Gesture Object Detection and Recognition Based on YOLOv11

Jian Xu[1,a,*], Heyao Chen[2], Xingpeng Xiao[3], Mengyuan Zhao[4], Bo Liu[5]

[1]Electrical and Electronics Engineering, University of Southern California, Los Angeles, USA
[2]Computer Science and Technology, Beijing University of Posts and Telecommunications, Beijing, China
[3]Computer Application Technology, Shandong University of Science and Technology, Qingdao, China
[4]Information System, Northeastern University, Oakland, USA
[5]Computer Software Engineering, Northeastern University, Boston, USA
a. xuj2979@gmail.com
**corresponding author*

Abstract: This article explores the application of YOLOv11 algorithm in gesture recognition field to evaluate its performance in human-computer interaction (HCI). By introducing the YOLOv8 model and corresponding dataset for training, we obtained the confusion matrix predicted by the model, which shows that the model can accurately recognize most gestures, although there are a few cases of misidentification. When the IoU threshold is 0.5, the average accuracy (mAP) of the model steadily improves with the progress of training, indicating that the overall performance of the model in gesture detection tasks has been enhanced. In addition, even under stricter evaluation conditions where the IoU threshold was increased from 0.5 to 0.95, mAP still showed an upward trend, although the growth rate was not as significant as mAP50, which still demonstrated the improvement in model performance. Through the detection of the test set images, we found that the YOLOv11 model can effectively recognize gestures and accurately interpret their meanings, demonstrating high accuracy. This study not only demonstrates the potential of YOLOv11 in gesture recognition tasks, but also provides a new technological path for the future development of HCI field. Overall, the YOLOv11 algorithm has demonstrated strong performance and accuracy in gesture recognition, providing a more natural and intuitive way for interaction between smart devices and humans.

Keywords: YOLOv11, Gesture recognition, Human-computer interaction

1. Introduction

Gesture recognition, as an important research direction in the field of human-computer interaction (HCI), is a technology that enables communication with computers or other intelligent devices by analyzing human gestures and actions. With the rapid development of technology, especially the rapid advancement of computer vision and machine learning technology, the application scope of gesture recognition is constantly expanding [1]. From the initial game control to the current fields of smart homes, virtual reality (VR), augmented reality (AR), etc., gesture recognition technology has played an indispensable role in improving user experience and interaction efficiency. Therefore, researchers

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

have conducted extensive exploration and innovation in gesture recognition algorithms and technologies [2].

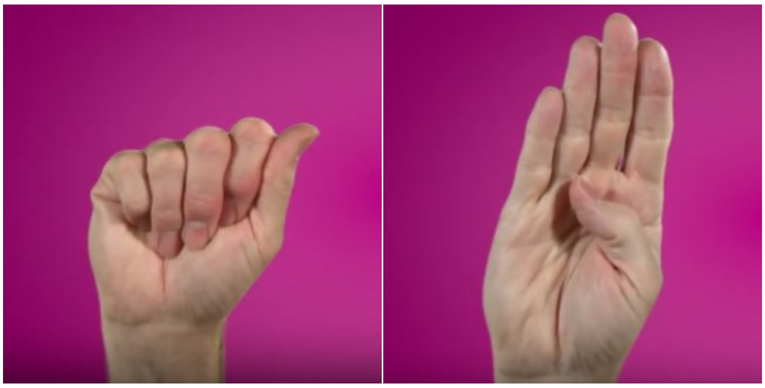
The research background of gesture recognition covers multiple aspects. Firstly, with the popularity of mobile devices, smart homes, and wearable devices, users' demand for natural and intuitive interaction methods has become increasingly evident. Compared to traditional input methods such as keyboards and mice, gesture recognition can provide a more natural communication experience, especially in scenarios that require fast and instant interaction, showing significant advantages. In addition, in the research field, gesture recognition can serve as an interactive tool in virtual reality environments and can also be widely applied in assistive technologies for people with physical disabilities, helping them better interact with their surroundings. Therefore, the research on gesture recognition not only has important academic value, but also has broad practical significance.

The importance of object detection algorithms is indispensable in gesture recognition technology. YOLO (You Only Look Once) is a real-time object detection algorithm that has received widespread attention due to its superior accuracy and speed. YOLO achieves efficient and fast real-time detection by treating the object detection process as a single regression problem. The application of YOLO algorithm in gesture recognition greatly enhances the feasibility and practicality of the model. Firstly, YOLO is able to accurately recognize gestures in complex backgrounds, especially when classifying gestures among numerous targets, its classification ability is particularly outstanding. In addition, YOLO's high real-time performance enables it to process real-time video streams, allowing gesture recognition to proceed smoothly in dynamic and changing scenes.

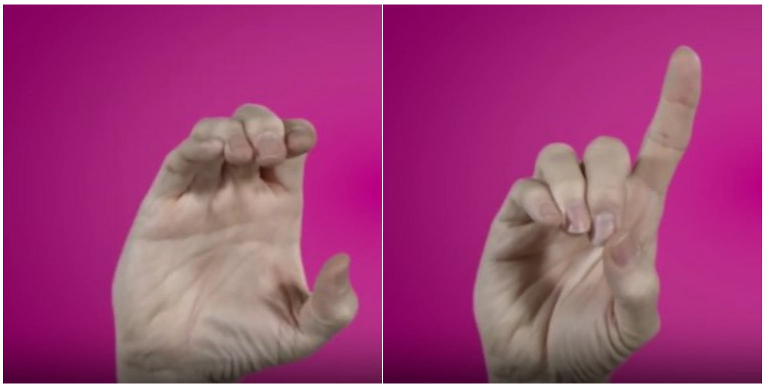
By using the YOLO algorithm, the gesture recognition system can quickly detect the position and state of the hand, and thus recognize different gestures. This real-time feedback not only improves the user experience, but also provides possibilities for various interactive applications [3]. For example, in virtual reality and augmented reality applications, users can interact with virtual objects through simple gestures, enhancing immersion and interactivity. In addition, the multitask learning ability of YOLO algorithm also provides a solution for the complexity of gesture recognition, enabling the system to detect other objects while recognizing gestures, achieving more diverse application scenarios. This article applies the most advanced YOLOv11 algorithm in the YOLO series to gesture recognition tasks and observes the detection performance of the algorithm [4].

2. Data sources

The dataset used in this article is the Ziyou dataset, which contains different gestures and their corresponding meanings. There are a total of 26 gestures from A to Z, representing 26 English letters. As shown in Figure 1. Figure 1. (a) represents A, Figure 1. (b) represents B, Figure 1. (c) represents C, and Figure 1. (d) represents D.



(a) (b)



(c) (d)

Figure 1: Dataset Introduction.

3. Method

3.1. YOLO model

YOLO (You Only Look Once) is an important innovative technology in the field of object detection, which has undergone multiple iterations and continuous optimization since its first proposal in 2016. This model quickly became a popular choice in the field of object detection due to its concise and efficient architecture. The development process of YOLO series models can be divided into several important stages, and the release of each version marks a significant technological advancement. The structure of the YOLO model is shown in Figure 2.

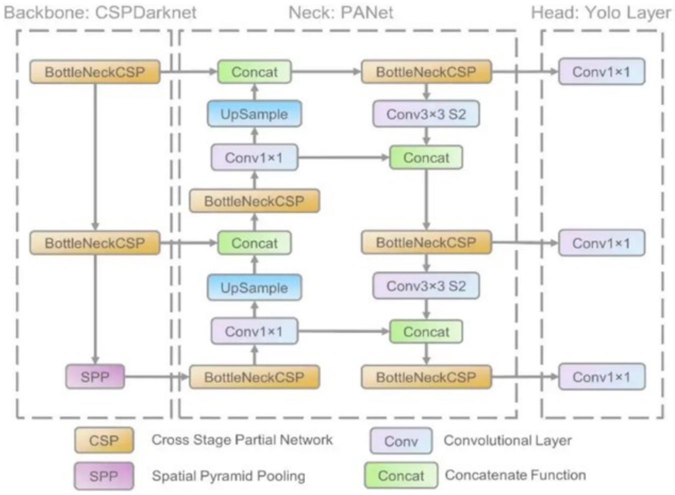


Figure 2: The structure of the YOLO model.

The original YOLO model was released in 2016, with the core idea of treating object detection as a regression problem rather than a traditional combination of classification and localization. YOLO achieves an end-to-end detection method by dividing the input image into grids and simultaneously predicting bounding boxes and assigning class labels in each grid. This method significantly improves detection speed, making real-time object recognition possible and promoting the widespread implementation of related applications. Although YOLO has advantages in speed, there is still room for improvement in small object detection and accuracy. [5]

Subsequently, YOLOv2 was released in 2017, which further improved the performance of the model by introducing the concepts of multi-scale training and anchor boxes. YOLOv2 also uses high-resolution input images to improve its ability to detect small targets, and supports multiple types of target detection commonly used in object detection, adapting to more complex scenes. Through these improvements, this version has achieved a better balance between accuracy and speed, laying the foundation for subsequent YOLO versions [6].

In 2018, YOLOv3 was launched, further enhancing the model's feature extraction capability. YOLOv3 introduces cross layer feature fusion, enabling the model to transfer features between different scales, resulting in a significant improvement in the model's ability to detect small targets. In addition, YOLOv3 adopts a deep convolutional neural network (Darknet-53) as the feature extractor, which improves the overall detection performance.

With the development of the YOLO series, the YOLOv4 released in 2020 has once again taken object detection performance to a new level. YOLOv4 improves feature extraction and information flow by adopting many advanced technologies such as amplitude enhancement, spatial attention mechanism, CSPNet, etc. In addition, YOLOv4 introduces more efficient data augmentation techniques during the training process, which enhances the robustness of the model in various environments. This version not only maintains fast detection speed, but also significantly improves accuracy, making it widely used in various fields, including security, autonomous driving, etc.

In 2021, the YOLO family welcomed a new member - YOLOv5. In YOLOv5, developers not only improved the model architecture, but also introduced a more flexible implementation framework that supports multiple resolution and size options, greatly enhancing the usability and deployability of the model. YOLOv5 also utilizes the latest patterns and optimization methods to further improve accuracy and speed, while bringing better performance through adaptive anchor box and automatic mixed precision training techniques. In various competitions, YOLOv5 has demonstrated excellent performance and gained widespread popularity in the open source community [7].

Subsequently, YOLOv6 was released and further optimized in terms of performance and application temperature, especially for applications on edge devices. YOLOv6 provides a more efficient solution for mobile devices and embedded systems, enabling better performance in real-time object detection on-site.

The release of the latest version YOLOv11 once again marks a significant technological breakthrough for the YOLO family. YOLOv11 further optimizes the balance between speed, accuracy, and model size by combining more advanced deep learning scheduling strategies and adaptive training methods, enabling the model to have higher stability and accuracy when processing large-scale data. In addition, YOLOv11 pays special attention to the processing of small targets and complex backgrounds, adopting various innovative feature enhancement techniques to ensure reliability in various environments [8].

3.2. YOLOv11

YOLOv11 is the latest version in the YOLO (You Only Look Once) series, aimed at pushing the boundaries of object detection technology, improving detection accuracy and efficiency. This version has undergone deep level optimization based on previous generations, combining the latest technologies and theories of modern deep learning to form a more efficient and flexible object detection model. The principle of YOLOv11 can be analyzed in detail from the following aspects. The structure of the YOLOv11 model is shown in Figure 3.

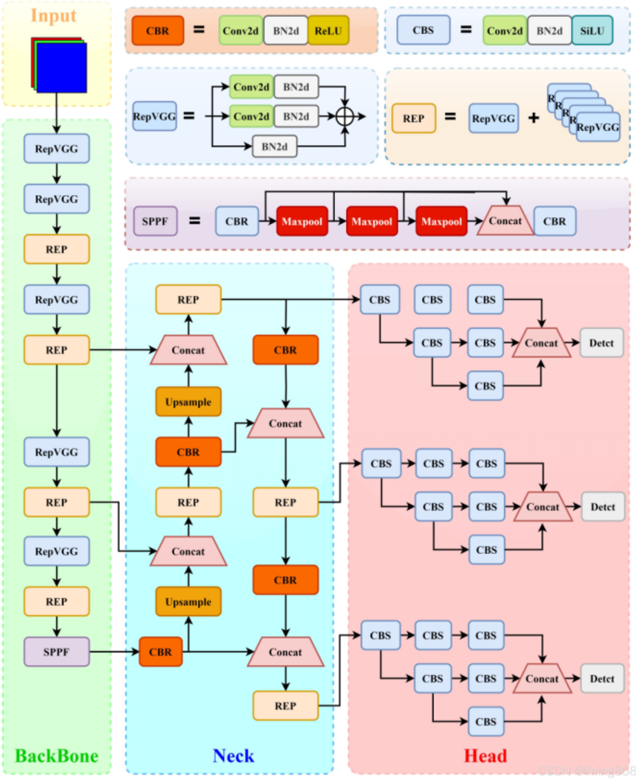


Figure 3: The structure of the YOLOv11 model.

Firstly, YOLOv11 continues the core concept of the YOLO series models, which treats object detection tasks as a regression problem and directly predicts bounding boxes and assigns class labels on the image. Unlike previous versions of YOLO, YOLOv11 has made significant improvements in network architecture. It introduces a new feature extractor that combines convolutional layers with attention mechanisms, enhancing its ability to model contextual information and spatial relationships. This design enables YOLOv11 to more accurately identify and locate targets, especially in detection tasks with complex backgrounds or small targets [9].

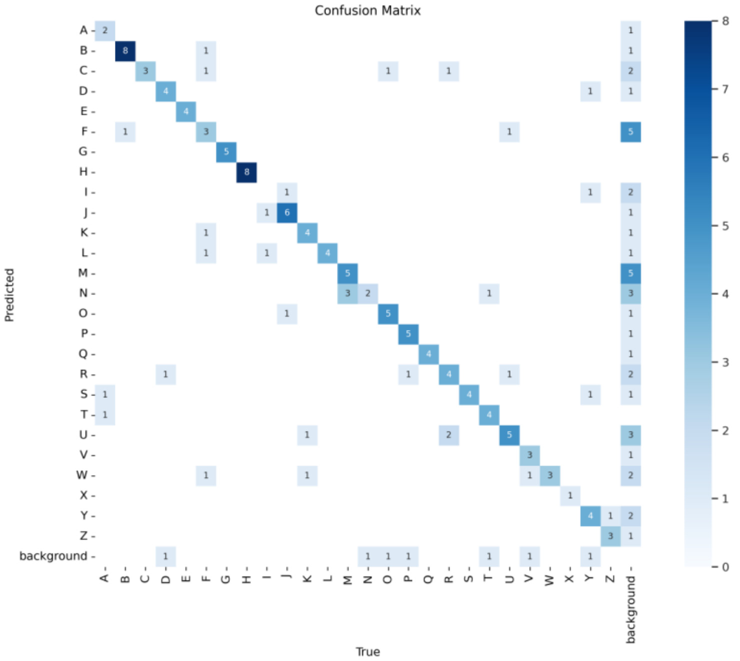
Secondly, YOLOv11 has also been optimized for multi-scale feature fusion. With the latest feature fusion techniques, YOLOv11 can effectively integrate feature information from different layers, enabling the model to simultaneously detect targets at multiple scales. This multi-scale detection mechanism not only improves the sensitivity of the model to small targets, but also enhances its applicability in different scenarios. Through this approach, YOLOv11 can handle the complexity of input images at different resolutions, achieving a combination of practicality and flexibility in various application scenarios.

In addition, YOLOv11 adopts advanced data augmentation techniques during the training process, especially adaptive data augmentation (ADA) methods, which can dynamically adjust the data augmentation strategy according to the development of the model, making the training process more efficient and robust. Through these enhancement measures, YOLOv11 can be trained under various environmental conditions, significantly improving the model's generalization ability, which directly affects the performance of the model in real-world applications [10].

4. Result

When conducting image object detection experiments using YOLOv11, the model parameters were set to 10 training rounds, the early stop patience value was set to 5, the batch size was set to 8, the input image size was imgsz=160, the model save was enabled with save=True, the cache type was cache=disk, the first GPU device was used with device=0, and the number of worker threads was set to 8.

Introducing the YOLOv8 model and dataset for training, first output the confusion matrix predicted by the model, as shown in Figure 4.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	background
A	2																										1
B		8				1																					1
C			3			1									1			1									2
D				4																					1		1
E					4																						
F		1				3															1						5
G							5																				
H								8																			
I										1															1		2
J									1	6																	1
K						1					4																1
L						1			1			4															1
M													5														5
N													3	2						1							3
O										1					5												1
P																5											1
Q																	4										1
R				1												1		4			1						2
S	1																		4						1		1
T	1																			4							
U											1							2			5						3
V																						3					1
W						1					1											1	3				2
X																								1			
Y																									4	1	2
Z																										3	1
background				1										1	1	1				1		1			1		

Figure 4: The confusion matrix.

According to the confusion matrix, the model can detect and recognize most gestures, with only a few cases experiencing recognition errors.

The changes in various parameters during the training process of the output model are shown in Figure 5.

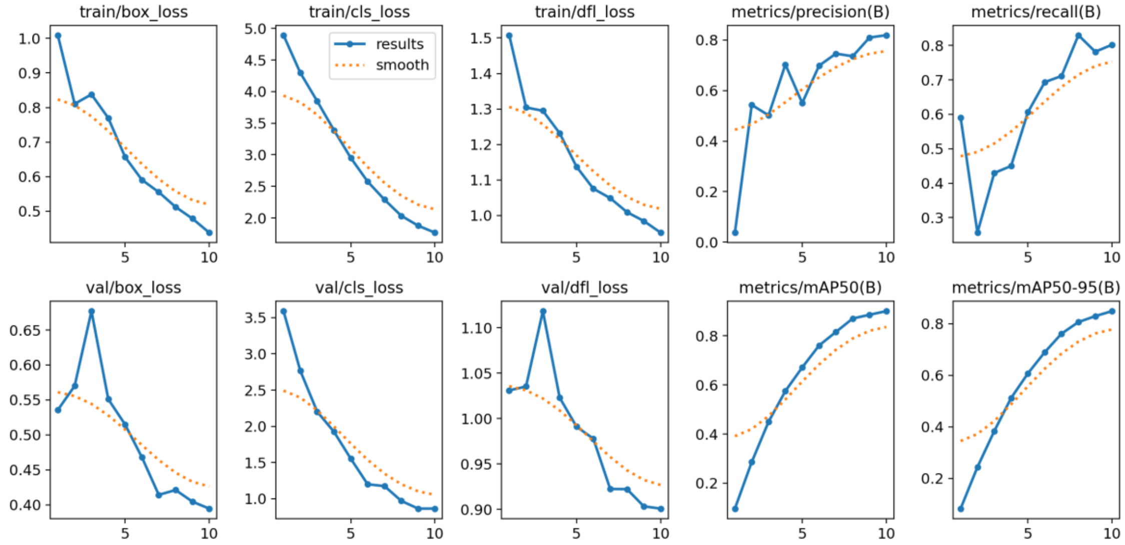


Figure 5: The changes in various parameters during the training process of the output model.

On train/box-loss, the bounding box loss shows a significant downward trend during the training process, indicating that the model is learning to locate targets more accurately.

On train/cls_loss, the classification loss (CLS loss) also showed a decreasing trend, indicating that the model gradually improved its recognition ability for the target category during the training process.

On train/dfl-loss, the distributed focus loss (dfl loss) is also decreasing, which may indicate an improvement in the accuracy of the model in predicting bounding boxes.

On metrics/precision (B), precision fluctuates during the training process, but the overall trend is upward, indicating that the model has improved in reducing false positives.

On metrics/recall (B), the recall rate fluctuates greatly during the training process, but overall shows an upward trend, indicating that the model has made progress in detecting more real targets.

On val/box-loss, the bounding box loss on the validation set fluctuates initially, but then shows a decreasing trend, which may indicate that the performance of the model on the validation set is also improving.

On val/cls_loss, the classification loss on the validation set fluctuates significantly in the initial stage, but gradually decreases afterwards, which may indicate that the model's classification performance on the validation set is gradually stabilizing and improving.

On val/dfl-loss, the distributed focus loss on the validation set fluctuates initially but then decreases, indicating that the model's bounding box prediction accuracy on the validation set may be improving.

On the metrics/mAP50 (B), the average accuracy (mAP) at an IoU threshold of 0.5 continued to increase during the training process, indicating an overall improvement in the model's performance in detection tasks.

On metrics/mAP50-95 (B), the average accuracy (mAP) of the IoU threshold from 0.5 to 0.95 is also increasing. Although the increase is not as significant as mAP50, it still indicates that the model's performance has improved under stricter evaluation criteria.

Use the images from the test set to detect the model, as shown in Figure 6. Use a model to recognize gestures in images and predict their meanings, and output confidence levels.

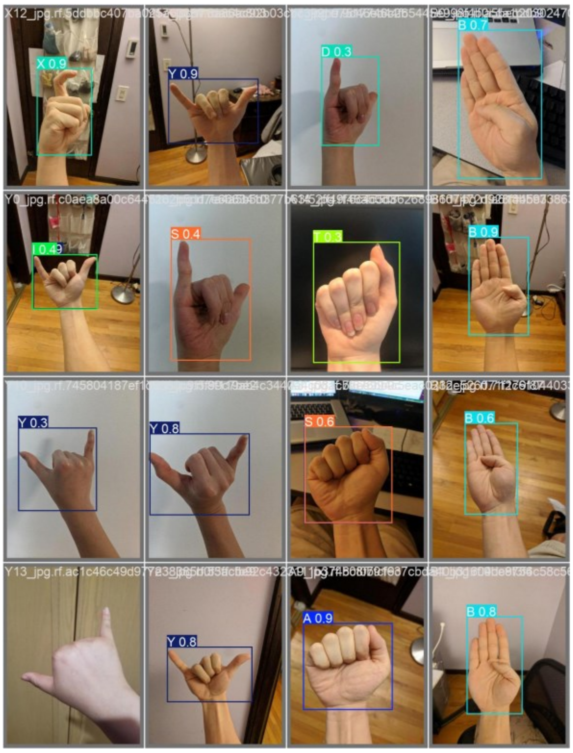


Figure 6: Test set results

According to the results of the test set, it can be seen that the YOLOv11 model in this article can recognize and predict gestures well, and accurately recognize the meaning represented by gestures. The accuracy of the model is very good.

5. Conclusion

In the field of human-computer interaction (HCI), gesture recognition technology achieves communication with computers and other intelligent devices by parsing human gestures and actions, which has important research value and application prospects [11]. This study adopts the latest progress in the YOLO series - YOLOv11 algorithm and conducts in-depth experimental research on gesture recognition tasks [12]. By introducing the YOLOv8 model and related datasets for training, we first generated the confusion matrix predicted by the model. The analysis results of the confusion matrix show that the YOLOv11 model can accurately detect and recognize gestures in most cases, although there may be recognition errors in certain specific situations [13].

During the training process, we observed a steady increase in the average accuracy (mAP) of the model when the IoU (Intersection over Union) threshold was 0.5, indicating a significant improvement in the overall performance of YOLOv11 in gesture detection tasks. In addition, when the IoU threshold was increased from 0.5 to 0.95, although the increase in mAP was small, it still maintained a growth trend, which further proves that YOLOv11 can still demonstrate good performance under stricter evaluation criteria [14].

In order to comprehensively evaluate the practical ability of the model, we used independent test set images to detect the YOLOv11 model [15]. The test results are encouraging, as the YOLOv11 model not only accurately recognizes gestures but also interprets the meaning behind them, demonstrating extremely high accuracy [16]. This achievement not only demonstrates the potential application of YOLOv11 in gesture recognition but also provides new ideas and methods for the development of future human-computer interaction technology [17].

In summary, this study successfully improved the accuracy and efficiency of gesture recognition by introducing the YOLOv11 algorithm. The YOLOv11 model has demonstrated excellent performance in gesture detection and recognition tasks, achieving high-precision gesture recognition on both the training and testing sets. This study not only provides strong technical support for the development of gesture recognition technology, but also provides valuable experience and reference for researchers in the HCI field [18]. With the continuous advancement and optimization of technology, we have reason to believe that YOLOv11 and its subsequent versions will play a more important role in the field of human-computer interaction, promoting more natural and efficient communication between smart devices and humans.

References

[1] Hussain, Muhammad. "YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection." Machines 11.7 (2023): 677.

[2] Chakravarthi, Bharathi Raja, et al. "Overview of the track on sentiment analysis for dravidian languages in code-mixed text." Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation. 2020.

[3] Talaat, Fatma M., and Hanaa ZainEldin. "An improved fire detection approach based on YOLO-v8 for smart cities." Neural Computing and Applications 35.28 (2023): 20939-20954.

[4] Terven, Juan, Diana-Margarita Córdova-Esparza, and Julio-Alejandro Romero-González. "A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas." Machine Learning and Knowledge Extraction 5.4 (2023): 1680-1716.

[5] Kim, Kyunghwan, Kangeun Kim, and Soyoon Jeong. "Application of YOLO v5 and v8 for Recognition of Safety Risk Factors at Construction Sites." Sustainability 15.20 (2023): 15179.

[6] Chakravarthi, Bharathi Raja, et al. "Corpus creation for sentiment analysis in code-mixed Tamil-English text." arxiv preprint arxiv:2006.00206 (2020).

[7] Sahafi, Ali, Anastasios Koulaouzidis, and Mehrshad Lalinia. "Polypoid lesion segmentation using YOLO-V8 network in wireless video capsule endoscopy images." Diagnostics 14.5 (2024): 474.

[8] Shi, Jiayou, et al. "Multi-Crop Navigation Line Extraction Based on Improved YOLO-v8 and Threshold-DBSCAN under Complex Agricultural Environments." Agriculture 14.1 (2023): 45.

[9] Liu, Hui, et al. "Research on Weed Reverse Detection Methods Based on Improved You Only Look Once (YOLO) v8: Preliminary Results." Agronomy 14.8 (2024): 1667.

[10] Cheng, Liang. "A Highly robust helmet detection algorithm based on YOLO V8 and Transformer." IEEE Access (2024).

[11] Chen, H., Shen, Z., Wang, Y. and Xu, J., 2024. Threat Detection Driven by Artificial Intelligence: Enhancing Cybersecurity with Machine Learning Algorithms.

[12] Liang, X., & Chen, H. (2019, July). A SDN-Based Hierarchical Authentication Mechanism for IPv6 Address. In 2019 IEEE International Conference on Intelligence and Security Informatics (ISI) (pp. 225-225). IEEE.

[13] Liang, X., & Chen, H. (2019, August). HDSO: A High-Performance Dynamic Service Orchestration Algorithm in Hybrid NFV Networks. In 2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS) (pp. 782-787). IEEE.

[14] Chen, H., & Bian, J. (2019, February). Streaming media live broadcast system based on MSE. In Journal of Physics: Conference Series (Vol. 1168, No. 3, p. 032071). IOP Publishing.

[15] Ke, Z., & Yin, Y. (2024). Tail Risk Alert Based on Conditional Autoregressive VaR by Regression Quantiles and Machine Learning Algorithms. arXiv preprint arXiv:2412.06193

[16] Ke, Z., Xu, J., Zhang, Z., Cheng, Y., & Wu, W. (2024). A Consolidated Volatility Prediction with Back Propagation Neural Network and Genetic Algorithm. arXiv preprint arXiv:2412.07223

[17] Yu, Q., Xu, Z., & Ke, Z. (2024). Deep Learning for Cross-Border Transaction Anomaly Detection in Anti-Money Laundering Systems. arXiv preprint arXiv:2412.07027

[18] Hu, Z., Lei, F., Fan, Y., Ke, Z., Shi, G., & Li, Z. (2024). Research on Financial Multi-Asset Portfolio Risk Prediction Model Based on Convolutional Neural Networks and Image Processing. arXiv preprint arXiv:2412.03618.