. The results showed that the sensitivity was 46.67%, the average was 90.80%, the recall rate was 92.15%, and the accuracy was 84.66%. Yolov5 performs similarly to other algorithms in intelligent item recognition. [29] also developed an application based on computer vision models to detect, count, and verify the status of bottled and canned products. YOLOv5 object detection model was chosen for the experiment. Various parameters were experimented with, the performance of the models was also evaluated, and the metrics were checked to determine the model that performed best. A dataset of product images in both good and bad conditions were used for training. Afterwards, the model that has been trained was put together to develop the application [29].

Yolov5 Algorithm was used to identify different road marking signs [30]. This study attempts to explore various contemporary object detectors, such as road marking signs and location detection [30]. The authors of the study provided a brief explanation of Yolov5 algorithm variants and how it was used for the identification of road sign markings, which ranges from Yolov5n to Yolo5x. Findings show that Yolov5m and Yolov5l achieved the best scores on mAP with 87%. Yolov5m proved to be the model that performed well with so much stability than when compared to the remaining models, with 76%, 86%, and 83% for precision, recall, and mAP, respectively, during the training stage. A grocery product detection and recognition framework based on deep learning technique was designed by [31]. The authors focused on the detection and recognition of many products placed at different locations on the shelf and off the shelf in grocery stores by identifying the label texts at once. The study designed the new framework, which comprises three parts: the detection of the retail product, followed by the detection of the text on the product, and eventually, the recognition of the text on the product [31].

The contributions of this study are to collect a digital advertisement images dataset, annotate the text locations, implement the various variants of YOLOv5 on the collected dataset, and also analyze the performance of the YOLOv5 model using the recall, precision, mAP50, and mAP50:95 metrics. Previous research has been done on images of billboards and signposts; this study is based on advertisement images in the social media space and the exploration of YOLOv5 for the automatic detection of texts.

3. MATERIALS AND METHOD

This paper introduces the framework for text detection from advertisement images, as shown in Fig 1.

The framework starts with the collection of digital advertisement image datasets, the annotation of the location of texts, image preprocessing, augmentation, model training, and then evaluation of the performance of the model.

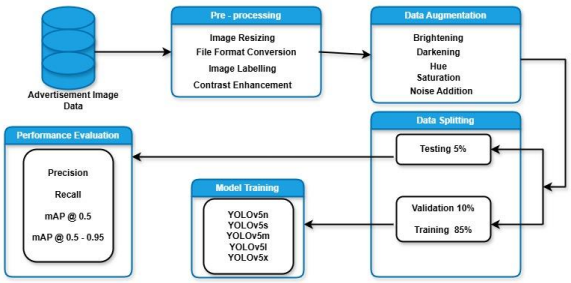

Fig 1: Text Detection from advertisement Images Framework

Digital advertisement images were collected from different social media such as Instagram and WhatsApp Status. A Total of 400 images were collected. The text locations in the images were annotated using the LabelImg annotation tool. A total of 19456 texts instances were annotated for use in this study. The images obtained were of different image sizes and file formats. The images were all resized to 512 by 512 pixels and in jpg file format. The contrast of the images was enhanced using the CLAHE algorithm. Due to the small amount of images collected for this study, the advertisement images were augmented using techniques such as Brightening, darkening, Hue, saturation, and addition of noise. After augmentation, the quantity of images in the dataset increased to 5207 images.

The augmented dataset was split into training, validation, and testing in the ratio 80:10:10, respectively. The YOLOv5 algorithm was implemented in this study for the text detection task. Fig 2 shows the model architecture. The architecture is divided into the backbone, the neck, and the head. The backbone handles the responsibility of extracting relevant features from different sizes necessary for accurate text detection from advertisement images. The neck handles the responsibility for the fusion of the feature maps from the backbone to generate feature maps P3, P4, and P5 which are used to detect texts of various sizes in the image. The head is responsible for the final prediction of text locations using the class confidence scores calculated and bounding box regression calculated from each feature map. The final output is the class confidence scores and the predicted bounding box coordinates.

Five variants of Yolov5 were explored in this study. They are: YOLOv5nano (n), YOLOv5small (s), YOLOv5medium (m), YOLOv5large (l) and YOLOv5xtra-large (x). The YOLOv5 variants differ in terms of the depth multiplier of the channels and the

width multiplier of the filter used during the model's training. The summary of the depths, and width are as shown in Table 1.

The experiment was carried out in the Google Colab environment, where we leveraged its free GPU resources (NVIDIA Tesla T4) for fast training of the models. The YOLOv5 is built around the PyTorch framework supported by Google Colab. The training (train.py) script from the official YOLOv5 GitHub source was used for best results. For the training of the models, pre-trained weights were loaded using the –weight argument alongside the input image size parameter (--img), which ensured that the models were trained using the 512 x 512 pixels. All the models were trained for 30 epochs, each with a batch size 16 and Adam optimizer. During training, after each epoch, validation is conducted to monitor the ability of the model to generalize.

Table 1: YOLOv5 variants Parameters [32], [33]

Name of Model	*Depth Multiplier*	*Width Multiplier*
YOLOv5n	*0.33*	*0.25*
YOLOv5s	*0.33*	*0.50*
YOLOv5m	*0.67*	*0.75*
YOLOv5l	*1.0*	*1.0*
YOLOv5x	*1.33*	*1.25*

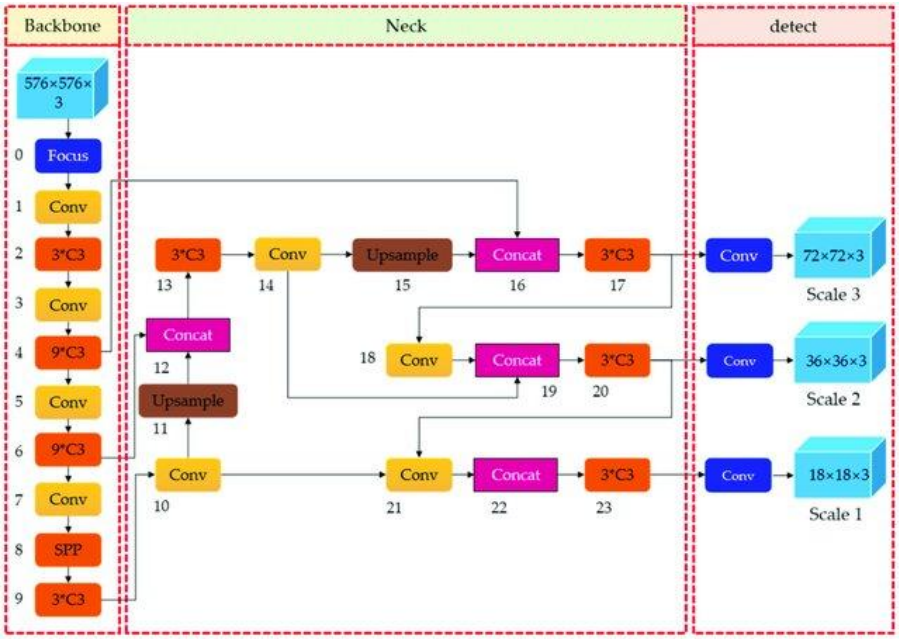

Fig. 2 YOLOv5 Architecture[32]

4. RESULTS AND DISCUSSION

The performance of the trained models for text detection from advertisement images is evaluated based on the Precision, Recall, mAP50, and mAP50-95 metrics. The metrics are based on the true positive rate (TP), False Positive rate (FP), True Negative (TN), and False negative rates (FN). The TP is the instance where the model correctly localizes the texts. The FN are the instances where the model localizes the text locations incorrectly. The FP are the instances where the model localizes another object as a text. The TN are the instances where other objects are correctly localized as other objects and not text. The performance metrics used in this study are derived from the TP, TN, FP, and FN. The Precision metric evaluates how well the percentage of predicted positives are positives, as given in Equ 1. The Recall metric evaluates the fraction of the true negatives predicted corresponding to the actual true negatives in the test set, given in Equ 2. The mAP@0.5 and mAP@0.5 – 0.95 evaluate the model's performance at different Intercept of Unions (IoU) between the predicted text locations and the ground truths.

$$Precision = \frac{TP}{TP + FP}\ x\ 100 \qquad Equ.1$$

$$Recall = \frac{TP}{TP + FN}\ x\ 100 \qquad Equ.2$$

Table 2: Result of Yolov5 variants on Digital Advertisement Images

Model	*P (%)*	*R (%)*	*mAP@ .50 (%)*	*mAP@.50-.95 (%)*
YOLOv5n	*91.2*	*89.1*	*92.8*	*55.3*
YOLOv5s	*93.3*	*92.6*	*95.5*	*60.1*
YOLOv5m	*95.7*	*94.2*	*97.2*	*63.9*
YOLOv5l	*96.0*	*95.6*	*97.8*	*65.8*
YOLOv5x	*98.9*	*97.7*	*99.4*	*81.4*

The results for the precision and recall for the YOLOv5n model are 91.2% and 89.1%, respectively. The mAP@0.5 score is 92.8%, while mAP@0.5 – 0.95 is 55.3%. This shows that despite its high precision and recall, when a more stringent IoU threshold was used to evaluate the performance of YOLOv5n, the performance was very poor. The YOLOv5s performed better than YOLOv5n with 93.3%, 92.6%, 95.5%, and 60.1% for precision, recall, mAP@0.5, and mAP@0.5 – 0.95 respectively. YOLOv5m further performed better than the previous models with a precision and recall of 95.7%, and 94.2%. The mAP@0.5 and mAP@0.5 – 0.95 recorded 97.2% and 63.9%, respectively. The performance of the three models above is due to the fewer parameters and small architecture used for training. Compared to YOLOv5n, YOLOv5s, and YOLOv5m, the performance of Yolov5l showed that its larger configuration improved the model's robustness. With a performance of 96.0% precision, 95.6% recall, 97.8% mAP@0.5, and 65.8% for mAP@0.5 – 0.95, results show that with the increasing depth and width, the performance tends to get better. The YOLOv5x has the largest depth and width multiplier parameters, and its performance for detecting texts in advertisement images shows its remarkable accuracy in its localization. It achieved a precision of 98.9%, recall of 97.7%, mAP@0.5 of 99.4%, and mAP@0.5 – 0.95 of 81.4%. The summary of the results of the YOLOv5 variants is shown in Table2. The results from the experiments show that the deeper the architecture, the better the performance of the model. For complex situations that contain small and clustered texts, the experiment shows that deeper architecture is better for accurate text detection in advertisement images.

5. CONCLUSIONS

The paper focused on the investigation of the performance of the five variants of YOLOv5 for the detection of texts from digital advertisement images. Our study discovered that models with deeper architecture tend to achieve better detection accuracy. However, there is a need to develop a more accurate model for text detection in advertisement images as a mAP@0.5-0.95 is low since the accuracy of the detected texts is of high importance. Accurate text detection is essential because accurate text detection will lead to accurate text recognition in cropped text instances from advertisement images. Further research could investigate improving the YOLOv5 algorithm to further achieve better accuracy of text detection in advertisement images.

6. REFERENCES

[1] Z. Said, "The Effect Of Advertisement On Brand Image (The Case Of Bank Of Abyssinia)," St. Mary's University, 2023.

[2] Oluwaseun Abiola Ajiva, Onyinye Gift Ejike, and Angela Omozele Abhulimen, "The critical role of professional photography in digital marketing for SMEs: Strategies and best practices for success," *Int. J. Manag. Entrep. Res.*, vol. 6, no. 8, pp. 2626–2636, 2024, doi: 10.51594/ijmer.v6i8.1410.

[3] M. Torbarina, "Human Face And Cognitive Load Effects On Advertisement Attention Grabbing And Attention Guiding," 2022.

[4] O. Olubusola Temiloluwa, T. Usman Moyosore, and B. Mustapha Tosin, "Investigating the effect of sales promotion on customer patronage of household appliances within Lagos metropolis," *IROCAMM-International Rev. Commun. Mark. Mix*, vol. 5, no. 2, pp. 119–129, 2022, doi: 10.12795/irocamm.2022.v05.i02.07.

[5] R. Pieters and M. Wedel, "Attention Capture and Transfer in Advertising: Brand, Pictorial, and Text-Size Effects," *J. Mark.*, vol. 68, no. 2, pp. 36–50, 2004, doi: 10.1509/jmkg.68.2.36.27794.

[6] S. Yao and K. Zhu, "Combating product label misconduct: The role of traceability and market inspection," *Eur. J. Oper. Res.*, vol. 282, no. 2, pp. 559–568, 2020, doi: 10.1016/j.ejor.2019.09.031.

[7] E. Gothai, S. Bhatia, A. M. Alabdali, D. K. Sharma, B. R. Kondamudi, and P. Dadheech, "Design Features of Grocery Product Recognition Using Deep Learning," *Intell. Autom. Soft Comput.*, vol. 34, no. 2, pp. 1231–1246, 2022, doi: 10.32604/iasc.2022.026264.

[8] N. Chabane, M. A. Bouaoune, R. A. S. Tighilt, B. Mazoure, N. Tahiri, and V. Makarenkov, "Using Clustering and Machine Learning Methods to Provide Intelligent Grocery Shopping Recommendations," in *Studies in Classification, Data Analysis, and Knowledge Organization*, 2023, pp. 83–91, doi: 10.1007/978-3-031-

09034-9_10.
[9] A. Ur Rehman, I. Gallo, and P. Lorenzo, "A Food Package Recognition Framework for Enhancing Efficiency Leveraging the Object Detection Model," in *ICAC 2023 - 28th International Conference on Automation and Computing*, 2023, pp. 1–6, doi: 10.1109/ICAC57885.2023.10275193.
[10] H. Li, D. Doermann, and O. Kia, "Automatic text detection and tracking in digital video," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 147–156, 2000, doi: 10.1109/83.817607.
[11] X. C. Yin, X. Yin, K. Huang, and H. W. Hao, "Robust text detection in natural scene images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 970–983, 2014, doi: 10.1109/TPAMI.2013.182.
[12] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading Text in the Wild with Convolutional Neural Networks," *Int. J. Comput. Vis.*, vol. 116, no. 1, pp. 1–20, 2016, doi: 10.1007/s11263-015-0823-z.
[13] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object Detection in 20 Years: A Survey," *arXiv Prepr. arXiv1905.05055*, 2019, [Online]. Available: http://arxiv.org/abs/1905.05055.
[14] Q. Ye and D. Doermann, "Text Detection and Recognition in Imagery: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1480–1500, 2015, doi: 10.1109/TPAMI.2014.2366765.
[15] W. Wu, J. Xing, C. Yang, Y. Wang, and H. Zhou, "Texts as Lines: Text Detection with Weak Supervision," *Math. Probl. Eng.*, vol. 2020, 2020, doi: 10.1155/2020/3871897.
[16] J. A. Choi and K. Lim, "Identifying machine learning techniques for classification of target advertising," *ICT Express*, vol. 6, no. 3, pp. 175–180, 2020, doi: 10.1016/j.icte.2020.04.012.
[17] L. M. Scott, "Images in Advertising: The Need for a Theory of Visual Rhetoric," *J. Consum. Res.*, vol. 21, no. 2, p. 252, 1994, doi: 10.1086/209396.
[18] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9905 LNCS, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.
[19] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, vol. 42, no. 2, pp. 318–327, doi: 10.1109/TPAMI.2018.2858826.
[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017, doi: 10.1109/TPAMI.2016.2577031.
[21] R. Khanam and M. Hussain, "What is YOLOv5: A deep look into the internal features of the popular object detector," *arXiv Prepr. arXiv2407.20892*, 2024, [Online]. Available: http://arxiv.org/abs/2407.20892.
[22] A. Jawaharlalnehru *et al.*, "Target Object Detection from Unmanned Aerial Vehicle (UAV) Images Based on Improved YOLO Algorithm," *Electron.*, vol. 11, no. 15, p. 2343, 2022, doi: 10.3390/electronics11152343.
[23] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10778–10787, doi: 10.1109/CVPR42600.2020.01079.
[24] E. Hassan, Y. Khalil, and I. Ahmad, "Learning Feature Fusion in Deep Learning-Based Object Detector," *Sci. World J.*, vol. 2020, 2020, doi: 10.1155/2020/7286187.
[25] S. Pei and M. Zhu, "Real-Time Text Detection and Recognition," *arXiv Prepr. arXiv2011.00380*, 2020, [Online]. Available: http://arxiv.org/abs/2011.00380.
[26] D. Haifeng and H. Siqi, "Natural scene text detection based on YOLO V2 network model," in *Journal of Physics: Conference Series*, 2020, vol. 1634, no. 1, p. 12013, doi: 10.1088/1742-6596/1634/1/012013.
[27] Y. L. Chaitra, R. Dinesh, M. Jeevan, M. Arpitha, V. Aishwarya, and K. Akshitha, "An Impact of YOLOv5 on Text Detection and Recognition System using TesseractOCR in Images/Video Frames," in *IEEE International Conference on Data Science and Information System, ICDSIS 2022*, 2022, pp. 1–6, doi: 10.1109/ICDSIS55133.2022.9915927.
[28] L. J. P. Manalo, E. A. S. Ornillo, and J. B. G. Ibarra, "YOLOv5-Based Smart Item Recognition System for Grocery Shopping," in *2024 IEEE 4th International Conference on Electronic Communications, Internet of Things and Big Data, ICEIB 2024*, 2024, pp. 660–665, doi: 10.1109/ICEIB61477.2024.10602671.
[29] P. Herrera-Toranzo, J. Castro-Rivera, and W. Ugarte, "Detection and Verification of the Status of Products Using YOLOv5.," in *ICSBT*, 2023, pp. 83–93.
[30] C. Dewi, R. C. Chen, Y. C. Zhuang, and H. J. Christanto, "Yolov5 Series Algorithm for Road Marking Sign Identification," *Big Data Cogn. Comput.*, vol. 6, no. 4, p. 149, 2022, doi: 10.3390/bdcc6040149.
[31] P. Selvam and J. A. S. Koilraj, "A Deep Learning Framework for Grocery Product Detection and Recognition," *Food Anal. Methods*, vol. 15, no. 12, pp. 3498–3522, 2022, doi: 10.1007/s12161-022-02384-2.
[32] Z. Li, X. Tian, X. Liu, Y. Liu, and X. Shi, "A Two-Stage Industrial Defect Detection Framework Based on Improved-YOLOv5 and Optimized-Inception-ResnetV2 Models," *Appl. Sci.*, vol. 12, no. 2, p. 834, 2022, doi: 10.3390/app12020834.
[33] J. Wang and J. Li, "Bogie-YOLO: A key component detection model for high-speed train bogies," in *Journal of Physics: Conference Series*, 2024, vol. 2816, no. 1, p. 12067, doi: 10.1088/1742-6596/2816/1/012067.