Implementation of a Convolutional Neural Network (CNN)-based Object Detection Approach for Smart Surveillance Applications

Weiguo Ni*
School of UAV, Guangzhou Civil Aviation College, Guangzhou 510000, Guangdong, China

***Abstract*—In the realm of smart surveillance systems, a fundamental technique for tracking and evaluating consumer behavior is object detection through video surveillance. While existing research underscores object detection through deep learning techniques, a notable gap exists in adapting these methods to effectively capture and recognize small, intricate objects. This study addresses this gap by introducing a customized methodology tailored to meet the nuanced requirements of accurate and lightweight detection for small objects, especially in scenarios prone to visual complexity and object similarity challenges. The primary objective is to furnish a vision-based object identification method designed for surveillance applications in smart stores, with a particular focus on locating jewelry objects. To achieve this, a Convolutional Neural Network (CNN)-based object detector utilizing YOLOv7 is employed for precise object detection and location extraction. The YOLOv7 network undergoes rigorous training and verification on a unique dataset specifically curated for this purpose. Experimental results affirm the efficacy of the proposed object identification method, demonstrating its capacity to detect items relevant to smart surveillance applications.**

Keywords*—Smart surveillance; lightweight object detection; YOLOv7; small object recognition; vision-based identification

I. Introduction

Cameras and image sensors are frequently deployed in smart surveillance systems so that automated object identification techniques may be used to automatically detect and identify various objects in smart environment analysis [1, 2]. Such automatic object recognition techniques often need sophisticated image/data processing tools and algorithms [3]. As a result, developing low-complexity automated object identification algorithms for use in urban surveillance applications becomes crucial [4, 5].

For computer vision applications, deep learning-based approaches, including Convolutional Neural Networks (CNNs), are among the finest solutions [6, 7]. Applications like object categorization and image segmentation have made significant strides thanks to CNNs [8]. Additionally, CNNs contain convolution layers that handle feature extraction; they are resilient to shifts and distortions in the image; they use less memory; training is simpler; and, as a result of the fewer parameters, they are better and quicker [9].

Object detection and monitoring in IoT Smart Shop Surveillance Systems have witnessed significant advancements in recent years. Current technologies utilize a combination of computer vision, IoT devices, and machine learning algorithms to enhance security, customer experience, and operational efficiency in retail environments. The integration of IoT devices enables real-time data collection from various sensors and cameras while object detection algorithms process this data to identify and track objects of interest.

In previous studies, various methods have been explored for object detection and monitoring in IoT Smart Shops [10, 11]. Traditional approaches, such as handcrafted features and rule-based algorithms, have limitations in handling complex and diverse scenarios. However, deep learning-based methods, particularly the CNNs, have gained immense popularity [12, 13]. Deep learning models can automatically learn and extract relevant features from raw data, making them capable of handling complex object detection tasks. The ability of deep learning models to handle large-scale datasets and their superior performance in terms of accuracy have attracted researchers to explore and develop new approaches based on these techniques.

Despite the advancements, there are still some limitations and research gaps in the field. One major challenge is the requirement for low computational costs and high accuracy rates [14, 15]. Many IoT devices have limited processing power and memory, making it necessary to develop lightweight deep-learning algorithms that can achieve high accuracy while maintaining computational efficiency. Additionally, the lack of publicly available datasets specifically designed for IoT Smart Shop Surveillance Systems poses another challenge for researchers.

To address these limitations, researchers have focused on developing lightweight deep learning models and utilizing algorithms like YOLO (You Only Look Once) for efficient object detection. YOLO-based algorithms offer real-time object detection capabilities with relatively low computational requirements [16, 17]. Using custom datasets, researchers can train these models on specific Smart Shop Surveillance scenarios, encompassing various objects and environmental factors.

In this study, 1170 images were collected for our custom dataset from the Internet and our capturing webcam-based process, and the image augmentation process in our custom dataset preparation. The dataset is used for training and evaluating a model based on the YOLOv7 network. This model has generated a weight and is used to perform object recognition using the YOLOv7 model on our custom dataset.

This study introduces novel contributions in the field of computer vision and deep learning for IoT Smart Shop Surveillance Systems. It innovates by developing a lightweight deep learning model tailored to the limited computational resources of IoT environments. The creation of a custom dataset, incorporating 1170 images with meticulous preparation, stands out as a unique aspect, emphasizing the study's commitment to robust methodology. The adoption of the YOLOv7 network architecture for object recognition further highlights the innovative application of state-of-the-art technologies to address surveillance challenges in a Smart Shop context.

In terms of research contributions, three potential areas of focus are identified. Firstly, the development of a lightweight deep learning model tailored for IoT Smart Shop Surveillance Systems, considering the low computational resources available. Secondly, the creation of a custom dataset that represents realistic scenarios encountered in Smart Shop Surveillance. This dataset can enable the training and evaluation of the proposed model. Finally, conducting thorough experimental evaluations to assess the performance of the model in terms of accuracy, real-time detection, and computational efficiency. By addressing the research gap through the proposed lightweight deep learning model, custom dataset, and rigorous performance evaluations, researchers can contribute to advancing object detection and monitoring in IoT Smart Shop Surveillance Systems. The outcome of this research can lead to improved security, customer experience, and operational efficiency in retail environments, promoting the widespread adoption of IoT-based surveillance systems in the retail industry.

The structure of this paper is as follows; Section I presents the introduction. The proposed approach discusses in Section II. Section III involves experimental results, and Section IV concludes the paper.

II. Related Works

This section reviews the related works that are focused on object detection in video-based surveillance systems.

Mneymneh et al. [18] introduced a vision-based framework for intelligent monitoring of hardhat wearing on construction sites. The framework utilizes computer vision techniques to detect and track the presence of hardhats on individuals within the construction site environment. It involves stages of image acquisition, pre-processing, detection, and monitoring to identify and track individuals wearing hardhats. The study's limitations include reliance on a specific color-based segmentation approach, vulnerability to challenging lighting conditions, and the absence of exploration of other safety equipment detection. Addressing these limitations can enhance the framework's accuracy and broaden its applicability in ensuring compliance with safety regulations on construction sites.

Lu et al. [19] presented a real-time object detection algorithm for video. The algorithm utilizes a combination of deep learning techniques, including Convolutional Neural Networks (CNNs) and feature extraction methods, to detect objects in video frames. The proposed algorithm achieves high detection accuracy and real-time performance by optimizing the architecture and leveraging parallel processing capabilities. However, the limitation of the study is that the algorithm's performance may be affected by complex scenes with occlusions or high object density, which can lead to missed detections or false positives. Further research could focus on improving the algorithm's robustness in challenging video scenarios to enhance its overall effectiveness in real-world applications.

The authors in [20] presented a methodology based on deep learning for object detection in video surveillance, specifically focusing on the identification of small objects that are handled similarly. The proposed methodology utilizes binary classifiers and leverages deep learning techniques to achieve accurate object detection. The approach demonstrates promising results in detecting small objects in challenging video surveillance scenarios. However, a limitation of the study is that the proposed methodology may face challenges when dealing with highly cluttered scenes or objects that have similar visual characteristics but different semantic meanings. Further research could explore techniques to address these limitations and improve the methodology's robustness in handling complex scenarios, ultimately enhancing its applicability in video surveillance applications.

Alrowais et al. [21] developed a deep transfer learning-enabled intelligent object detection approach for crowd density analysis in video surveillance systems. The proposed method utilizes deep learning techniques and transfers learning to detect and analyze crowd density in video footage. By leveraging pre-trained models and fine-tuning them on specific crowd density datasets, the approach achieves accurate object detection and density analysis. The results demonstrate the effectiveness of the method in crowd density estimation. However, a limitation of the study is the reliance on pre-trained models that may not fully capture the diverse range of crowd dynamics and behaviors. Further research could focus on developing customized models or incorporating additional data augmentation techniques to improve the algorithm's performance and robustness in capturing different crowd scenarios.

According to review of previous studies, Addressing the research challenge of achieving high accuracy and lightweight object detection, particularly for small objects like jewelry, requires a tailored approach. While existing studies focus on object detection using deep learning techniques, adapting these methods to effectively capture and recognize small, intricate objects remains a gap. Current methodologies may struggle in cluttered scenes or with objects sharing similar visual characteristics.

III. Proposed Approach

This study selects jewelry objects and implements the approach for these kinds of objects involving rings and earnings. We suggest a detection method based on YOLOv7 to create a model that could recognize jewelry [9].

A. YOLO

The YOLO model, which stands for "You Only Look Once," is one stage object detector [17]. YOLO predicts the

positions of the bounding boxes and the classes of the bounding boxes [19]. Objectness of the bounding boxes after feature-stripping image frames through a backbone, combining and blending features in the neck, and objectness prediction in the head of the network [22]. To arrive at its ultimate forecast, YOLO employs post-processing using NMS [23]. Fig. 1 illustrates the basic concept of YOLO network architecture.

B. Dataset

Our dataset consists of two various categories of jewelry (earrings and rings), as mentioned earlier. The dataset includes images from the Internet and our captured images in the real environment. We gathered webcam photos from two fixed cameras that were placed in various places. We chose images with a variety of model kinds, sizes, resolutions, orientations, and sample counts in each picture. One thousand one hundred seventy example photos of two types of jewelry pieces in various orientations, rotations, and scales are included in our unique dataset. Some sample images from our dataset are shown in Fig. 2.

C. Training and Testing

It is usually a good idea to start with a model that has already been trained using extremely big datasets and then utilize the model's weights to train an object detector [24]. Even if the learned weights don't contain the items needed for this specific experiment, to deal with the training process in the YOLO network, initial weights are taken from a pretrained model that includes weights from the known dataset [25]. In this study, we use an initial weight that trained a model from the COCO dataset.

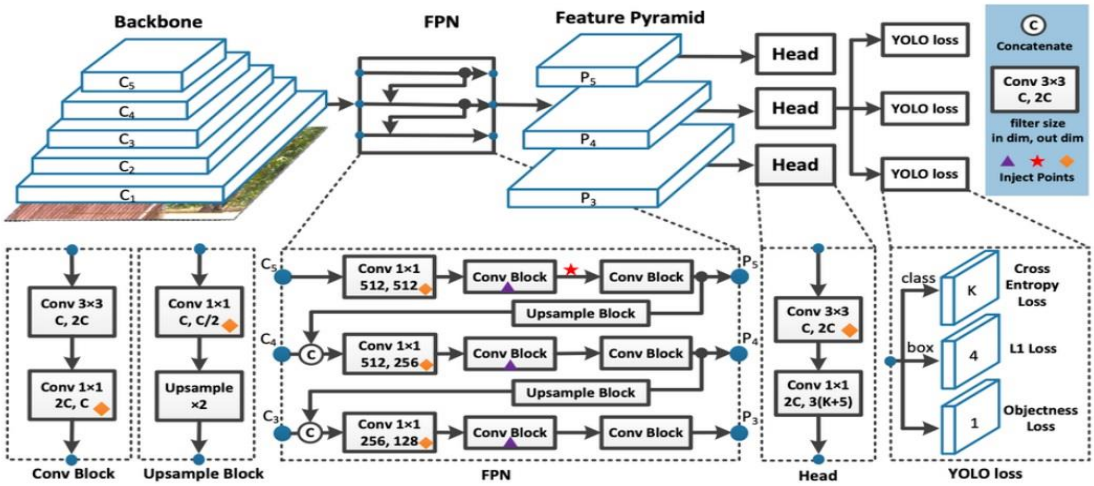


Fig. 1. YOLO network architecture [23].

Fig. 2. Some image samples from the dataset.

IV. Experimental Results

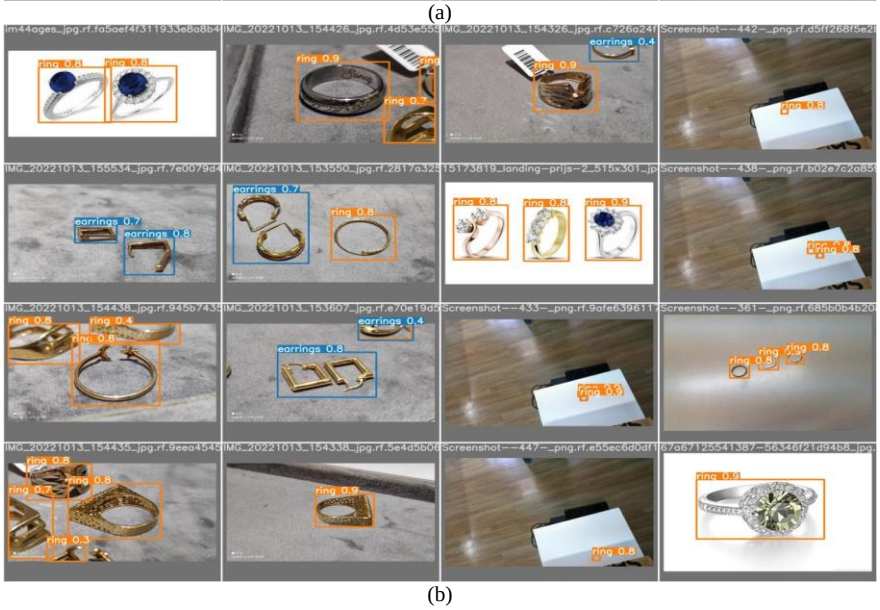
In this section, we present the experiment's details, and then we show the training results using pretraining weights and compare the three models of YOLOv7. Fig. 3 shows the result of the implementation of our proposed approach.

In the following, some analyses are presented to justify why the YOLOv7 can be presented accurate results and has superiority to apply in real-time requirements. To prove this superiority, some visual representations in graphs are illustrated. Using these graphs, a comparison of performance results is shown to visually demonstrate the superiority justifications. Inspired by [26], the graph depicts a comparison of YOLOv7, YOLOv4, PPYOLOE, YOLOX, YOLOR, YOLOv5, and Transformers object detectors. The X-axis represents the Inference Time, indicating the time taken for object detection, while the Y-axis represents the average precision (%) of the detectors.

Average precision (AP) is a commonly used metric in object detection that measures the accuracy of a model in localizing and classifying objects. It calculates the precision at various recall levels, considering the trade-off between precision and recall. Higher AP values indicate better performance and accuracy in object detection.

Inference Time refers to the time taken by the model to perform object detection on a given input. It reflects the efficiency and speed of the algorithm in processing frames or images in real-time applications. Lower inference times are desirable for real-time object detection scenarios. Fig. 4 presents the comparison of object detectors [26].

(a)



(b)

Fig. 3. Experimental results (a) and (b).

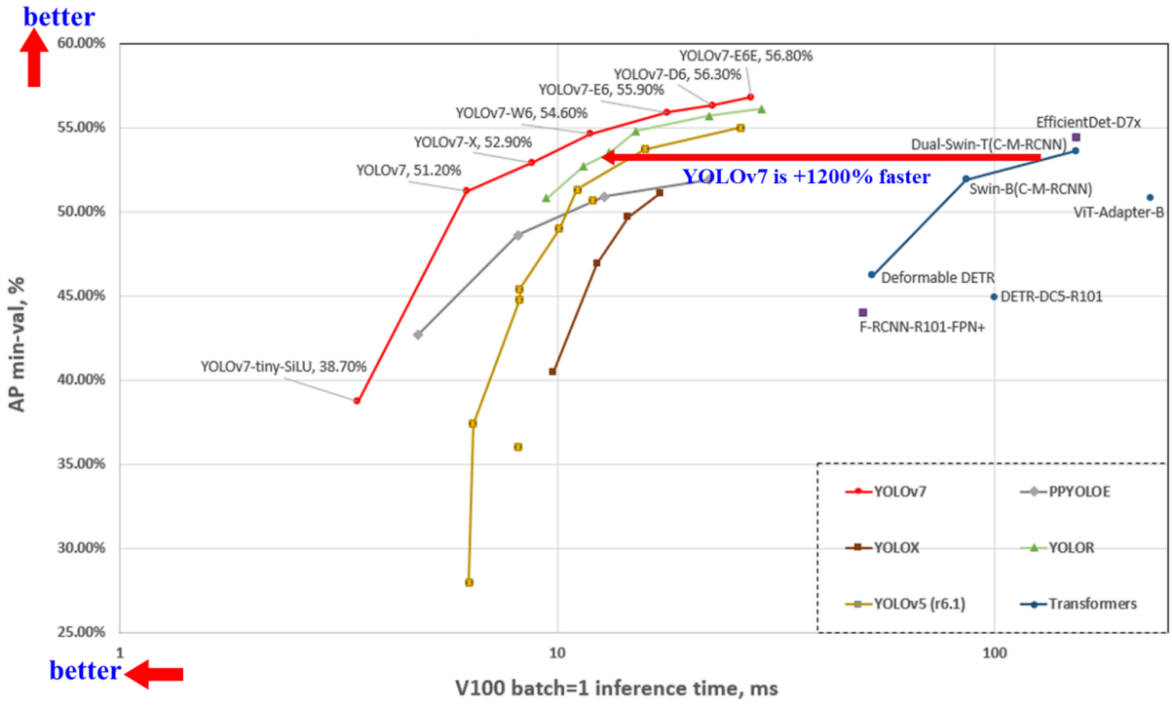


Fig. 4. Comparison of object detectors [26].

As shown in Fig. 4, it can be observed that YOLOv7 outperforms the other object detectors in terms of precision rate. At similar inference times, YOLOv7 consistently achieves higher average precision compared to PPYOLOE, YOLOX, YOLOR, YOLOv5, and Transformers. This suggests that YOLOv7 demonstrates superior accuracy in detecting and recognizing objects across various scenarios.

The YOLOv7 utilizes an efficient single-pass detection pipeline that eliminates the need for time-consuming region proposal techniques. This allows YOLOv7 to process images and videos in real-time without compromising accuracy. Other detectors may achieve lower inference times but often at the expense of reduced precision. Furthermore, YOLOv7 incorporates advanced training strategies, including data augmentation techniques and optimization methods like focal loss and learning rate scheduling. These strategies enhance the model's ability to generalize and accurately detect objects under diverse conditions, contributing to its superior precision rate.

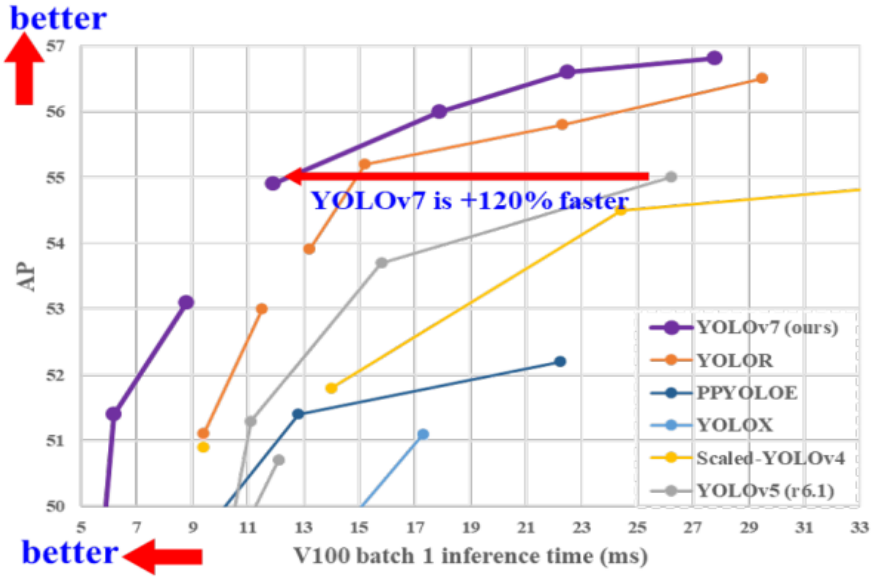
As a result, the graph demonstrates that YOLOv7 outperforms PPYOLOE, YOLOX, YOLOR, YOLOv5, and Transformers in terms of precision rate. The advanced architecture, efficient detection pipeline, and advanced training strategies employed by YOLOv7 contribute to its effectiveness and accuracy in object detection.

In the following, performance analysis and comparison of AP are presented to justify why the YOLOv7 is better than other object detector algorithms. Fig. 4 demonstrates the comparison of object detectors in real-time [26].

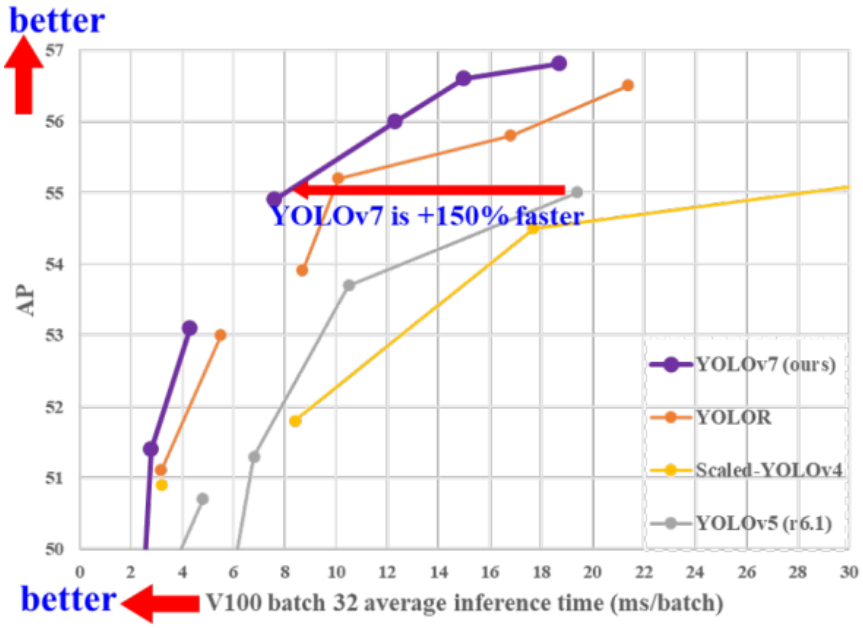
Fig. 5 illustrates the comparison of object deters in real-time conditions based on the presented comparison graphs. The graph illustrates a comparison of real-time object detection algorithms, including YOLOv7, PPYOLOE, YOLOX, YOLOR, Scaled-YOLOv4, YOLOv5, and Transformers. The X-axis represents the Inference Time, indicating the time taken for object detection, while the Y-axis represents the average precision (AP) of the detectors.

As shown in Fig. 4, it can be observed that YOLOv7 performs better than the other object detection algorithms in terms of precision rate while maintaining real-time capabilities. YOLOv7 achieves higher AP values at similar inference times compared to PPYOLOE, YOLOX, YOLOR, Scaled-YOLOv4, YOLOv5, and Transformers. This indicates that YOLOv7 provides more accurate object detection results. Furthermore, YOLOv7 demonstrates its effectiveness in real-time conditions by achieving low inference times while maintaining high precision. It strikes a balance between accuracy and speed, making it suitable for real-time applications where accuracy and efficiency are crucial.

As depicted in Fig. 4 and Fig. 5, the comparative analysis highlights the exceptional performance of YOLOv7 in the realm of object detection. The YOLOv7 stands out by demonstrating superior precision rates while seamlessly maintaining real-time capabilities. The discernible advantage lies in its ability to achieve higher Average Precision (AP) values when compared to a spectrum of other prominent object detection algorithms, including PPYOLOE, YOLOX, YOLOR, Scaled-YOLOv4, YOLOv5, and Transformers. This distinction is particularly noteworthy as it underscores YOLOv7's efficacy in delivering heightened accuracy without compromising on the efficiency required for real-time applications. The observed performance superiority positions YOLOv7 as a compelling choice for tasks demanding both precision and prompt response, further solidifying its standing as a leading solution in the field of object detection algorithms.



(a)



(b)

Fig. 5. Comparison of object detectors in real-time.

Finally, the graph demonstrates that YOLOv7 outperforms PPYOLOE, YOLOX, YOLOR, Scaled-YOLOv4, YOLOv5, and Transformers in terms of precision rate while maintaining real-time capabilities. Its advanced architecture, coupled with efficient inference times, makes YOLOv4 an effective and efficient choice for real-time object detection applications.

V. Conclusion

In this paper, a robust object identification technique based on YOLOv7 is developed for applications in smart surveillance. The first step involves the meticulous preparation of a custom dataset, with labels configured to adhere to the YOLOv7 format. The chosen technique demonstrates remarkable accuracy in identifying and categorizing jewelry objects, specifically rings and money. The network efficiently captures coordinates of resulting bounding boxes, enabling precise object identification within frames. While the current study focuses on training YOLOv7 for two types of jewelry, future research directions could involve expanding the model to encompass a broader spectrum of jewelry classes. This extension would enhance the model's versatility and applicability in diverse contexts within the realm of smart surveillance. Additionally, exploring real-time applications of the YOLOv7-based methodology remains an open problem, presenting an avenue for further investigation into its efficiency and effectiveness in dynamic surveillance scenarios. Moreover, investigating strategies to improve the model's adaptability to varying lighting conditions and complex backgrounds stands as

a potential area for future research, contributing to the refinement of its performance in real-world surveillance applications.

ACKNOWLEDGMENT

This work was supported by Youth Innovation Talent Project of Department of Education of Guangdong Provincial (2023)-Uav Fi intelligent monitoring system based on deep learning. (Item Number：2023KQNCX158)

REFERENCES

[1] M. Mukherjee, I. Adhikary, S. Mondal, A. K. Mondal, M. Pundir, and V. Chowdary, "A vision of IoT: applications, challenges, and opportunities with Dehradun perspective," in Proceeding of international conference on intelligent communication, control and devices, 2017: Springer, pp. 553-559.

[2] G. T. S. Ho, Y. P. Tsang, C. H. Wu, W. H. Wong, and K. L. Choy, "A computer vision-based roadside occupation surveillance system for intelligent transport in smart cities," Sensors, vol. 19, no. 8, p. 1796, 2019.

[3] I. Saradopoulos, I. Potamitis, S. Ntalampiras, A. I. Konstantaras, and E. N. Antonidakis, "Edge Computing for Vision-Based, Urban-Insects Traps in the Context of Smart Cities," Sensors, vol. 22, no. 5, p. 2006, 2022.

[4] L. Hu and Q. Ni, "IoT-driven automated object detection algorithm for urban surveillance systems in smart cities," IEEE Internet of Things Journal, vol. 5, no. 2, pp. 747-754, 2017.

[5] M. J. Akhtar et al., "A Robust Framework for Object Detection in a Traffic Surveillance System," Electronics, vol. 11, no. 21, p. 3425, 2022.

[6] K. Bjerge, H. M. Mann, and T. T. Høye, "Real - time insect tracking and monitoring with computer vision and deep learning," Remote Sensing in Ecology and Conservation, vol. 8, no. 3, pp. 315-327, 2022.

[7] A. Aghamohammadi, M. C. Ang, E. A. Sundararajan, K. W. Ng, M. Mogharrebi, and S. Y. Banihashem, "Correction: A parallel spatiotemporal saliency and discriminative online learning method for visual target tracking in aerial videos," Plos one, vol. 13, no. 3, p. e0195418, 2018.

[8] J. Du, "Understanding of object detection based on CNN family and YOLO," in Journal of Physics: Conference Series, 2018, vol. 1004: IOP Publishing, p. 012029.

[9] M. Hatab, H. Malekmohamadi, and A. Amira, "Surface defect detection using YOLO network," in Proceedings of SAI Intelligent Systems Conference, 2020: Springer, pp. 505-515.

[10] J. Xu et al., "Design of smart unstaffed retail shop based on IoT and artificial intelligence," IEEE Access, vol. 8, pp. 147728-147737, 2020.

[11] L. Sharma and N. Lohan, "Internet of things with object detection: Challenges, applications, and solutions," in Handbook of Research on Big Data and the IoT: IGI Global, 2019, pp. 89-100.

[12] A. Hazarika, S. Poddar, M. M. Nasralla, and H. Rahaman, "Area and energy efficient shift and accumulator unit for object detection in IoT applications," Alexandria Engineering Journal, vol. 61, no. 1, pp. 795-809, 2022.

[13] W. Cao et al., "CNN-based intelligent safety surveillance in green IoT applications," China Communications, vol. 18, no. 1, pp. 108-119, 2021.

[14] W. Fang, L. Wang, and P. Ren, "Tinier-YOLO: A real-time object detection method for constrained environments," IEEE Access, vol. 8, pp. 1935-1944, 2019.

[15] G. Wang, H. Ding, Z. Yang, B. Li, Y. Wang, and L. Bao, "TRC - YOLO: A real - time detection method for lightweight targets based on mobile devices," IET Computer Vision, vol. 16, no. 2, pp. 126-142, 2022.

[16] A. A. Mei Choo Ang, Kok Weng Ng, Elankovan Sundararajan, Marzieh Mogharrebi, Teck Loon Lim, "Multi-core Frameworks Investigation on A Real-Time Object Tracking Application," Journal of Theoretical & Applied Information Technology, 2014.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779-788.

[18] B. E. Mneymneh, M. Abbas, and H. Khoury, "Vision-based framework for intelligent monitoring of hardhat wearing on construction sites," Journal of Computing in Civil Engineering, vol. 33, no. 2, p. 04018066, 2019.

[19] S. Lu, B. Wang, H. Wang, L. Chen, M. Linjian, and X. Zhang, "A real-time object detection algorithm for video," Computers & Electrical Engineering, vol. 77, pp. 398-408, 2019.

[20] F. Pérez-Hernández, S. Tabik, A. Lamas, R. Olmos, H. Fujita, and F. Herrera, "Object detection binary classifiers methodology based on deep learning to identify small objects handled similarly: Application in video surveillance," Knowledge-Based Systems, vol. 194, p. 105590, 2020.

[21] F. Alrowais et al., "Deep Transfer Learning Enabled Intelligent Object Detection for Crowd Density Analysis on Video Surveillance Systems," Applied Sciences, vol. 12, no. 13, p. 6665, 2022.

[22] T. Diwan, G. Anirudh, and J. V. Tembhurne, "Object detection using YOLO: challenges, architectural successors, datasets and applications," Multimedia Tools and Applications, pp. 1-33, 2022.

[23] Roboflow. "YOLOv7 Breakdown." https://blog.roboflow.com/pp-yolo-beats-yolov4-object-detection/ (accessed).

[24] C. Liu, Y. Tao, J. Liang, K. Li, and Y. Chen, "Object detection based on YOLO network," in 2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC), 2018: IEEE, pp. 799-803.

[25] D. Garg, P. Goel, S. Pandya, A. Ganatra, and K. Kotecha, "A deep learning approach for face detection using YOLO," in 2018 IEEE Punecon, 2018: IEEE, pp. 1-4.

[26] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7464-7475.