Advances in Engineering Innovation (2024) Volume 11 EWA Publishing
Published online: 26 September 2024 DOI: 10.54254/2977-3903/11/2024116

Tracking abandoned objects in surveillance video based on human trajectory

Xiyuan Yang

Xiamen University Tan Kah Kee College

2608550708@qq.com

Abstract. Intelligent video surveillance is a significant research area in computer vision, with the detection and tracking of abandoned objects in public spaces attracting considerable public attention. However, existing detection technologies still face challenges, including false alarms, missed detections, poor adaptability to environmental conditions, and an inability to monitor in real-time. This paper proposes a method for tracking abandoned objects in surveillance video by analyzing human trajectories. The method addresses three main challenges: detection, tracking, and identification of abandoned objects. First, the paper discusses the inter-frame comparison method and the YOLOv8-based object detection algorithm for identifying abandoned objects and determining their categories. Additionally, it explores person re-identification technology and its application in identifying the owner of an abandoned object. Finally, the movement trajectory of the object's owner is analyzed to determine the intent of the person who placed it.

Keywords: abandoned objects, object detection, trajectory tracking, person re-identification

1. Introduction

With the rapid development of society, public safety has increasingly attracted attention, especially in complex environments such as airports, shopping malls, and schools. In these areas, dangerous objects may be disguised and deliberately left in crowded places, creating significant security risks. To ensure public safety, the automatic, rapid, and accurate detection and tracking of suspicious items left behind in surveillance video have become indispensable components of intelligent video surveillance systems.

However, most current surveillance systems still rely on manual verification. Given the vast amount of surveillance video data, this traditional method is not only inefficient but also incurs significant labor costs. Additionally, in practical applications, the detection of left objects is often hindered by factors such as lighting conditions and the occlusion of objects in the video. As technology advances, surveillance systems have been widely deployed in most public places. Thus, developing an efficient, stable, and adaptive abandoned object detection algorithm using surveillance video data holds great practical significance and application value. Such an algorithm can effectively reduce labor costs while enabling efficient and accurate detection of suspicious abandoned objects.

This paper discusses the issue of tracking abandoned objects. First, abandoned object detection will be achieved using the inter-frame comparison method and the YOLOv8-based object detection algorithm. Once the owner of the left object is identified, Person Re-identification technology will be employed to detect and analyze their trajectory. This analysis will help determine potential security risks and enable efficient and accurate detection and assessment of abandoned objects in surveillance video, providing robust support for enhancing public safety and responding to emergencies.

Regarding the feasibility of this project, the rapid development of computer vision technology offers strong technical support for the detection of suspicious abandoned objects. The powerful computing capabilities of modern computers and GPUs enable real-time processing of surveillance video. The growing availability of public datasets and labeling tools also provides valuable data support for model training. From a practical standpoint, surveillance systems have been widely deployed in many public places, providing the necessary hardware support for the detection and tracking of abandoned objects. Additionally, increasing public awareness of safety creates a broader market and more application scenarios for the detection of abandoned objects. Moreover, many countries and regions have established regulations and policies regarding public safety and surveillance, providing policy support for this type of research.

This paper is divided into four parts. The second part reviews classical literature, including previous research on object detection algorithms. The third part focuses on the inspection, tracking, and identification of abandoned objects. Finally, the fourth section concludes with a discussion of the limitations of this research and potential future directions.

2. Literature review

Multiple studies have been conducted on Re-ID and abandoned object detection algorithms.

In this article, Re-ID is employed to track the owners of abandoned objects. Person re-identification (Re-ID) is an intelligent video surveillance technology that identifies the same individual across different cameras [1]. Ye, M., Shen, J., et al. introduce the widely studied person Re-ID in a closed-world setting, focusing on three aspects: feature representation learning, deep metric learning, and ranking optimization. Additionally, they introduced a new evaluation metric to measure the cost of finding all correct matches [2]. Ming, Z., Zhu, M., et al. classify deep learning-based person Re-ID methods into four categories: depth metric learning, local feature learning, generative adversarial learning, and sequence feature learning. Furthermore, they subdivide these four categories according to their methodologies and motivations, discussing and comparing the advantages and limitations of each subcategory [3]. Person re-identification may be affected by changes in the camera's perspective, and the differences between individuals in public places are often subtle, which can reduce recognition accuracy. To address this, a person re-identification method based on feature fusion, incorporating a full-size feature deep Convolutional Neural Network (OSNet) [4], has been proposed to improve accuracy.

For abandoned object detection, a model using a multi-step method for dynamic detection is primarily trained with Fully Convolutional Networks (FCN), the RetinaNet algorithm, and the YOLO network [5]. This approach effectively addresses the complexity of abandoned objects and background changes, maintaining high reliability and stability in various environments. Another method, based on YOLOv5, accurately identifies small targets of abandoned objects more quickly [6]. This model is also well-suited for detecting targets of various small objects.

3. Main parts

The abandoned object tracking method of surveillance video based on person trajectory mainly includes three parts: abandoned object detection, abandoned object tracking and suspicious abandoned object identification.

3.1. Abandoned object detection

In the detection phase, the inter-frame comparison method is used to identify abandoned objects, followed by the application of the YOLOv8-based object detection algorithm to classify the object types in the surveillance video, such as a person, bag, etc.

3.1.1. Algorithmic model

1. Inter-frame Alignment Method

Inter-frame alignment is an object detection method in computer vision. The main concept is to detect and distinguish the foreground from the background by comparing the differences between each video frame and the background frame, thereby extracting the moving object.

Suppose the background image of the video is A, and the NTH frame image in the video sequence is A_n. For the corresponding pixels in the two frames, their gray values are $A(x,y)$ and $A_n(x,y)$, respectively. The gray values of the corresponding pixels in the two frames are subtracted, and their absolute values are taken to obtain the difference image D_n, as shown in Formula 1:

$$D_n(x,y) = |A_n(x,y) - A(x,y)| \tag{1}$$

The advantages of the inter-frame alignment method include its simplicity and intuitiveness, relatively low computational complexity, and the lack of need for complex mathematical models or large amounts of training data, making it easy to implement. However, the inter-frame comparison method is sensitive to changes in illumination. In environments with dynamic elements, such as swaying leaves or rippling water, the inter-frame comparison method may struggle to accurately distinguish between the foreground and background, potentially misclassifying dynamic background elements as foreground. This could result in the failure to accurately detect or identify moving targets in the foreground in some complex situations.

2. YOLOv8 Model

YOLOv8 is the latest major update to YOLOv5, open-sourced by Ultralytics on January 10, 2023. It is the most advanced YOLO model for image classification, object detection, and instance segmentation tasks.

The network architecture of the YOLOv8 model primarily includes three parts: the backbone network, neck network, and detection head [7]. In the backbone network, the image is down-sampled through convolution operations to extract features, with each convolution layer containing batch normalization and the SiLU activation function. In the neck network, the PA-FPN up-sampling stage convolution from YOLOv5 is removed, and the C3 module is replaced by the C2f module to enhance features and

transmit information to the head network. Finally, YOLOv8 uses decoupled detection heads to calculate the loss of regression and classification through two parallel convolution branches [8].

YOLOv8 continues the real-time detection advantages of the YOLO series and can maintain high frame rate (FPS) operation even with limited hardware configurations, making it especially suitable for scenarios with strict real-time requirements. It uses optimized multi-scale prediction technology and adaptive anchor adjustment, effectively improving the detection capabilities of objects of various sizes and accurately predicting the location and size of objects. YOLOv8 is highly flexible and scalable, compatible with various hardware platforms, including CPUs and GPUs, which enhances its practicality in real-world applications. However, YOLOv8 demands significant computing resources, and the complexity of its network architecture, along with the cumbersome training process, may pose challenges for model tuning and deployment.

Overall, YOLOv8 has become the preferred solution for object detection, image segmentation, and classification tasks due to its excellent speed, accuracy, and ease of use. It holds broad application prospects in the field of computer vision.

3.1.2. Main processes

1. Determine an Abandoned Object

The main idea is to take the first frame of the surveillance video and save it as 'background.jpg', representing the scene without any left objects, i.e., the background. The difference between each current frame and the background frame in the surveillance video is then calculated through pixel-wise subtraction to obtain the difference image, which represents the foreground. Suspicious abandoned objects are those that have been brought into the scene and left behind by individuals. Therefore, when the first frame is used as the background, the difference image can identify objects that were not originally in the scene, i.e., the foreground elements. The suspicious abandoned object can then be determined by analyzing the motion state of each element in the foreground across consecutive frames. Figure 1 illustrates the process of the inter-frame alignment method.

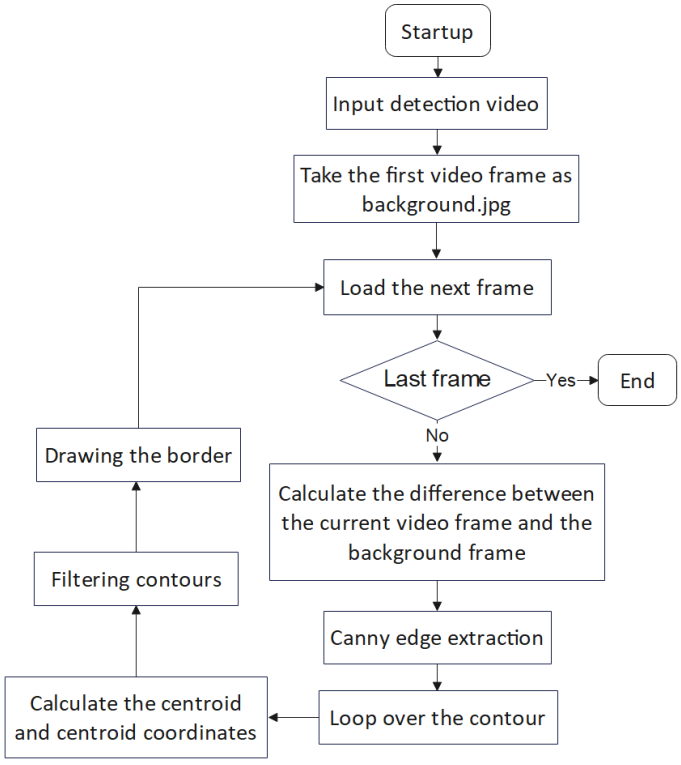

Figure 1. Flowchart of inter-frame alignment method.

Next, we define the contour of a foreground element with the following criteria: 1) its area is between 200 and 20,000 pixels; 2) the contour appears in the same position across multiple consecutive frames of the surveillance video; and 3) the count exceeds 100. If these conditions are met, it is determined that the surveillance video may contain left items. The system then records the time point, duration of stay, and location information of the remnants and alerts the user that there may be suspicious items in the surveillance video, as shown in Figure 2.

Figure 2. Determining an abandoned object.

2. Left-Behind Category Detection

Using the YOLOv8 model, a training dataset (including the Caltech Pedestrians dataset and the Bag dataset) is prepared by converting it to a format that matches the YOLOv8 model's input requirements. The images and annotation files are divided into training, validation, and test sets at a ratio of 8:1:1 and imported into the YOLOv8 model for training.

Video frames are extracted and preprocessed to meet the input requirements of the YOLOv8 model. These processed video frames are then input into the trained YOLOv8 model. The Darknet53 network within the YOLOv8 model extracts features from the video frames and generates a series of candidate boxes on the feature map. Confidence scores are calculated to predict the class of each candidate box, and finally, the bounding box of the detected object is drawn on the original video frame, with its class and confidence labeled. Figure 3 illustrates the YOLOv8 object detection process.

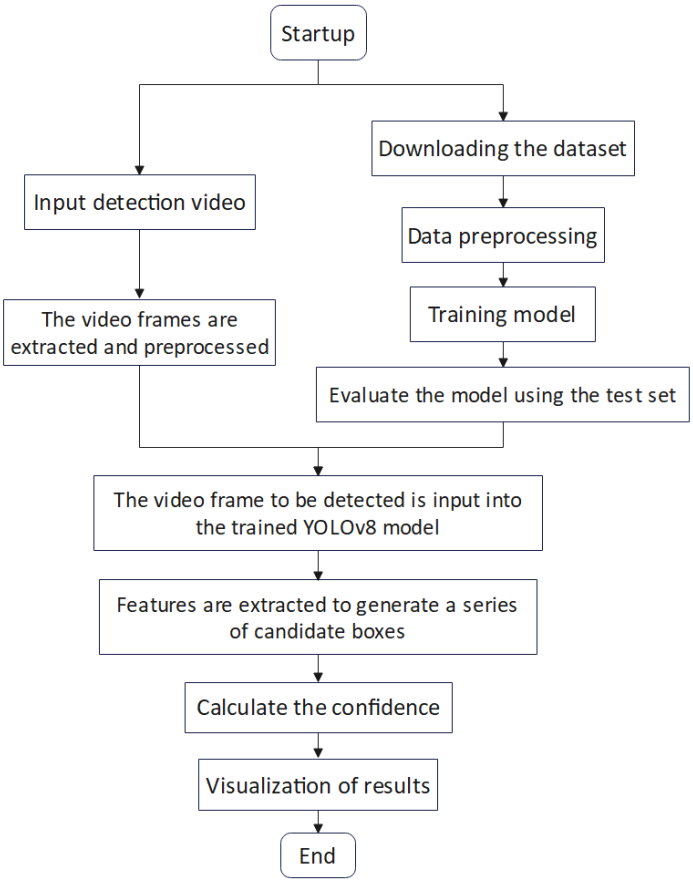

Figure 3. Flowchart of YOLOv8 object detection.

As shown in Figure 4, when the category of the moving object in the foreground is detected as "person" in the left image, the label "person" and its confidence score are displayed on the video frame. The right image shows a case where the object category is "bag".

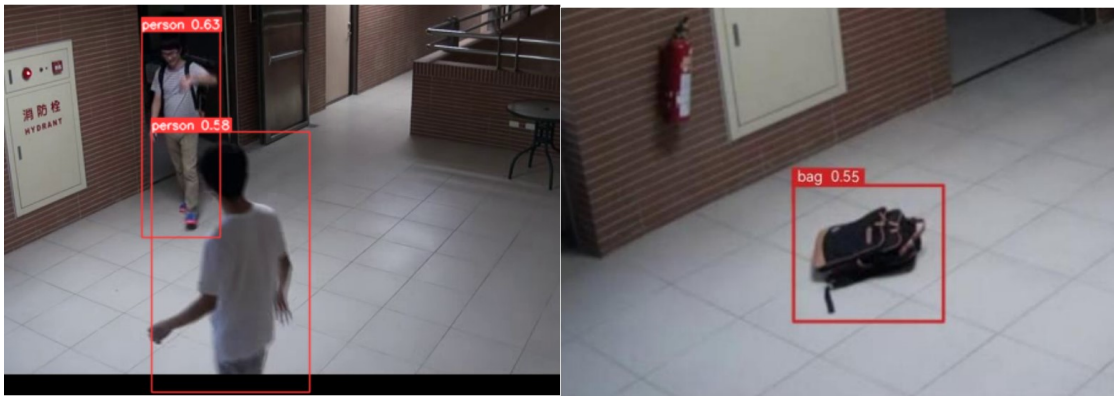

Figure 4. Detecting categories of abandoned objects.

3.2. Left-behind object tracking

The tracking process for an abandoned object involves two main steps: identifying the person who placed the abandoned object and detecting their trajectory. First, based on the location, time, and other relevant information of the left object, the most likely individual responsible is identified by calculating the overlap area and linear distance between the pedestrian and the abandoned object. Then, person re-identification technology is employed to detect and analyze the trajectory of the owner.

3.2.1. Identifying the person who left the item

Given the known time, duration, and location information of the abandoned object, the time period during which the object entered the monitored area can be backtracked in the surveillance video from the moment the object was detected. The overlap area between each pedestrian in the video and the bounding box of the abandoned object, as well as the linear distance between their positions, are calculated. The individual with the largest overlap area and the shortest linear distance is most likely the owner of the abandoned object.

As shown in Figure 5, the origin of the coordinate system is located in the upper left corner of the image, with the X-axis running horizontally to the right and the Y-axis running vertically downward. Let the upper left corner pixel coordinates of the abandoned object's bounding box be (x_1, y_1) and the lower right corner pixel coordinates be (x_2, y_2). For the bounding box of the ith pedestrian appearing in the video, let the upper left corner pixel coordinates be (x_{i1}, y_{i1}) and the lower right corner pixel coordinates be (x_{i2}, y_{i2}).

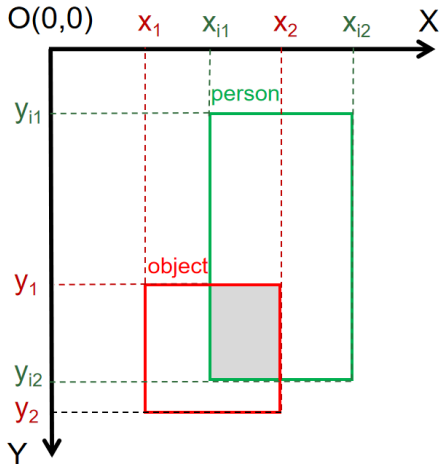

Figure 5. Coordinate system.

Then, the coordinates of the upper left corner of the intersection of the two bounding boxes (the gray area in Figure 5) are $x_{left} = max(x_1, x_{i1})$, $y_{left} = max(y_1, y_{i1})$, and the coordinates of the bottom-right pixel are $x_{right} = min(x_2, x_{i2})$, $y_{right} = min(y_2, y_{i2})$. Therefore, the overlap area of the left object's bounding box and the pedestrian's bounding box can be expressed by Formula 2:

$$A = \max\left(x_{right} - x_{left}, 0\right) \times \max\left(y_{right} - y_{left}, 0\right) \tag{2}$$

The center coordinate C_{obj} of the left object's bounding box is $\left(\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2}\right)$, and the center coordinate C_i of the pedestrian's bounding box is $\left(\frac{x_{i1}+x_{i2}}{2}, \frac{y_{i1}+y_{i2}}{2}\right)$. The straight-line distance c between the left object and the pedestrian can then be calculated using the Euclidean distance formula, as shown in Formula 3:

$$d_i = \sqrt{\left(\frac{x_1 + x_2}{2} - \frac{x_{i1} + x_{i2}}{2}\right)^2 + \left(\frac{y_1 + y_2}{2} - \frac{y_{i1} + y_{i2}}{2}\right)^2} \tag{3}$$

As illustrated in Figure 6, the red boundary box represents the left object, while the green box on the right has the largest overlap area and the shortest distance to the red box. Therefore, the pedestrian within the green box is most likely to have placed the left object.

Figure 6. Identifying the person who left the item.

3.2.2. Trajectory detection of the owner of the abandoned object

When the person who placed the object moves from one camera to another, person re-identification technology is used to identify pedestrians across cameras by comparing characteristics such as clothing and body posture. This helps confirm whether a pedestrian captured by different cameras is the same person, thereby enabling the detection of the trajectory of the owner of the abandoned object.

Person Re-Identification (ReID) is a technology that uses computer vision to determine whether a specific pedestrian appears in an image or video sequence, i.e., to identify the target pedestrian across possible sources and non-overlapping camera views. ReID is often considered a sub-problem of image retrieval, addressing the visual limitations of fixed cameras, and can be combined with pedestrian detection and tracking technologies. It has broad applications in intelligent video surveillance and security. Due to factors such as camera resolution and shooting angle, high-quality facial images are often not available in surveillance videos. An important feature of ReID is its ability to retrieve the target pedestrian across cameras when face recognition fails.

Although ReID research has been ongoing for many years, significant breakthroughs have only occurred in recent years with the advent of deep learning. From 2005 to 2013, traditional algorithms dominated the field. Since 2014, deep learning approaches have emerged, achieving significant advancements in pedestrian re-identification research, with performance indicators far exceeding those of traditional algorithms. However, recent developments indicate that while the latest algorithms have reached or even surpassed human-level performance, their performance has plateaued. Current research focuses mainly on complex environments, open settings, and unsupervised learning.

Despite these advances, person re-identification still faces many challenges. Factors such as human pose, illumination, resolution changes, and shooting conditions can impact performance. For example, different pedestrians may have similar body shapes and clothing. Additionally, the actual scenes in surveillance videos are often complex, with many surrounding objects and frequent occlusion. Variations in camera scenes and parameters, such as illumination and perspective changes, can lead to significant differences in the appearance of the same pedestrian across different cameras.

In summary, ReID technology can effectively detect and analyze the trajectory of the person who placed the object. First, using the YOLOv8 object detection algorithm, the pedestrian in each video frame is detected, and the bounding box is saved. Second,

the pedestrian re-identification algorithm searches through the video stream captured by all cameras based on the provided target pedestrian image, ultimately generating the motion trajectory of the target pedestrian, as shown in Figure 7.

Figure 7. The main processes of ReID.

Among them, the person re-identification algorithm mainly includes the following steps:

1. Extract Features

The pedestrian information in the candidate owner's image is converted into high-dimensional feature vectors. These vectors represent the pedestrian's appearance features (such as clothing, body posture, hairstyle, etc.) and shape features (such as size, height, etc.).

2. Feature Representation and Metric Learning

Through metric learning methods (such as contrastive loss, triplet loss, etc.), the extracted feature vectors are mapped to a new space where the vectors of the same pedestrian are close to each other, and the vectors of different pedestrians are far apart.

3. Feature Matching and Retrieval

The most similar feature vector to the query image is searched for in the gallery set. The similarity between the feature vectors (such as cosine similarity, Euclidean distance, etc.) is calculated to find the pedestrian image that best matches the query image. The retrieval results are then returned, as shown in Figure 8. This process allows for the generation of the action trajectory of the owner.

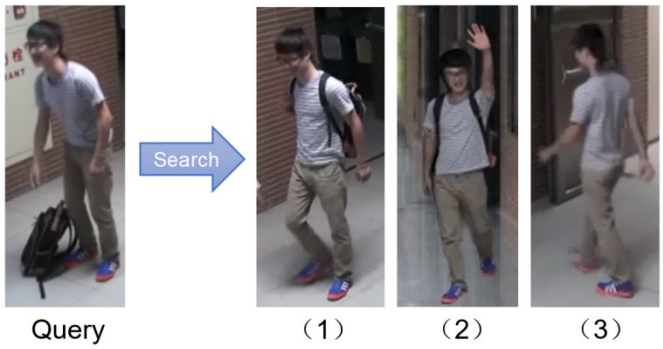

Figure 8. Pedestrian Matching.

3.3. Identification of suspicious abandoned objects

Based on the detected action trajectory of the owner, their intention is further analyzed. If the person does not disappear from all the monitoring images, returns to the location of the abandoned object, and ultimately leaves the scene with the abandoned object, it may indicate that the item was temporarily placed. However, if the person leaves the item and ultimately disappears from all surveillance video images, it is suspected that the person deliberately left the suspicious item, as shown in Figure 9.

Camera1 Camera2

Camera3 Camera4

Figure 9. The Placer Disappears.

4. Conclusion

This paper focuses on the problem of abandoned object tracking in surveillance video based on human trajectories. The issue is addressed from three main aspects. First, abandoned objects in surveillance video are detected using the inter-frame comparison method and the YOLOv8-based object detection method. Second, the principle of person re-identification technology is described and applied to identify the owner of the abandoned object. To tackle the complex problem of abandoned object tracking, this paper proposes a detection scheme that combines an innovative inter-frame comparison method with a YOLOv8-based object detection algorithm. This approach offers significant improvements in the efficiency and accuracy of abandoned object detection, enhancing both detection precision and speed. After identifying the owner of the abandoned object, person re-identification technology is used to detect and analyze the owner's trajectory, providing a robust basis for assessing potential security risks. The proposed method contributes to improving public safety and emergency response capabilities, demonstrating substantial practical significance and value.

However, this study has several limitations that suggest directions for future research. The accuracy of this approach primarily depends on the quality and quantity of surveillance video. Low-resolution, blurred, or incomplete video data can adversely affect abandoned object detection and tracking performance. Additionally, the detection algorithms and techniques discussed require significant computational resources when processing large volumes of monitoring data, which may limit real-time application in systems with constrained computing power. Moreover, the dataset used does not encompass all possible object types and scenarios, potentially restricting the model's generalizability.

Despite these limitations, important lessons can be drawn for practice and further research. First, incorporating advanced deep learning architectures or expanding the diversity of training data could enhance performance across various environments and object types. Second, integrating additional sensor data, such as audio or infrared, could provide more comprehensive information. Third, addressing public safety monitoring challenges can benefit from interdisciplinary collaboration with law enforcement, traffic management, and other relevant fields.

References

[1] Wei, W. , Yang, W. , Zuo, E. , Qian, Y. , & Wang, L. . (2022). Person re-identification based on deep learning - an overview. *Journal of visual communication & image representation*(Jan.), 82.

[2] Ye, M. , Shen, J. , Lin, G. , Xiang, T. , & Hoi, S. C. H. . (2021). Deep learning for person re-identification: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99), 1-1.

[3] Ming, Z. , Zhu, M. , Wang, X. , Zhu, J. , Cheng, J. , & Gao, C. , et al. (2022). Deep learning-based person re-identification methods: a survey and outlook of recent works. *Image and vision computing*(Mar.), 119.

[4] Zhou, K. , Yang, Y. , Cavallaro, A. , & Xiang, T. . (2019). Omni-scale feature learning for person re-identification.

[5] Hou Rui, & Tang Zhenmin. (2023). A multi-step method for road remnants detection based on deep learning. *Computer & Digital Engineering*, 51(8), 1756-1760.

[6] Anonymous. (2023). A Precise Identification Method for Small Targets of Legacy Based on Yolov5. CN117315233A.

[7] Yu Fei, Xu Bin, Wang Ronghao, Han Hequan. (2023). Rotating chain plate detection algorithm based on improved yolov8. *Manufacturing automation*, 45(9), 212-216.

[8] Su Jia, Jia Ze, Qin Yichang & Zhang Jianyan. (2024). Improved YOLOv8 algorithm for industrial surface defect detection. *Computer Engineering and Applications* (14), 187-196.