DLINE JOURNALS

Journal of Digital Information Management

JDIM

ISSN 0972-7272

A Focus on the Deep Learning-based Intelligent Video Surveillance System

Weigang Zhang*, Youzi Li
Xi'an Mingde Institute of Technology
xi'anShaanxi,710124,China
zhangweig_23@126.com

ABSTRACT: *This paper focuses on applying a deep learning-based intelligent video surveillance system, particularly emphasising using the YOLOv7 model for object detection. By reviewing the development of intelligent video surveillance technology, we recognize the importance of deep learning in computer vision. The structure and characteristics of the YOLOv7 model are detailed, including the input layer, backbone network layer, feature fusion layer, and output layer. To validate the model's performance, we conducted experiments on the VOC dataset, and the results show that the YOLOv7 model achieved an average detection accuracy of 0.89 on this dataset. The experimental results demonstrate the efficiency, accuracy, and robustness of the YOLOv7 model in intelligent video surveillance, providing essential references for optimizing and applying intelligent video surveillance systems.*

Subject Categories and Descriptors: [I.2.11 Distributed Artificial Intelligence]; I.2.10 [Vision and Scene Understanding]: Video analysis; [I.5 PATTERN RECOGNITION]; I.5.1 Models

General Terms: Energy Vehicle Charging Models, Genetic Algorithms, Algorithm Applications

Received: 10 April 2024, Revised 21 August 2024, Accepted 6 September 2024

Keywords: Deep Learning, Intelligent Video Surveillance System, Object Detection, YOLOv7 Model

Review Metrics: 0/6; Review Score: 4.65; Inter-reviewer Consistency: 81.4%

DOI: https://10.6025/jdim/2024/22/4/130-136

1. Introduction

Intelligent video surveillance systems are important applications in modern security, combining video monitoring technology with artificial intelligence to achieve real-time target detection, tracking, and analysis. With the rapid development of computer vision and deep learning technology, intelligent video surveillance systems have shown tremendous potential and prospects in security monitoring, traffic management, smart city construction, and other fields. Traditional video surveillance systems often rely on manual operations, which are inefficient and susceptible to subjective factors. The rise of deep learning technology, especially break throughs in object detection algorithms, brings new opportunities for developing intelligent video surveillance systems. Object detection is a crucial research direction

in computer vision, aiming to accurately locate and recognize the objects of interest from images or videos [1]. In object detection tasks, the You Only Look Once (YOLO) series models have received significant attention due to their high efficiency in real-time detection and accurate performance. The YOLO (You Only Look Once) series [2-9] of models, which utilize end-to-end detection, provide prediction outcomes directly. Their remarkable detection performance has made them a favoured framework for various industrial uses. Currently, YOLO has undergone multiple iterations, and YOLOv7, as a new version, is expected to have extensive applications in intelligent video surveillance systems [2]. The YOLOv7 model accomplishes object detection tasks through deep Convolutional Neural Networks (CNNs). Compared to other object detection algorithms, YOLOv7 maintains fast detection speed while preserving high detection accuracy, making it an auspicious choice for intelligent video surveillance systems.

This study explores the application of the YOLOv7 model in common object detection tasks within intelligent video surveillance systems. By deeply analyzing the performance and advantages of the YOLOv7 model inobject detection, it is expected to achieve efficient and accurate detection and tracking of targets in intelligentvideo surveillance systems [3]. Through experiments and analysis in this study, we hope to promote the development of intelligent video surveillance systems in fields such as security monitoring and urban management, making positive contributions to social security, stability, and intelligent construction. In summary, research on applying a deep learning-based intelligent video surveillance system, especially in common object detection using the YOLOv7 model, has important theoretical and practical significance. This research is expected to drive technological advancements in intelligent video surveillance systems, bringing more innovative applications to fields like security and urban management and becoming vital for smart city construction.

2. Intelligent Video monitoring systems

Video monitoring represents a dynamic field of study. Detection and tracking of objects in video surveillance systems typically rely on background estimation and subtraction methods. The main emphasis of current video surveillance systems is on utilizing video compression technology to effectively combine or save images from numerous cameras onto large storage devices.

The smart video monitoring system designed to trigger alerts during unusual events in the footage is categorized into two approaches: one utilizing computers and the other employing compact devices. The computer-driven smart video monitoring system performs a range of detections and has been the subject of research. Nevertheless, these systems carry installation and upkeep expenses, significant power usage, and risks of personal data breaches, making them less suitable for practical application in real-world settings. Camera plurality influences object tracking in intelligent video surveillance systems. While designing intelligent video systems, Zhang used GPU to reduce processing time and conduct distributed processing in the IP networks. [4]. The sparse random projection (SRP) is an advanced model that uses *scikit-learn* reduced processing time through image compression that projects high-dimensional video frames into low-dimensional partial spaces. [5]

2.1. Object Detection in the Video Surveillance Systems

As mentioned by Jie Xu [6], deep-learning methods for both object detection have many obvious benefits. Deep-learning techniques are simpler to implement and offer superior scalability compared to traditional machine-learning approaches. [7]. Deep-learning techniques can learn more intricate characteristics by utilizing various levels of representation. In the face recognition process, FaceNet [8] with Multi-task Cascaded Convolutional Networks (MTCNN), Zhange et al [9] found to produce significantly high accuracy without being too hardware-consuming. Numerous current object detection algorithms demonstrate encouraging levels of accuracy. [10-14]

3. Current Research Status of Object Detection Based on Video Surveillance

The application research of intelligent video surveillance systems at home and abroad is thriving, and significant progress has been made in deep learning technology. Countries worldwide attach great importance to the development of intelligent video surveillance and have initiated multiple intelligent video surveillance projects. The Defense Advanced Research Projects Agency (DARPA) in the United States jointly conducted the VSAM project with several prestigious domestic universities. This project, led by Carnegie Mellon University, primarily focuses on researching human behavior recognition technology applicable to battlefields and public places. The VSAM system obtains the positions of individuals and performs body part segmentation to establish features describing human postures, ultimately predicting and assessing human actions based on these features [15]. In Europe, the Engineering and Physical Sciences Research Council (EPSRC) in the United Kingdom has funded the BEHAVE project, which is dedicated to predicting abnormal or crime-oriented behaviors based on video sequences. The project aims to distinguish normal and abnormal behaviors in crowds by detecting, understanding, and differentiating similar types of interactions and analyzing crowd scenes [16]. Additionally, the French National Institute for Research in Computer Science and Automation (INRIA) has led the WILLOW group, focusing on research in object

and scene understanding, aiming to address the ultimate scientific challenges in computer vision and apply real-time and reliable object detection and scene understanding technologies to defense, entertainment, healthcare, human-computer interaction, and other fields [17]. INRIA also led another project called Pulsar, which focuses on activity recognition and mainly researches human, animal, or vehicle behaviors. The intelligent video surveillance service market in China is a vast and diverse field with broad demand. Although China started relatively late in this field, it has made rapid progress and is currently in a high-speed development stage. Many universities and organizations have invested substantial human and material resources in intelligent video surveillance technology, achieving remarkable results. For example, Professor Shan Shiguang of the Chinese Academy of Sciences developed and open-sourced a face recognition system, laying the foundation for domestic face recognition technology research [18]. Chinese universities have also made significant progress in intelligent video-processing technology. Computer vision research laboratories in universities such as Shanghai Jiao Tong University, Tsinghua University, and Zhejiang University have been dedicated to the research of intelligent video processing technology, contributing to the development of this field. Nanjing University even established the first artificial intelligence college in China, further promoting the development of intelligent video processing technology. The academic exchange environment in China is also very active. Renowned journals such as the Journal of Computer Science and Technology and the Journal of Automation have published many high-quality papers on target detection, behavior recognition, and other technologies, providing a good communication platform for domestic scholars. In China, technology companies focusing on intelligent video surveillance systems are rapidly developing, with two companies, Hikvision and Dahua Technology, ranking among the top five globally in annual revenue and growth rate [19]. These companies offer leading monitoring equipment and overall solutions, providing reliable support for constructing and applying intelligent video surveillance systems. Moreover, companies like SenseTime, Megvii, and iFlytek, specializing in the research and development of artificial intelligence products, have also performed prominently in intelligent video surveillance, securing multiple financing rounds [20]. Their facial recognition services are widely used in major airports and train stations for passenger identity verification, ensuring passenger safety while saving time during boarding and enhancing the verification efficiency of airports and train stations. In conclusion, intelligent video surveillance technology has enormous potential for development in the Chinese market and has already achieved significant progress. The active investments and collaborations of universities, academic organizations, and technology companies have laid a solid foundation for intelligent video surveillance systems research, development, and application. With continuous technological advancements and the expansion of application scenarios, intelligent video surveillance systems are expected to play a more significant role in various fields, providing more intelligent solutions for social security and management.

4. Intelligent Video Surveillance Model Based on YOLOv7

4.1. Theoretical Framework of YOLOv7 Network Model

The YOLO (You Only Look Once) model is a deep learning-based object detection algorithm proposed by Joseph Redmon and his team in 2015. Its uniqueness lies in transforming the object detection task into a single end-to-end regression problem, achieving real-time and efficient object detection. The origins of YOLO can be traced back to the R-CNN series algorithms. In R-CNN, candidate regions are first extracted using methods such as selective search, and then each region is classified and located. Although R-CNN made significant progress in accuracy, it was slow and unsuitable for real-time applications. YOLO emerged as a solution by directly predicting bounding boxes and class probabilities using convolutional neural networks, avoiding the complex process of generating candidate regions. The YOLO model divides the input image into a grid, and each grid predicts a fixed number of bounding boxes and their corresponding class probabilities, achieving end-to-end object detection [21]. The YOLO model has been continuously evolving. YOLOv2 (YOLO9000) introduced technologies such as Anchor Box and Darknet-19 based on YOLO, improving detection accuracy. YOLOv3 introduced multi-scale prediction, FPN (Feature Pyramid Network), and other techniques to enhance detection accuracy and applicability. The widely used model is YOLOv7, which performs well in both detection speed and accuracy.

The YOLOv7 model is an object detection algorithm composed of four network levels. First, the input layer defines the size of the original three-channel image as 640x640. Next, the backbone network layer adopts deep convolution to extract features at different scales. The feature fusion layer fuses the feature maps from different scales in-depth to enhance the feature representation ability further. Finally, after non-maximum suppression, the output layer outputs the predicted anchor box coordinates, classes, and confidences. The design of the YOLOv7 model includes multiple highly integrated convolutional operation modules: the ELAN module, MP module, and SPPCSPC module. These integrated structures are vital in optimizing the model structure and enhancing feature extraction capabilities, enabling the YOLOv7 model to perform well in object detection tasks. The ELAN module effectively increases information propagation and importance weighting of features by introducing additional long connections and attention mechanisms, enhancing the model's perception ability. The MP module uses multi-scale pooling

operations to capture object information at different scales, improving the model's adaptability to different-sized objects. The SPPCSPC module combines spatial and channel pyramid pooling, further enhancing feature representation and equipping the model with better perception and recognition accuracy. The YOLOv7 model structure is illustrated in Figure 1.

The YOLOv7 model combines four network levels: input layer, backbone network layer, feature fusion layer, and output layer. It fully utilizes the advantages of integrated modules such as ELAN, MP, and SPPCSPC, achieving efficient and accurate performance in object detection tasks. This makes the YOLOv7 model a highly regarded advanced algorithm in object detection, demonstrating significant advantages and application potential in practical use.

4.2. Intelligent Video Object Detection Model based on YOLOv7

The process of using the YOLOv7 model for intelligent video object detection can be divided into the following steps. First, it is necessary to prepare a video dataset for training and testing. The dataset should contain multiple video files that cover various scenarios with target objects to be detected. Each video frame needs to be annotated with bounding boxes indicating the positions of the objects and their corresponding class labels. Data preprocessing is required before starting the training. Preprocessing includes operations such as resizing the images, data augmentation (e.g., random cropping, rotation, flipping), and normalization to better adapt to the model's input requirements and enhance the diversity of data samples. Before training, the YOLOv7 model needs to be constructed. The structure of the YOLOv7 model includes components such as the input layer, backbone network layer, feature fusion layer, and output layer. Depending on the requirements of the objects detection task, hyperparameters and layers of the model can be adjusted to achieve better detection performance. The preprocessed data is then fed into the YOLOv7 model for training. During training, the model continuously adjusts its parameters through backpropagation algorithms to make the predicted results as close as possible to the ground truth annotations. Cross-entropy loss function and gradient descent algorithms are typically used for model training. Fine-tuning of the model is required during the training process. Techniques such as learning rate decay, transfer learning, and model fusion can improve the model's performance and generalization capability. After training, the model must be validated and evaluated using test data. Test videos are input to the model, and the model outputs detection results, including the positions and class labels of the detected objects. By comparing the prediction results with the ground truth annotations, performance metrics such as accuracy,

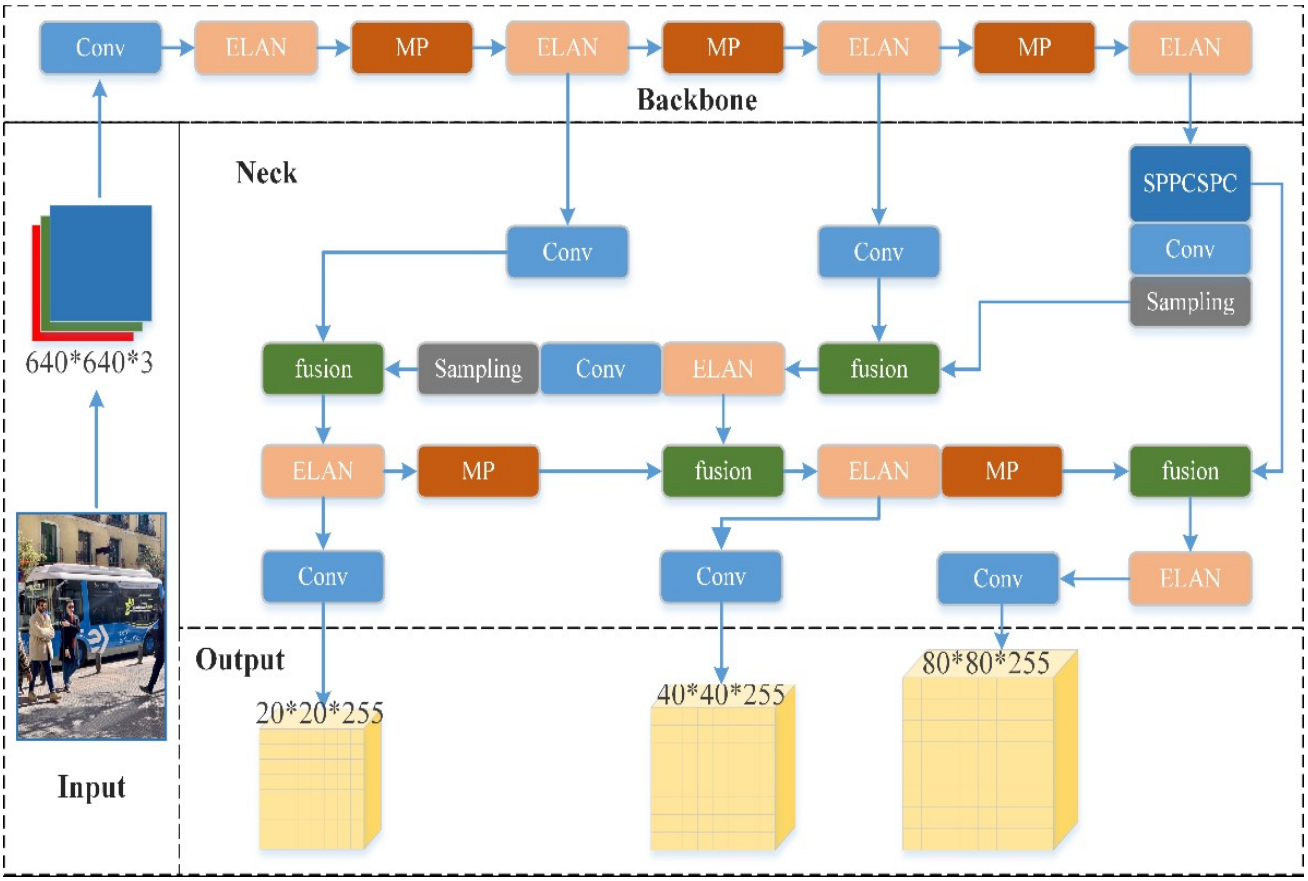

Figure 1 . The Theoretical Structure of YOLOv7 Model

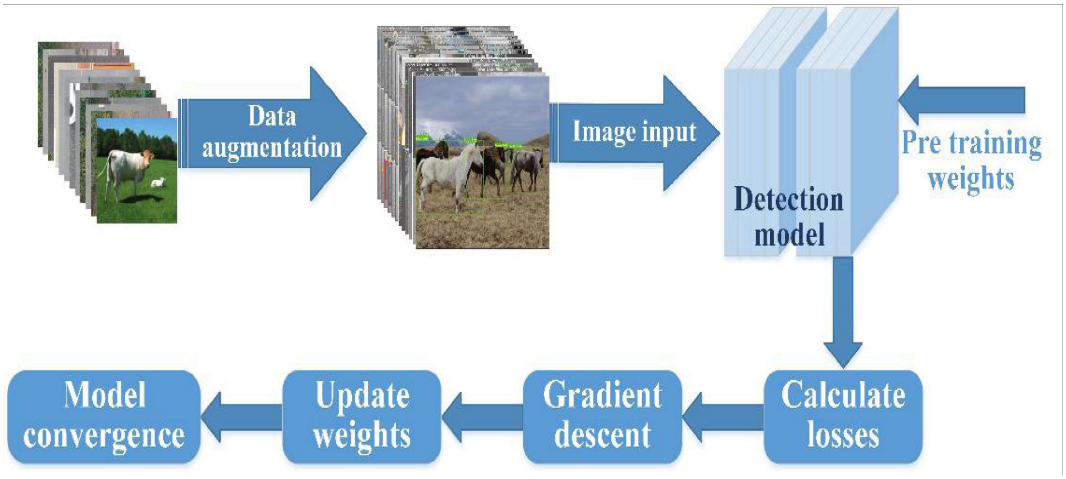

Figure 2. Flowchart of Intelligent Object Detection based on YOLOv7 Model

recall, and F1-score are calculated to evaluate the model's performance. The process of using this model for defect detection is shown in Figure 2.

Finally, the trained YOLOv7 model is applied to the intelligent ideo surveillance system. The model is embedded into the video surveillance system to perform real-time object deetection and tracking on video streams. The detected target information can be used in security monitoring, traffic management, and people counting applications.

5. Intelligent Object Detection based on YOLOv7 Model

This paper aims to research the application of intelligent video surveillance systems based on deep learning chnology, specifically using the YOLOv7 model. To verify the model's performance, the VOC dataset is selected for model evaluation. The VOC dataset is a widely used public dataset in object detection, containing annotated images of various object categories in real scenes. It is a classic object detection dataset that includes 20 different object categories. These categories cover common objects and scenes in daily life, such as aeroplanes, bicycles, birds, boats, bottles, buses, cars, cats, chairs, cows, dining tables, dogs, horses, motorcycles, people, potted plants, sheep, sofas, trains, and televisions. Each object category corresponds to objects in real scenes, such as cars, animals, and furniture. The diversity and richness of the VOC dataset make it an important benchmark for object detection algorithm research and evaluation. Many well-known object detection models are trained and tested on this dataset. By accurately detecting these object categories, the intelligent video surveillance system can achieve more precise and efficient object

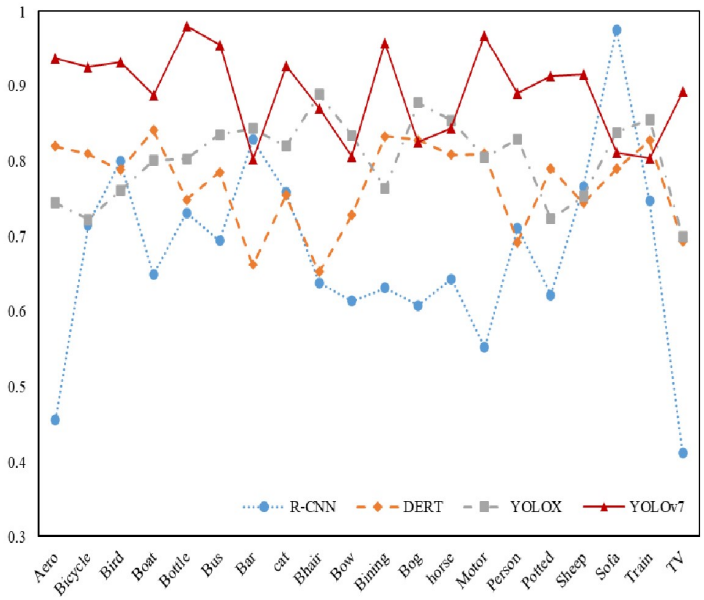

Figure 3. Recognition Rates of 20 Common Targets in the VOC Dataset

recognition and tracking, providing more reliable security and intelligent services for various application scenarios. To verify the advantages of the YOLOv7 model applied in video detection tasks, this paper also conducts comparative experiments with common R-CNN, DERT, and YOLOX object detection models, and the experimental results are shown in Figure 3. In the VOC dataset, after experimental comparisons, the detection average accuracy of the R-CNN model is 0.68, the DERT model is 0.77, and the YOLOX model is 0.8. In contrast, the YOLOv7 model achieved the highest detection average accuracy of 0.89. Firstly, the R-CNN model performs relatively low, possibly due to its adoption of candidate region selection and multi-step object detection process. This leads to higher computational complexity and longer inference time, affecting the model's accuracy. Next, the DERT model has a relatively high accuracy but still lags behind the YOLOX and YOLOv7 models. The DERT model may have some limitations in network structure and feature extraction, leading to its relatively limited performance in object detection tasks. Secondly, the YOLOX model performs well in the VOC dataset with an accuracy of 0.80, showing strong object detection capability. YOLOX adopts a series of optimization strategies, such as DETR and CIOU loss, to improve YOLOv7, resulting in improved detection accuracy. Finally, the YOLOv7 model achieved the highest detection average accuracy of 0.89 in the VOC dataset. YOLOv7 incorporates a series of integrated convolutional operation modules, such as the ELAN module, MP module, and SPPCSPC module, optimizing the model structure and feature extraction capability and demonstrating outstanding performance in object detection tasks. In conclusion, the experimental results show that the YOLOv7 model achieved the best detection average accuracy in the VOC dataset, demonstrating high accuracy and robustness. This makes it suitable for efficient and accurate target recognition and tracking in intelligent video surveillance systems. These results provide valuable references for further optimising and applying smart video surveillance systems.

6. Conclusions

In this paper, we focus on the application research of intelligent video surveillance systems based on deep learning to improve the efficiency and accuracy of object detection. For this purpose, we selected the YOLOv7 model as the main research tool and used the VOC dataset for evaluation. The demand for intelligent video surveillance systems continues to grow, making applying deep learning technology in this field increasingly important. The VOC dataset contains 20 target categories; each image has corresponding annotation information. Through experimental comparisons, we found that the YOLOv7 model achieved the highest detection average accuracy in the VOC dataset, reaching 0.89. This demonstrates the excellent performance of the YOLOv7 model in object detection tasks, indicating its broad application prospects in tasks, indicating its broad application prospects in intelligent video surveillance systems.

In summary, this paper proposes the application research of intelligent video surveillance systems based on deep learning, focusing on the YOLOv7 model. The experimental results show that the YOLOv7 model performs well in the VOC dataset, with efficient, accurate, and robust characteristics. This makes it suitable for efficient detection and tracking of targets in intelligent video surveillance systems. This provides strong support for the optimization and application of intelligent video surveillance systems.

References

[1] Sneha, K. A. (2022). Hyperspectral imaging and target detection algorithms: A review. Multimedia Tools and Applications, 81 (30) 44141–44206.

[2] Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016). You only look once: Unified, real-time object detection. *In:* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (p. 779–788). Las Vegas, NV, USA.

[3] Jocher, G. (2020). YOLOv5 by Ultralytics. Retrieved from https://github.com/ultralytics/yolov5

[4] Redmon, J., Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint, arXiv:1804.02767.

[5] Jocher, G., Chaurasia, A., Qiu, J. (2023). Ultralytics YOLO. Retrieved from https://github.com/ultralytics/ultralytics

[6] Wang, C. Y., Bochkovskiy, A., Liao, H. Y. M. (2023). 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 (p. 7464–7475). Vancouver, BC, Canada.

[7] Wang, C. Y., Yeh, I. H., Liao, H. Y. M. (2024). YOLOv9: Learning what you want to learn using programmable gradient information. arXiv preprint, arXiv:2402.13616.

[8] Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., and Ding, G. (2024). YOLOv10: Real-time end-to-end object detection. arXiv preprint, arXiv:2405.14458.

[9] Bochkovskiy, A., Wang, C. Y., and Liao, H. Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint, arXiv:2004.10934.

[10] Xu, S., Zhu, J., Jiang, J., et al. (2020). Sea-surface floating small target detection by multi-feature detector based on isolation forest. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 704–715.

[11] Hou, Q., Wang, Z., Tan, F., et al. (2021). RISTDnet: Robust infrared small target detection network. IEEE Geoscience and Remote Sensing Letters, 19, 1–5.

[12] Zhang, S., et al. (2012). On design and implementing a high definition multi-view intelligent video surveillance system. In Proceedings of the IEEE International Conference on Signal Processing, Communication and Computing (pp. 353–357).

[13] Paglinawan, C. C., et al. (2018). Optimization of vehicle speed calculation on Raspberry Pi using sparse random projection. In Proceedings of the IEEE International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management.

[14] Xu, J. (2021). A deep learning approach to building an intelligent video surveillance system. Multimedia Tools and Applications, 80, 5495–5515.

[15] Lin, T. Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dolla, P. (2014). Microsoft COCO: Common objects in context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T. (Eds.), Computer vision – ECCV 2014 (pp. 740–755). Springer International Publishing.

[16] Schroff, F., Dmitry, K., and Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[17] Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10), 1499–1503.

[18] Girshick, R. (2015). Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[19] Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[20] Gool, L. V., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The Pascal visual object classes (VOC) challenge. International Journal of Computer Vision, 88, 303–338.

[21] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9), 1904–1916.

[22] Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (Eds.), Advances in neural information processing systems 28 (pp. 91–99). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf

[23] Kumar, S., and Das, S. K. (2019). Target detection and localization methods using compartmental model for Internet of Things. IEEE Transactions on Mobile Computing, 19(9), 2234–2249.

[24] Gillis, D. B. (2020). An underwater target detection framework for hyperspectral imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 1798–1810.

[25] Ma, J., Tang, L., Xu, M., et al. (2021). STDFusionNet: An infrared and visible image fusion network based on salient target detection. IEEE Transactions on Instrumentation and Measurement, 70, 1–13.

[26] Zhao, B., Wang, C., Fu, Q., et al. (2020). A novel pattern for infrared small target detection with generative adversarial network. IEEE Transactions on Geoscience and Remote Sensing, 59(5), 4481–4492.

[27] Chang, C. I. (2021). Hyperspectral anomaly detection: A dual theory of hyperspectral target detection. IEEE Transactions on Geoscience and Remote Sensing, 60, 1–20.

[28] Huang, S., Cornelis, B., Devolder, B., et al. (2020). Multimodal target detection by sparse coding: Application to paint loss detection in paintings. IEEE Transactions on Image Processing, 29, 7681–7696.

[29] Yu, C., Liu, Y., Wu, S., et al. (2022). Pay attention to local contrast learning networks for infrared small target detection. IEEE Geoscience and Remote Sensing Letters, 19, 1–5.