IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 January 2024, accepted 21 January 2024, date of publication 25 January 2024, date of current version 6 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3358450

RESEARCH ARTICLE

Indirect Vaccine Box Localization in Small to Medium Obstructed Cold Storages via Worker Tracking With VCS-YOLOv5

CHEN LIANG[1,2,3,4], WEI YANG[1,2,3], LONGLONG PANG[1,2,3,4], ZHUOZHANG ZOU[1,2,3,4], AND QUANGAO LIU[1,2,3,4]

[1]Key Laboratory of Networked Control Systems, Chinese Academy of Sciences, Shenyang 110016, China
[2]Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China
[3]Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110016, China
[4]School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China

Corresponding author: Wei Yang (epicard@163.com)

This work was supported by the Natural Science Foundation of Gansu Province under Grant 21JR11RA067.

ABSTRACT Within the realm of public health, the end-to-end traceability and monitoring of vaccines play an indispensable role in ascertaining the safety and efficacy of vaccines, especially the precise localization of vaccines in the vaccine cold storage. However, challenges such as limited space, dense stacking of boxes, and frequent obstructions in the vaccine cold storage, particularly in Small to Medium Cold Storage (SMCS), pose significant obstacles to effective localizing. Existing vaccine box localizing methods in cold storage, like manual localizing, Radio Frequency Identification (RFID) technology, and traditional visual localizing, struggle with obstructions and inefficiencies, leading to limited accuracy and real-time update capabilities. This paper introduces an innovative solution for vaccine box localization in obstructed environment within SMCS, leveraging computer vision technology. Specifically, to address the challenge of accurately locating vaccine boxes in densely stacked and heavily obstructed SMCS, this paper exploits the strong correlation between the vaccine boxes and workers during the storage process. The vaccine box is indirectly located by focusing on the less numerous and less obstructed cold storage workers. Furthermore, to enhance the tracking accuracy of the workers, the YOLOv5 model was modified, resulting in the development of the Vaccine Cold Storages YOLOV5 (VCS-YOLOv5) model tailored for obstructed environment in SMCS. Additionally, the final location of the vaccine box is determined by a behavior recognition model, identifying instances where the workers' hands are not in contact with the vaccine box. Extensive experiments confirm that VCS-YOLOv5 sets a new benchmark in vaccine box localization and worker tracking, significantly surpassing the performance of standard models in accuracy and real-time effectiveness.

INDEX TERMS YOLOv5, object tracking, vaccine box location, small to medium obstructed cold storage, behavior recognition.

I. INTRODUCTION

Vaccines play an irreplaceable role in safeguarding public health because vaccination is the most cost-effective and efficient public health intervention tool for the prevention and control of infectious diseases [1]. An active vaccine is crucial for ensuring its immune functionality. Thus, to ensure the safety and efficacy of each vaccine, rigorous monitoring and management are required in the research and production stage as well as the storage and transportation stage. In the storage and transportation stage, the cold storage unit, which serves as a transitional repository for vaccines, plays a significant role in the vaccine cold chain distribution. In the cold storage, the precise localization of vaccine boxes is helpful for effective vaccine management. Consequently, enhancing the ability to precisely localize vaccines within cold storage facilities is

The associate editor coordinating the review of this manuscript and approving it for publication was Shovan Barma.

vital for improving vaccine quality, reducing wastage, optimizing supply chain efficiency, and ensuring public health safety.

Existing vaccine management primarily relies on attaching traceability codes to vaccine packaging. Subsequently, these codes are scanned throughout vaccine distribution and cold chain storage to acquire relevant data. Besides, due to limited investment and maintenance costs, most vaccine cold storages, especially in Small to Medium cold storages (SMCS), still depends on manual management using handheld PDAs (Personal Digital Assistants) for operations like stock entry, inventory checks, and dispatch. This management approach cannot precisely determine the location of vaccine boxes in the storage units, potentially leading to the oversight of vaccines nearing expiration. Hence, finding an accurate method to localize these vaccine boxes in vaccine cold storage is paramount.

Due to limited research specifically locating vaccine boxes in cold storage, this study explores the potential solutions to vaccine boxes localization in vaccine cold storages, such as IoT (Internet of Things) locating technology and automated equipment locating solutions commonly used in general logistics. These methods provide a baseline for addressing the unique challenges of locating vaccine boxes in cold storage environments. The IoT locating technique employs electronic tags attached to goods, which communicate with readers via radio waves to update their position information. Zeng et al. [2] proposed a smart decision-support system architecture based on RFID (Radio Frequency Identification) technology, which collects the location information of real-time goods through RFID tags. Liu et al. [3] designed a hazardous materials intelligent storage management system using UWB (Ultra-Wide Band). By deploying UWB nodes and installing UWB tags on goods, the system can determine the location of goods within the warehouse in real-time. However, due to dense stacking of vaccine boxes and limited space in SMCS, IoT technologies like RFID face challenges in SMCS, including signal transmission block and restricts on the placement and range of readers. Meanwhile, although UWB is more accurate, it struggles with higher costs and complex installation in these constrained spaces. For automated equipment locating solutions, automation devices like storage robots, AGVs (Automated Guided Vehicles), and automatic sorting systems are widely applied in the logistics field can also be introduced to the vaccine cold storage. Shi et al. [4] designed a smart robot storage logistics system adaptable to highly dynamic environmental changes, and Liu et al. [5] applied multiple AGVs for more efficient warehouse management. However, automated technologies such as storage robots and AGVs often struggle in the dense and confined spaces of SMCS. Besides, specialized adaptations for cold and humid conditions also improve their complexity and cost. It is worth mentioning that disease prevention and control centers predominantly use SMCS, so these technologies and equipment are not suitable for the widespread promotion in SMCS.

In addition to these technologies and equipment, computer vision technology has been widely applied to identify the object and determine the location of goods with its efficiency, accuracy, and cost-effectiveness in recent years. Yang et al. [6] used an improved RetinaNet algorithm for the precise locating of boxes within warehouses, aiding robot stacking and unstacking. However, this method encounters limitations in distinguishing boxes with similar external features, a common environment in SMCS environments where vaccine boxes often appear identical. Zou and Liu [7] proposed an image instance segmentation algorithm based on the improved Mask R-CNN to enhance object recognition and precise localization in complex environments. It can accurately calculate the target's spatial position but performs poorly when objects are severely obstructed. These computer vision methods struggle with challenges such as distinguishing similar-looking vaccine boxes and handling obstructed conditions. Consequently, these methods are not directly applicable for locating vaccine box in SMCS environments.

To address these challenges mentioned above, this paper proposes a comprehensive framework for accurately locating vaccine boxes in SMCS environments with obstructions. First, the traceability codes on the boxes are recognized by industrial cameras as workers transport vaccine boxes into cold storage. Then, to address the challenge of accurately locating vaccine boxes in densely stacked and heavily obstructed SMCS, this paper exploits the strong correlation between the vaccine boxes and workers during the storage process. The vaccine box is indirectly located by focusing on the less numerous and less obstructed cold storage workers. The YOLOv5 detector followed by the DeepSORT algorithm is employed for workers tracking. Furthermore, to enhance the tracking accuracy of the workers, the YOLOv5 model was modified, resulting in the development of the Vaccine Cold Storages YOLOV5 (VCS-YOLOv5) model tailored for obstructed condition in SMCS. Additionally, the final position of the vaccine box is determined by the ResNet behavior recognition model. Specifically, when a worker transitions from transporting to a non-transporting state, indicated by their hands no longer being in contact with the vaccine box, the worker's location at that moment is considered as the final position of the vaccine box. For clarity, 'localization' specifically refers to determining the exact position of a vaccine box within the cold storage at a precise moment, while, 'tracking' refers to continually determining the precise locations of workers in the cold storage in this paper.

The main contributions of this paper are summarized as follows:

1) This study proposes a comprehensive framework for vaccine box localization in Small to Medium Cold Storages (SMCS). This framework integrates traceability code recognition, worker tracking, worker behavior recognition, and vaccine box localization, addressing the need for real-time precise localization of vaccine boxes. Experiments demonstrate the superior performance of this framework, setting a new benchmark in this field.

2) This study advances the field by offering a cost-effective, computer vision-based vaccine boxes localization alternative to existing methods in SMCS. The approach not only reduces operational costs by minimizing the need for extensive hardware infrastructure but also fills a critical research gap in accurately locating vaccine boxes in SMCS.

3) Given the challenges faced by existing computer vision methods in distinguishing visually similar vaccine boxes, locating the same vaccine box in a video stream, and handling occlusion, this paper transforms the problem of vaccine box localization into a worker tracking issue by leveraging the strong correlation between vaccine boxes and worker. Consequently, it indirectly achieves accurate locating of vaccine boxes. Due to the relatively small number of workers, their distinct features and the difficulty of complete occlusion, this approach effectively overcomes the aforementioned challenges.

4) To further enhance the precision of the YOLOv5 algorithm in tracking workers, this paper has made improvements to the YOLOv5 model from three perspectives: generalization, feature extraction, and bounding box modeling under occlusion. This paper proposed the VCS-YOLOv5 model tailored for occlusion environment in SMCS, significantly enhancing the performance in terms of tracking workers.

The rest of this paper is organized as follows. Section II outlines several related works. The proposed method is presented in detail in Section III. Section IV is devoted to the presentation of the experimental results and the analysis performed on the dataset. Section V discusses the experimental results, contrasts them with previous studies, explores limitations, and suggests future research directions. Section VI summarizes this study and discusses future work.

II. RELATED WORK

A. IoT LOCALIZATION TECHNOLOGIES

In the realm of IoT localization, technologies such as RFID, UWB, Bluetooth, and ZigBee have established themselves as cornerstone solutions [8]. RFID, leveraging radio frequency waves, excels in inventory locating and asset management, known for its wide coverage and ease of deployment [9]. UWB technology, characterized by its use of extremely short radio waves, stands out for its high precision in indoor positioning, proving invaluable in complex environments [10]. Bluetooth, commonly used for short-range communication, has been adapted for localization with notable success in consumer applications, offering a balance between range and accuracy [11]. ZigBee, operating on low-power digital radio, is renowned for its efficiency and network flexibility, making it ideal for smart home and industrial applications [12]. However, the efficacy of these technologies diminishes significantly in SMCS with obstructed environments. RFID, while cost-effective, often struggles with signal interference and limited range in dense environments [13]. UWB, despite its accuracy, faces challenges in terms of high cost and power consumption, limiting its widespread adoption [14]. Bluetooth, commonly affected by signal attenuation and physical obstructions, can suffer in precision. ZigBee, although efficient, is constrained by its low data rate and susceptibility to interference in crowded radio frequency environments [15]. In contrast, this paper leverages advanced computer vision techniques to effectively overcome these limitations. By tracking workers for indirect vaccine box localization, the method bypasses traditional IoT issues like signal interference. This approach not only improves localization accuracy in obstructed cold storages but also minimizes hardware reliance, offering a more scalable and cost-effective solution.

B. VISION-BASED LOCALIZATION TECHNIQUES

Vision-based localization techniques, particularly those within the YOLO (You Only Look Once) series, have gained substantial attention in recent years. The YOLO architecture, renowned for its high-speed real-time object detection, leverages deep convolutional neural networks to classify and localize objects in a single forward pass [16]. This foundational work has spurred a series of developments and enhancements in the YOLO series [17], [18], [19], each aiming to improve aspects like detection accuracy, speed, and robustness in varied environments. Moreover, advancements in vision-based techniques against occlusions have furthered the scope of applications in complex environments [20]. A notable example is the use of part-based models, as discussed in the paper by Wang et al. [21], which are effective in localizing objects even when they are partially obscured. These models operate by recognizing and combining different parts of an object, a strategy that proves advantageous in maintaining localization continuity in the presence of occlusion. However, despite these advancements, such techniques still encounter significant challenges in environments with high-density obstructions, like SMCS. In these environments, the common issue of similar objects overlapping, combined with limited space, presents a substantial hurdle for conventional vision-based localization methods. These methods often struggle to distinguish between closely stacked items, which leads to inaccuracies in both object localization and identification. In response to these challenges, this paper takes a novel direction compared to traditional vision-based localization techniques. It leverages the correlation between the movement of cold storage workers and the position of vaccine boxes, introducing an indirect locating method. By focusing on the less obstructed and more distinguishable features of workers, this method, supported by the tailored VCS-YOLOv5 model, effectively overcomes the limitations posed by dense stacking and occlusions.

C. MULTI-OBJECT TRACKING TECHNIQUES

Multi-object tracking (MOT) techniques have seen substantial advancements, evolving to address increasingly complex tracking environments. The prevalent approach in this field

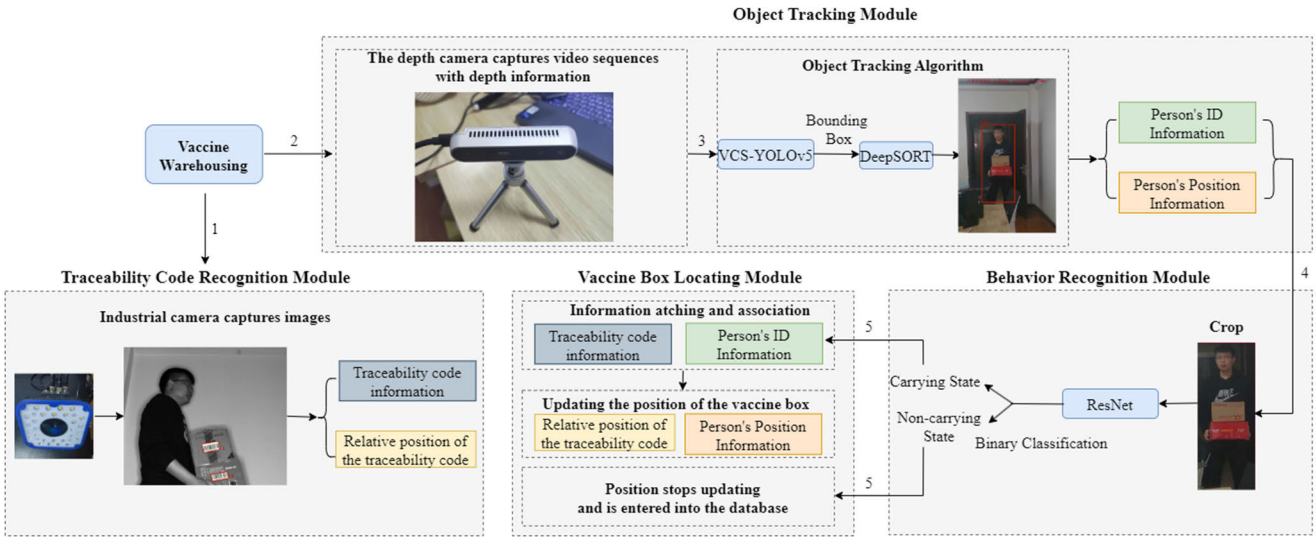


FIGURE 1. The overall framework of the vaccine box localization.

is the Tracking by Detecting methodology, which fundamentally consists of two components: Detection and Embedding. Based on the implementation of these components, algorithms can be categorized into two series: Separate Detection and Embedding (SDE) and Joint Detection and Embedding (JDE). In SDE algorithms, the detection and embedding processes are distinct, allowing for more focused optimizations in each step. This separation often leads to more efficient and accurate tracking, making SDE methods widely preferred in practical applications. DeepSORT [22] is a prime example of an SDE-based algorithm, known for its enhanced tracking accuracy through deep learning features. However, while SDE methods have shown significant progress, they heavily rely on the precision of the detector. In complex and confined environments like vaccine cold storages, where obstructions are prevalent, the accuracy of the detector can considerably diminish. This paper addresses this challenge by employing the VCS-YOLOv5 model, which has been specifically optimized for the unique conditions of vaccine cold storage environments. The enhanced detection capabilities of VCS-YOLOv5 ensure more accurate and reliable tracking in these challenging environments, thereby overcoming the limitations typically encountered by standard SDE methods.

III. METHODOLOGY

This study focuses on the task of accurately locating each unique vaccine box in RGB-D images. Given an RGB-D image, the objective is to identify and return the three-dimensional coordinates of every vaccine box, each distinguished by its own traceability code. This chapter introduces a framework encompassing four key parts: traceability code recognition module, object tracking module, behavior recognition module, and vaccine box localization module. These form the subsections of this chapter. A significant aspect of this work involves enhancing the YOLOv5 model, producing the specialized VCS-YOLOv5 for vaccine cold storage environments.

A. OVERALL FRAMEWORK DESCRIPTION

This study proposes a novel framework to localize vaccine boxes in small to medium cold storages (SMCS) with dense stacking and obstructions. Recognizing the challenges of direct locating in such environments, this study shifts focus to the correlation between workers and vaccine boxes. This method tracks fewer, less-obstructed cold storage workers, using their locations as proxies for the vaccine boxes. This indirect approach significantly improves the localization accuracy in obstructed environments. To ascertain the final location of the vaccine boxes, this study defines two distinct states for the workers: 'transporting' and 'non-transporting'. During the 'transporting' state, the vaccine box and the worker are regarded as a unified entity, allowing for real-time updates of the vaccine box's location. This state is identified by a behavior recognition model, which detects when a worker is handling the box. Conversely, the 'non-transporting' state is recognized when the worker's hands are no longer in contact with the vaccine box, indicating that the box has been placed at a location. In conclusion, this framework provides a practical and accurate solution for locating vaccine boxes in obstructed cold storage environments.

The overall framework proposed in this paper is illustrated in Figure 1. Firstly, an industrial camera automatically captures and recognizes the trace code on the vaccine box entering the cold storage, obtaining the information of the trace code and its relative position on the box surface. Next, a depth camera is used to continuously capture video sequences with depth information. The video sequence is then inputted into the VCS-YOLOv5 algorithm to obtain the detection box of the workers. The DeepSORT algorithm is then employed for real-time tracking of workers, retrieving the workers' ID information and three-dimensional position

data. Subsequently, the workers are cropped from the original video sequence, and the ResNet algorithm is employed to classify and recognize the workers' actions. If the current worker is identified as being in a transporting state, the traceability code information is matched and associated with the worker's ID information. The three-dimensional position data of the worker is used to represent the three-dimensional location of the vaccine batch being transported, realizing a preliminary location of the vaccine box in the vaccine cold storage. Using the relative position information of the trace code, the three-dimensional location data of each vaccine box is then restored, achieving precise location of the vaccine box in the cold storage. If it's recognized that the current worker transitions from a transporting state to a non-transporting state, the update of the vaccine box's location is stopped, and the information is recorded in the database.

B. TRACEABILITY CODE RECOGNITION MODULE

An industrial camera is placed behind the cold storage door. When workers transport vaccine boxes into the cold storage, the industrial camera automatically recognizes the traceability code on the vaccine box. It returns the information of the traceability code as well as its relative position on the box, which is used for subsequent locating of each vaccine box.

C. OBJECT TRACKING MODULE

1) VCS-YOLOV5 MODEL

Using YOLOv5 as an object detector in the vaccine cold storage, and directly locating the vaccine box based on the DeepSORT model, there are problems like difficulty in distinguishing vaccine boxes with similar appearances, inability to locating the same vaccine box in the video stream, and challenges in handling occlusion. By leveraging the high correlation between the vaccine box and the workers, this paper transforms the vaccine box locating problem into worker tracking problem, indirectly achieving vaccine box positioning. However, when YOLOv5 is used as an object detector to track worker in the vaccine cold storage, there are still some shortcomings:

First, YOLOv5 is an anchor-based algorithm, relying on manually designed hyperparameters, such as the aspect ratio, area, and number of anchor boxes. This mechanism doesn't adapt well to the complex and changing dataset of vaccine cold storage environments. Second, due to the intricate environment in the vaccine cold storage, workers will still be partially obscured, resulting in decreased algorithm detection accuracy. Lastly, due to the low color contrast between the workers and the surrounding environment, and significant noise interference, false negatives and false positives are likely.

To address the above issues and further improve the detection accuracy of the YOLOv5 algorithm in the vaccine cold storage, this paper proposes the VCS-YOLOv5 model tailored for small and medium-sized vaccine cold storage occlusion environments:

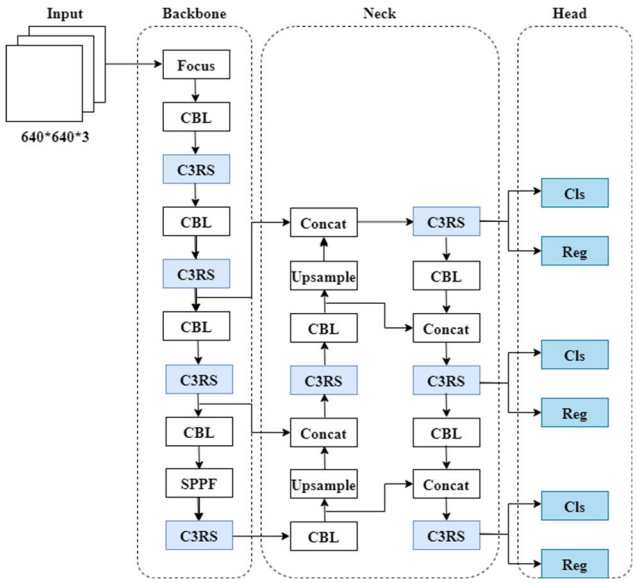


FIGURE 2. VSC-YOLOv5 network structure.

Anchor-free Detection Mechanism: To enhance the model's generalization capability on the vaccine cold storage environment dataset, the concept of the anchor-free model, FCOS [23], is borrowed. The VCS-YOLOv5 model adopts an anchor-free detection mechanism.

Optimized Network Structure: The overall network structure is depicted in Figure 2. Initially, this paper designs a C3RS module, a combination of C3, RepResBlock, and SE. The C3RS module introduces the RepResBlock structure to replace the Res-Unit structure of the C3 module in YOLOv5. By integrating multi-scale feature fusion, the model not only enhances its feature extraction capability but also maintains the inference speed of the original network structure. Furthermore, after the Concat connection, the C3RS module incorporates the channel attention SE module, further strengthening the model's feature representation capability. Subsequently, the paper substitutes the network's activation function with the Mish activation function, enhancing the model's non-linear representation capability. Finally, the coupled detection head (Coupled Head) is replaced by a decoupled detection head (Decoupled Head), allowing classification and regression tasks to learn independently, facilitating the enhancement of the model's feature representation capability.

Optimized Matching Strategy: This paper refers to and draws inspiration from the dynamic matching strategy of Task Align Assigner in the TOOD detector. This strategy strengthens the alignment and interaction of classification and regression tasks, enabling simultaneous acquisition of the highest classification scores and the most precise bounding boxes.

Optimized Loss Function: To improve the model's detection precision under occlusion environment, this paper introduces the Distribution Focal Loss (DFL) loss function, further

amplifying the model's capability to model bounding boxes in the intricate environment of the vaccine cold storage.

2) ANCHOR-FREE DETECTION MECHANISM

In the intricate environment of a vaccine cold storage, factors such as the camera installation position, the distance between the camera and the workers, and the posture of the workers can influence the proportion of workers present in the images. Conventional anchor-based object detection methods require manual design of anchor box aspect ratios, areas, and quantities to better meet the detection needs of various-sized workers members in the vaccine cold storage environment. However, in practical applications, given that the environment of the vaccine cold storage frequently changes, these manually designed parameter settings might not be accurate, leading to a decline in object detection performance and affecting the model's generalization capability.

To address the above issues, drawing inspiration from the anchor-free model FCOS, the VCS-YOLOv5 model proposed in this paper adopts an anchor-free detection mechanism, granting the model greater flexibility and adaptability to swiftly respond to changes in the vaccine cold storage environment.

Initially, every grid location inside the ground truth bounding box is considered as a positive sample in this paper. For each of these locations, the distances to the top, bottom, left, and right sides of the real bounding box are predicted, denoted as $[t, b, l, r]$. As illustrated in Figure 3, assuming a location inside the actual bounding box is (x, y), with the top-left corner of the actual box being (x_l, y_l) and the bottom-right corner being (x_r, y_r), the regression targets $[l_*, t_*, r_*, b_*]$ for this location can be represented as:

$$\begin{aligned} l_* &= x - x_l \\ t_* &= y - y_l \\ r_* &= x_r - x \\ b_* &= y_r - y \end{aligned} \tag{1}$$

Furthermore, this study employs the last three feature maps $[P_3, P_4, P_5]$ of the YOLOv5 network for prediction and sets four thresholds $[m_2, m_3, m_4, m_5]$. Subsequently, every positive sample location on the feature map P_i is traversed. For a positive sample (x, y) on feature map P_i, the distances $[t, b, l, r]$ from the current location to the real bounding box are first computed. The maximum value among them, denoted as m, is then determined, i.e., $m = \max(t, b, l, r)$. The range of threshold values into which m falls is ascertained to verify if m satisfies the conditions. If it meets the criteria, the location is designated as a positive sample; otherwise, it's labelled as a negative sample.

Through the above steps, objects of different scales are assigned to different feature layers for detection. Given that many overlaps occur between objects of disparate sizes, the VCS-YOLOv5 model not only possesses the capability for multi-scale prediction but also addresses the issue of object overlap.

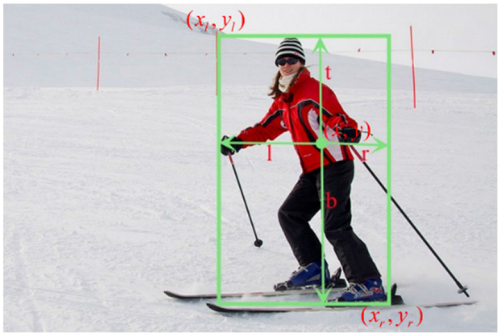


FIGURE 3. Model prediction principle.

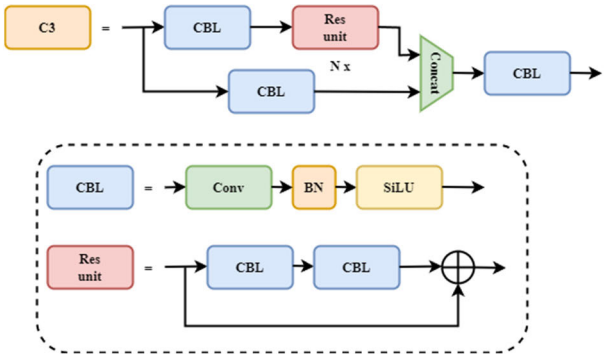


FIGURE 4. Basic principle of the C3 module.

3) OPTIMIZED NETWORK STRUCTURE

a: C3RS MODULE

In the environment of vaccine cold storage, many vaccine boxes are stacked together densely. Workers inevitably face obstructions from the boxes they are carrying or from their surroundings during the transportation process, increasing the difficulty of model recognition. Moreover, the cold storage itself is a relatively enclosed space. The lighting conditions are usually weak, and the color contrast between the target to be tested and its surroundings is not obvious. Coupled with the presence of noise interference, it is easy for the model to miss or misidentify objects. To further improve the detection accuracy of the YOLOv5 model in the vaccine cold storage environment, this paper has improved the C3 module in the YOLOv5 model and designed a C3RS module that combines C3, RepResBlock, and SE modules. With the introduction of only a small number of parameters, the model possesses stronger feature extraction capabilities, allowing it to effectively deal with the strong obstructions and weak contrast in the vaccine cold storage environment.

The C3 module is an essential component of the YOLOv5 network. Its main function is to increase the depth and receptive field of the network, enhancing the capability of feature extraction. The structure is divided into two branches. One branch passes through a standard convolutional module and multiple residual modules, while the other branch only passes through a standard convolutional module. Finally, a Concat operation is performed on the two branches, as shown in Figure 4.

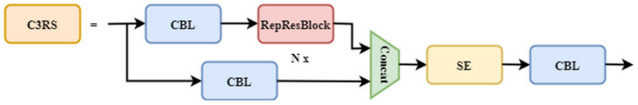


FIGURE 5. The overall structure of the C3RS module.

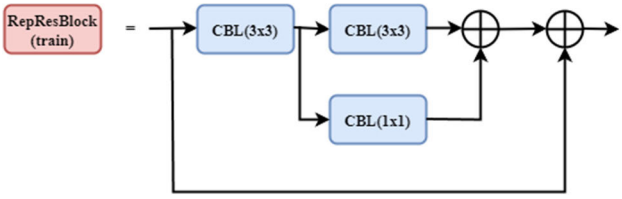


FIGURE 6. Structure of RepResBlock during training.

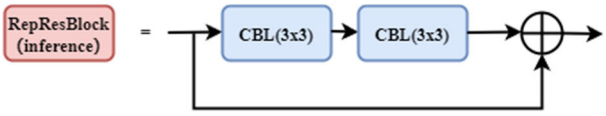


FIGURE 7. Structure of RepResBlock during inference.

The overall structure of the C3RS module is shown in Figure 5. Firstly, this paper introduces the RepResBlock [24] structure to replace the Res-Unit structure in the C3 module. The structures of RepResBlock during training and inference are shown in Figures 6 and 7, respectively. During model training, the RepResBlock structure introduces a standard convolutional branch based on the Res-Unit structure, which expands the receptive field of the model. After computation, the two branches are merged to complete the fusion of multi-scale features. With these improvements, the model can learn multi-scale feature information, further enhancing its feature extraction capability. However, introducing a new branch inevitably introduces additional parameters, leading to an increase in the model's computational cost and a decrease in inference speed. To avoid the adverse effects brought by the new branch, during the inference phase of the model, RepResBlock undergoes structural reparameterization. Specifically, the standard convolutional branch is padded as a convolution, and it is weight-fused with the convolution on the other branch, further optimizing the network structure, as shown in Figure 6. Through these improvements, the model can enhance its representational capability during training, and it will not introduce extra computational costs during deployment.

Furthermore, to emphasize the inter-channel relationships in the fused feature maps and enhance the contribution of crucial information to feature representation, this study integrates the SE [25] attention mechanism module after the C3 module. This mechanism explicitly models the interdependencies between feature channels. Through a learning approach, it automatically determines the importance of each feature channel. Consequently, it amplifies features with high contributions while suppressing less significant ones. The SE attention mechanism module is depicted in Figure 8.

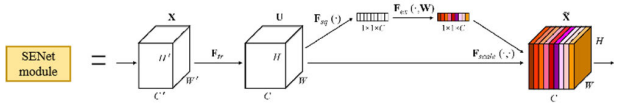


FIGURE 8. Basic principle of the C3 module.

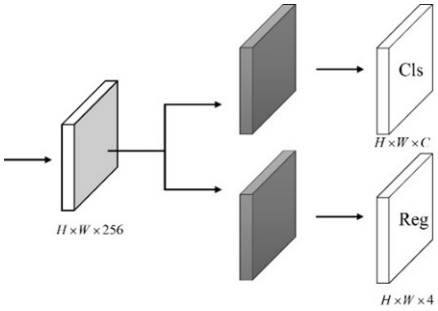


FIGURE 9. Structure of the decouple head.

b: DECOUPLED HEAD STRUCTURE

In object detection tasks, classification and regression are two relatively independent tasks. During the training process of the model, each focuses on different content and exhibits distinct preferences. Specifically, classification tasks emphasize the differences between individual objects, focusing on the texture features of the target, while regression tasks pay more attention to the contour and boundary features of the object. The original YOLOv5 algorithm employs a coupled prediction head that simultaneously handles classification and regression tasks. This not only reduces the detection accuracy of the model but also affects its convergence speed.

Inspired by the decoupled prediction head proposed by YOLOX [18], this study replaces the coupled prediction head of YOLOv5 with a decoupled one, allowing the classification (Cls) and regression (Reg) tasks to learn independently. This lets different sub-tasks focus on distinct features, as shown in Figure 9. Compared to the original YOLOv5, the improved decoupled prediction head enhances detection accuracy while accelerating the model's convergence speed.

c: MISH ACTIVATION FUNCTION

To further enhance the non-linear expressive capability of the YOLOv5 model, this paper adopts the Mish activation function [26] to replace the SiLu activation function within the network. Introduced by MISRA in 2019, the Mish activation function is novel, as shown in equation (2).

$$\text{Mish} = x \times \tanh(\ln(1 + e^x)) \tag{2}$$

While the Mish function establishes a lower boundary, it does not possess an upper one. As the values edge towards the extremes in either direction, its gradient approximates 1. This characteristic efficiently prevents the issue of slow convergence during network training caused by zero gradients. When benchmarked against activation functions like SiLu and Relu, Mish showcases a more refined and smoother transition, introducing heightened non-linear representation.

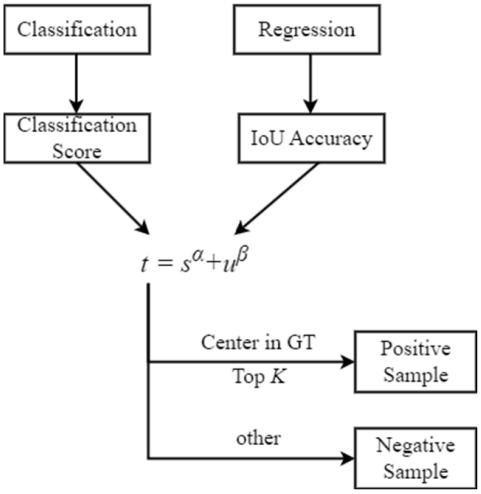


FIGURE 10. TAL dynamic matching strategy.

This leads to a commendable enhancement in the model's generalization and precision metrics.

4) OPTIMIZED MATCHING STRATEGY

The VCS-YOLOv5 model employs a geometry-based allocation strategy, considering every grid location (anchor point) within the real annotation box as a positive sample. It then allocates these to different feature layers for classification and regression based on the size scale of the detected object. However, the best anchor points for classification and localization are typically inconsistent. They can vary significantly based on the shape and features of the detected object. Since the geometry-based sample allocation strategy is task-independent, the chosen anchor points may struggle to make accurate and consistent predictions for both classification and regression tasks simultaneously.

Addressing the issues of the anchor-less YOLOv5 detector, this paper references and draws from the Task Align Learning (TAL) dynamic matching strategy found in the TOOD [27] detector. This strategy enhances the alignment and interaction of the detector's classification and regression tasks, obtaining boundary boxes with the highest classification scores and the most precise localization. Firstly, the decoupled prediction heads in the VCS-YOLOv5 model compute the classification score s and the location precision score u of the current prediction box separately for classification and regression tasks. Subsequently, the TAL dynamic matching strategy employs a high-order combination of the classification score and the location precision score to represent the comprehensive score t of the prediction box, expressed as $t = s^{\alpha} + u^{\beta}$. Here, α and β can control the influence degree of the classification score and location precision score on the comprehensive score, respectively. Finally, the top K prediction boxes with high comprehensive scores and central points within the real annotation boxes are selected as positive samples, with the remaining detection boxes being negative samples. The principle of the TAL dynamic matching strategy is illustrated in Figure 10.

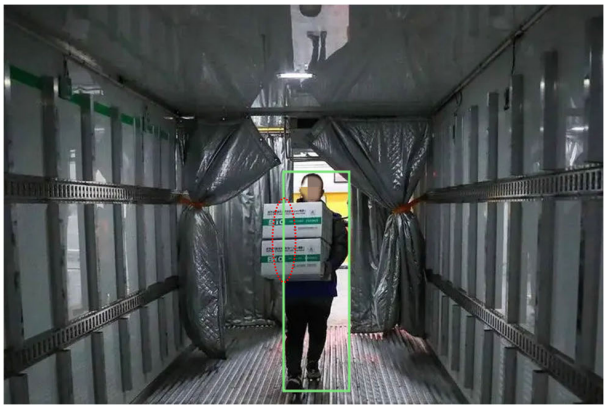


FIGURE 11. Workers obscured by vaccine boxes.

5) OPTIMIZED LOSS FUNCTION

During the process of workers transporting vaccine boxes inside the vaccine cold storage, due to the obstruction by the vaccine boxes held in their hands, the images captured by the camera cannot clearly indicate the boundaries of the workers. The annotated true boundary boxes have strong uncertainty, as shown in Figure 11. Current mainstream object detectors like YOLOv5 fundamentally model the object boundary boxes with a single Dirac distribution, that is, they directly predict the coordinates of the boundary boxes. They do not consider situations where the boundaries of the objects are not clear or complete due to obstructions, shadows, blurriness, etc. As a result, the flexibility and generalization capabilities of the detector are insufficient.

To address the issues in the YOLOv5 regression task, this study does not directly predict the absolute coordinates of bounding boxes. Instead, it predicts the probability distribution of bounding box coordinates. The Distribution Focal Loss (DFL) loss function [28] is introduced to optimize the probability distribution of bounding box coordinates, enhancing YOLOv5's ability to model bounding boxes in complex environment s such as vaccine cold storage. Specifically, an interval $[0, k]$ is first selected. After dividing the object positioning label $[l, r, t, b]$ by the down sampling multiple (stride) of the detection layer, the positioning label will fall into a sub-interval $[i, i+1]$ of the integer interval $[0, k]$. The SoftMax function is then used to discretize the interval $[0, k]$, achieving any form of discrete distribution. The network outputs $k+1$ values as the probability of falling on positioning interval nodes. When the positioning label falls in the interval $[i, i+1]$, the predicted bounding box coordinate distribution probabilities $p(i)$, $p(i+1)$ should theoretically be relatively large. Therefore, the DFL loss function is introduced to allow the network distribution to quickly focus near the label value. The formula for the DFL loss function is shown in Equation (3).

$$\mathrm{DFL}(P_i, P_{i+1}) = -(y_{i+1} - y)\log(P_i) - (y - y_i)\log(P_{i+1}) \tag{3}$$

Finally, after integrating the predicted probability distribution of the bounding box over the interval, this paper can

obtain the predicted value y^* of the bounding box. y^* can be considered as the expected value of the bounding box. Since discrete points of the interval are used in its definition, y^* can be expressed as:

$$y^* = \sum_{i=0}^{k} p(i) \cdot i \tag{4}$$

6) OBJECT TRACKING IMPLEMENTATION

The VCS-YOLOv5 algorithm can accurately detect workers inside the cold storage. By inputting the detection boxes obtained from the VCS-YOLOv5 algorithm into the DeepSORT algorithm, it ultimately achieves tracking of the workers. The DeepSORT algorithm returns the ID information of the workers as well as their three-dimensional location information.

D. BEHAVIOR RECOGNITION MODULE

To determine the final position of the vaccine box, this paper introduces a behavior recognition model to identify the actions of the workers. It needs to distinguish between transporting actions and non-transporting actions. When the worker is in the act of transporting, the vaccine box and the worker are treated as one entity, updating the position of the vaccine box in real time. When the worker transitions from transporting to non-transporting status, the current location of the worker is considered as the final position of the vaccine box, completing the locating of the vaccine box. Since transporting actions and non-transporting actions are relatively easy to distinguish, this paper adopts a behavior recognition scheme based on image classification. Firstly, the locating of the workers in the video stream is obtained through the object tracking algorithm. Then, each image is analyzed, and the corresponding area of the worker in the image is cropped and sent to the ResNet image classification network for behavior recognition. Specifically, the transporting state is defined by direct contact between the worker's hands and the vaccine box, accompanied by a noticeable lifting or moving action. Conversely, the non-transporting state is characterized by the absence of contact between the worker's hands and the vaccine box. When the category of the image corresponds to the specific action, it is believed that the person is in that state of action within a certain period.

E. VACCINE BOX LOCALIZATION MODULE

The RGB-D depth camera, using the VCS-YOLOv5 model, has obtained the position $O(u, v, d)$ of the worker's center point in the pixel coordinate system O. Here, u and v represent the pixel coordinates of the worker in the image, and z indicates the distance from that pixel coordinate to the worker. Once the worker's position is determined, the location of the vaccine box being transported in that batch is also ascertained.

To better understand and analyze the location information of the vaccine box in the vaccine cold storage environment, the pixel coordinate system is transformed into the camera coordinate system to obtain the three-dimensional spatial coordinates of the vaccine box in the vaccine cold storage. Firstly, the Zhang calibration method [29] is employed to calibrate the RGB-D depth camera and obtain the intrinsic matrix K of the camera. The intrinsic matrix is shown as follows:

$$K = \begin{bmatrix} f_x & 0 & cx \\ 0 & f_y & cy \\ 0 & 0 & 1 \end{bmatrix} \tag{5}$$

where f_x and f_y are the focal lengths of the camera along the x-axis and y-axis respectively, cx and cy represent the origin position of the pixel coordinate system. Next, this paper computes the three-dimensional spatial coordinates (x, y, z) of the target center pixel (u, v) in the camera coordinate system. The conversion formula is as follows:

$$\begin{aligned} X &= (\mathrm{u} - cx)^* d / f_x \\ Y &= (v - cy)^* d / f_y \\ Z &= d \end{aligned} \tag{6}$$

After converting the three-dimensional spatial position of this batch of vaccine boxes into the three-dimensional coordinates under the camera coordinate system, the three-dimensional position information of each vaccine box is restored based on the relative position information of the traceability code. This achieves precise locating of the vaccine boxes inside the vaccine cold storage.

IV. EXPERIMENTAL ANALYSIS

A. DATASETS

1) OBJECT DETECTION DATASET

This paper employs the widely used COCO dataset for object detection tasks to verify the efficacy and generalization capabilities of the VCS-YOLOv5 model. The COCO [30] dataset is a benchmark in the field of object detection, with over 330K images, 220K annotated images, and 1.5 million objects, spanning 80 object categories. The train2017 and val2017 sets from this dataset serve as training and validation datasets, with the former containing 118,287 images and the latter comprising 5,000 images.

2) OBJECT TRACKING DATASET

For the object tracking model's training and validation, this paper chose the MOT17 [31] pedestrian tracking dataset and a dataset collected from the SMCS environment. MOT17 is designed to evaluate the performance of multi-object tracking algorithms. It includes 21 high-definition videos from varied environments, split into training (7 videos) and testing sets (7 videos). These videos encapsulate various challenging scenes, such as dense crowds, occlusions, and changing perspectives. The dataset provides annotations including object IDs and locations. While each sequence varies in frame numbers, they range between 150 to 1,500 frames. Each sequence features multiple pedestrian objects, amassing a total of 7,251 instances, with 4,432 belonging to the training

set and 2,849 to the testing set. Additionally, to enhance the model's tracking capability in the SMCS environment, this paper has gathered 14 video clips, with 7 chosen randomly for training and the remaining 7 for testing. Each video averages 300 frames, presenting approximately 3 targets per frame.

3) BEHAVIOR RECOGNITION DATASET

This paper has chosen the UAV-Human [32] dataset and another dataset from the SMCS to train and validate the behavior recognition model. The UAV-Human is a large-scale human behavior understanding dataset featuring 67,428 multimodal video sequences and 119 action recognition entities. Furthermore, to boost the model's behavior recognition capability in the SMCS, this paper has expanded the dataset with 14 video clips. This research focuses on identifying personnel behaviors and categorizing them into carrying and non-carrying actions. Firstly, the UAV-Human dataset was processed using an object tracking model to isolate and capture individuals in the videos, producing the required training images. Positive samples were carrying actions, with the negatives being other actions. Given the potential redundancy in video-to-image conversion, this paper sampled every 8 frames for positives. For negative samples, this paper selected based on a 1:4 positive-negative ratio, especially emphasizing actions prone to misidentification, such as picking up objects, rubbing hands, and squatting.

B. EXPERIMENTAL ENVIRONMENT

The experiments were conducted on a computer equipped with an Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50GHz (344G RAM) and 8 RTX 3070 GPUs, running Ubuntu 20.04. The software environment includes Cuda11.3, PyTorch1.11.0, Python3.8, and TensorRT 8.0.3.4. A single RTX 3070 GPU was utilized during testing with a batch size set to 1.

C. EVALUATION METRICS

For the object detection model, this paper evaluated detection outcomes using four metrics: mean average precision (mAP), model parameter count, model size, and detection speed. For the object tracking model, this paper employed the Multiple Object Tracking Accuracy (MOTA), the harmonic means of identification precision and recall (IDF1), and frames processed per second (FPS) to gauge tracking performance. For behavior recognition, the accuracy (Acc) and FPS were the chosen metrics.

D. OBJECT DETECTION EXPERIMENT

1) COMPARATIVE EXPERIMENT

To verify the effectiveness of the algorithm presented in this paper, this paper conducted comparative experiments between the VCS-YOLOv5 model and other mainstream object detection algorithms, including the original YOLOv5, YOLOX, and YOLOv7. These algorithms were tested on the COCO dataset and a combined dataset of MOT17 and the SMCS scene. The experimental results are shown in Table 1.

TABLE 1. Comparison of algorithm experiments.

Dataset	Model	mAP (%)	Parameters (M)	GFLOPs	FPS
COCO	YOLOv5-L	49.0	46.5	109.3	98.9
COCO	YOLOX-L	50.1	54.2	155.6	75.3
COCO	YOLOv7-L	51.0	37.62	106.08	94.6
COCO	VCS-YOLOv5	**51.3**	48.2	118.7	80.2
MOT17 & SMCS	YOLOV5-L	46.7	46.5	109.3	98.9
MOT17 & SMCS	YOLOX-L	48.3	54.2	155.6	75.3
MOT17 & SMCS	YOLOv7-L	49.4	37.62	106.08	94.6
MOT17 & SMCS	VCS-YOLOv5	**52.9**	48.2	118.7	80.2

From Table 1, the VCS-YOLOv5 model achieved the best detection accuracy on both the COCO dataset and the fused dataset of MOT17 and SMCS. On the COCO dataset, the mAP value of the proposed method in this paper is 51.3%, which is 2.3%, 1.2%, and 0.3% higher than that of YOLOv5, YOLOX, and YOLOv7, respectively. On the fused dataset of MOT17 and vaccine cold storage, the mAP value of this method is 52.9%, which is 6.2%, 4.6%, and 3.5% higher than YOLOv5, YOLOX, and YOLOv7, respectively. Compared with YOLOX, the model parameters of this method decreased by 11%, and the computational complexity decreased by 23%. Compared with YOLOv5 and YOLOv7, there was a slight increase in the number of parameters and computational complexity, which is due to the introduction of the SE module and the decoupled detection head. In terms of model inference speed, this method achieves real-time detection, but it is slightly slower than the detection speed of YOLOv5 and YOLOv7.

Through comparison, while ensuring real-time detection, this method only increased a small number of parameters and computational complexity and achieved the best detection accuracy on multiple datasets. This further verifies the universality and effectiveness of the method proposed in this paper.

Visual Analysis of object detection Results:

For a more intuitive comparison of the detection effects of different algorithms, using the real annotation boxes as a benchmark, this paper contrasted the detection performance of the original YOLOv5 algorithm with that of the VCS-YOLOv5 algorithm. The detection results are displayed in Figure 12. In each row of images, the sequence is as follows: original image, annotated image, YOLOv5 detection results, and VCS-YOLOv5 detection results.

In environments where the workers are highly occluded, have unclear contours, or are of a smaller scale (as shown in rows 1 and 2), the original YOLOv5 algorithm may miss some detections. In contrast, the method proposed in this paper can effectively distinguish overlapping targets.

When the workers are not facing the camera directly (as seen in row 3), the original YOLOv5 algorithm exhibits notable misses. The method proposed in this paper can detect targets from different angles.

In situations where the workers are only slightly occluded (as shown in row 4), neither the original YOLOv5 algorithm nor the method from this paper misses any detections.

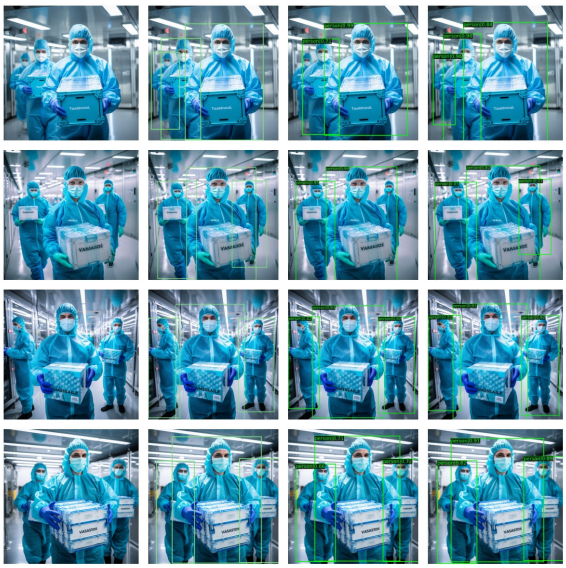


FIGURE 12. Visual analysis of object detection results.

TABLE 2. Comparison of ablation experiment results.

Model	mAP (%)	Parameters(M)	GFLOPs	FPS
YOLOv5-L	49.0	46.5	109.3	98.9
+Anchor-free	49.3(+0.3)	46.2	108.4	99.8
+C3RS	49.9(+0.6)	46.4	108.6	95.2
+Decouple Head	50.3(+0.4)	47.3	118.7	82.4
+Mish	50.6(+0.3)	48.2	118.7	80.6
+TAL	51.1(+0.5)	48.2	118.7	80.2
+DFL	**51.3**(+0.2)	48.2	118.7	80.2

However, the accuracy and confidence levels of the method proposed in this paper are notably higher.

2) ABLATION STUDY

To verify the effectiveness of the improvements introduced in this paper, we conducted ablation experiments on the COCO dataset to explore the enhancement effect of each improvement on the overall model. Since each improvement is not entirely independent and there are too many combinations of improvements, it is challenging to conduct a comprehensive analysis. Therefore, this paper demonstrates the effectiveness of each improvement incrementally. Using the original YOLOv5 network as the baseline, five sets of ablation experiments were conducted on the COCO dataset. The environment and parameter settings were kept consistent across all experiments. The model inference speed (FPS) was tested under TensorRT-FP16, excluding the time for data preprocessing and post-processing (NMS) after model output. The experimental results are shown in Table 2.

Firstly, this paper selects the YOLOv5 model as the baseline for comparison in the subsequent five sets of experiments. The detection mAP value is 49.0%, with 46.5M parameters, 109.3G GFLOPs, and an inference speed of 98.9 FPS. Experiment 1 transforms the YOLOv5 model into an anchor-free detector. The model's accuracy improved by 0.3 percentage points, with a slight reduction in parameters and computational cost, and a minor increase in inference speed. In Experiment 2, the C3 module is replaced with the C3RS module. The improved model's accuracy increased by 0.6%. Due to the introduction of the SE module, the model's parameters and computational cost rose slightly, and the inference speed decreased marginally. Experiment 3 replaced the coupled prediction head with a decoupled prediction head, resulting in a 0.4% increase in model accuracy. However, this introduced new computational costs, increasing the model's parameters by 0.9M, GFLOPs by 10G, and decreasing the inference speed to 82.4 FPS. Experiment 4 replaced the SiLu activation function in the network with the Mish activation function. Without changing the model's parameters and computational cost, the model's accuracy improved by 0.3%, proving the effectiveness of introducing the Mish activation function. Experiment 5 introduced the TAL dynamic matching strategy. Without changing the model's parameters and computational cost, the model's accuracy improved by 0.5%, and the inference speed remained almost unchanged. Experiment 6 introduced the DFL loss function. Compared with the improved model in Experiment 5, the model's accuracy increased by 0.2%. The model's parameters, computational cost, and inference speed remained virtually unchanged.

TABLE 3. Comparison of ablation experiment results.

Model	检测 mAP (%)	MOTA	IDF1	FPS
YOLOv5-DeepSORT	46.7	50.2	52.4	93.8
YOLOX-DeepSORT	48.3	51.8	55.8	70.4
YOLOv7-DeepSORT	49.4	56.7	60.5	89.5
VCS-YOLOv5-DeepSORT	**52.9**	**58.4**	**64.6**	76.3

E. OBJECT TRACKING EXPERIMENT

The detection results of the object detection algorithm need to be input into the DeepSORT algorithm to achieve multi-object tracking. To verify the performance of the VCS-YOLOv5 algorithm in pedestrian multi-object tracking, the tracking results of this algorithm are compared with those of YOLOv5-DeepSORT, YOLOX-DeepSORT, and YOLOv7-DeepSORT algorithms. Tests are conducted on the MOT17 and SMCS Fusion datasets, with results shown in Table 3.

From Table 3 compared with the three algorithms YOLOv5-DeepSORT, YOLOX-DeepSORT, and YOLOv7-DeepSORT, the algorithm presented in this paper has improved the MOTA metric by 8.2%, 6.6%, and 1.7% respectively. The IDF1 metric has also been improved by 12.2%, 8.8%, and 4.1% respectively. This method ensures real-time tracking while achieving the best tracking accuracy.

Visual Analysis of object tracking Results:

For a clearer and more intuitive presentation and analysis of the tracking effects of VCS-YOLOv5-DeepSORT, this study visualizes the tracking results on the vaccine cold storage dataset, as shown in Figure 13.

During the movement of the workers, the appearance information of the workers changes continuously with the scene. However, the prediction box can still be stably associated with the detection box and maintain the ID information unchanged. This indicates that the algorithm proposed in this

FIGURE 13. Visual analysis of object tracking results.

TABLE 4. Behavior recognition experiment results.

Dataset	Accuracy(%)	FPS
UAV-Human dataset	92.6	72.5
Vaccine cold storage scene dataset	96.8	72.5

study achieves good tracking results in the complex environment of the vaccine cold storage.

F. BEHAVIOR RECOGNITION EXPERIMENT

The object tracking algorithm can obtain the coordinate position of the pedestrian target in the image as well as the target's ID number. By using the coordinate position of the pedestrian target, the corresponding area in the image is cropped and fed into the ResNet image classification network to recognize the behavior of the person and determine whether the person is in a carrying state. To verify the performance of the behavior recognition algorithm, this paper tests datasets collected from the UAV-Human dataset and the vaccine cold storage scene separately, with the results shown in Table 4.

From Table 4, it can be observed that the accuracy of the algorithm presented in this study reached 92.6% on the UAV-Human dataset. On the vaccine cold storage scene dataset, the accuracy of this paper's algorithm reached 96.8%. In terms of model inference speed, even after adding behavior recognition, the algorithm can still meet the requirements of real-time recognition.

Visual Analysis of behavior recognition Results:

For a clearer and more intuitive presentation and analysis of the behavior recognition effect of the algorithm proposed in this study, we have visualized the behavior recognition results on the vaccine cold storage dataset, as illustrated in Figure 14. From the behavior recognition results, it can be observed that the algorithm of this paper can effectively identify the behavior of the workers under high occlusion environment, whether they are carrying or not, meeting the application requirements.

FIGURE 14. Visual analysis of behavior recognition results.

TABLE 5. Vaccine box location experiment.

Number	Actual value(x,y,z)/cm	Measurements (x,y,z)/cm	Error/cm
1	(100,100,100)	(99,106,102)	6.4
2	(200,100,100)	(198,105,98)	5.7
3	(100,200,100)	(101,207,99)	7.1
4	(100,100,200)	(98,105,199)	5.4
5	(200,200,200)	(199,206,198)	6.4

G. VACCINE BOX LOCATION EXPERIMENT

To validate the accuracy of the methodology introduced in this article for the task of vaccine box locating, this paper first chose five spots within the vaccine cold storage that could represent the spatial distribution inside. Using a laser distance meter, this paper acquired the precise coordinates of each spot, which served as the actual values. Next, test personnel simulated the behavior of cold storage workers by standing at the previously marked spots. The proposed method in this paper was then employed to locate the vaccine box, and the outcomes from the algorithm were taken as the measured values. Finally, for each spot, the Euclidean distance between the measured value and the actual value was computed, determining the overall error in coordinates. The results can be seen in Table 5. Based on Table 5, it's evident that the maximum error amounted to 7.1 cm, the minimum was 5.4 cm, and the average error stood at 6.2 cm. These findings indicate that the method can meet the precision requirements for locating vaccine boxes.

V. DISCUSSION

The experimental results of this paper indicate that the proposed VCS-YOLOv5 model demonstrates significant performance advantages in object detection and tracking tasks compared to YOLOv5, YOLOX, and YOLOv7. Ablation experiments further confirm the effectiveness of the improvements in VCS-YOLOv5. Additionally, the results of behavior recognition experiments show that this method can effectively distinguish between transporting and non-transporting actions of workers. Finally, the vaccine box localization experiment

demonstrates that this method meets the precision requirements for locating vaccine boxes in medium and small cold storages, offering a new solution for practical applications.

Compared to IoT technologies like RFID and UWB, this solution effectively overcomes signal interference in cluttered environments and minimizes hardware dependence, enhancing scalability and cost-effectiveness. Against traditional YOLO series vision-based localization, VCS-YOLOv5 shows significant advantages in handling dense obstacles and occlusions in SMCS. Additionally, this solution offers more precise tracking in confined spaces compared to multi-object tracking technologies like DeepSORT.

However, it is essential to note that this paper did not conduct experimental comparisons with IoT localization technologies such as RFID and UWB due to limitations in experimental conditions. Instead, discussions on their limitations are based on descriptions in existing literature. This may lead to some discrepancies between the conclusions drawn and their actual application performance. Additionally, the experimental datasets and results used in this study mainly come from controlled environments, which may not fully reflect the complexity and variability of the real world. For example, if the traceability code of a vaccine box is obscured upon entering the cold storage, it may lead to inaccurate locating of the box. Besides, this study recorded the relative position of vaccine boxes upon their entry into the cold storage by scanning the traceability code, aiming for precise subsequent localization. However, if these boxes are moved during shelving, such as being placed on different levels, the initial relative position information might become inaccurate, thereby impacting the final localization accuracy. In future work, it is necessary to further expand research to include more real-world environment testing, especially in more complex and unpredictable environments. This will help validate and improve the applicability and robustness of the current methods. Additionally, exploring the integration with other technologies like RFID and UWB is needed to overcome the limitations of a single technology in specific situations.

In summary, the new framework proposed in this paper can meet the needs of accurate vaccine box localization in SMCS. The VCS-YOLOv5 model demonstrates significant performance advantages in object detection and tracking tasks over traditional methods, as effectively supported by experiments. This work also shows unique strengths in handling complex environments and occlusions, yet faces challenges related to experimental conditions and dataset applicability. Future research will focus on more comprehensive testing in complex real-world environments and exploring integration with other technologies to overcome current method limitations, better adapting to practical application needs.

VI. CONCLUSION

This paper introduces an innovative solution for vaccine box localization in obstructed environment within SMCS, leveraging computer vision technology. Specifically, to address the challenge of accurately locating vaccine boxes in densely stacked and heavily obstructed SMCS, this paper exploits the strong correlation between the vaccine boxes and workers during the storage process. The vaccine box is indirectly located by focusing on the less numerous and less obstructed cold storage workers. Furthermore, to enhance the tracking accuracy of the workers, the YOLOv5 model was modified, resulting in the development of the Vaccine Cold Storages YOLOV5 (VCS-YOLOv5) model tailored for obstructed environment in SMCS. Additionally, the final location of the vaccine box is determined by a behavior recognition model, identifying instances where the workers' hands are not in contact with the vaccine box. To verify the effectiveness of the algorithm proposed in this paper, comparative experiments were conducted on multiple datasets. The experimental results indicate that the VCS-YOLOv5 model demonstrates significant performance advantages in object detection and tracking tasks compared to YOLOv5, YOLOX, and YOLOv7. Specifically, in the vaccine box localization task, the detection accuracy of VCS-YOLOv5 improved by 6.2% compared to the original YOLOv5; in the task of tracking cold storage workers, the tracking accuracy increased by 8.2%; and in the task of recognizing worker behavior, the accuracy reached 96.8%. The average error in vaccine box localization was 6.2 cm, and the overall positioning and tracking speed of the algorithm was 72 FPS, achieving the requirements for real-time localization of vaccine boxes. In future work, it is necessary to further expand the research to include more real-world environment testing, particularly in more complex and unpredictable environments. At the same time, exploring the integration with other technologies like RFID and UWB is also needed to overcome the limitations of a single technology in specific situations.

REFERENCES

[1] B. Wang and Y. Zhong, "Prospects in evaluating the safety and effectiveness of SARS-CoV-2 vaccines in humans," *Bull. Nat. Natural Sci. Found. China*, vol. 34, no. 5, pp. 581–587, 2020.

[2] Y. Zeng, X. Chen, R. Li, and H.-Z. Tan, "UHF RFID indoor positioning system with phase interference model based on double tag array," *IEEE Access*, vol. 7, pp. 76768–76778, 2019, doi: 10.1109/ACCESS.2019.2921560.

[3] B. Liu, G. Chen, H. Yuan, Z. Li, L. Chen, and L. Ma, "Research on UWB-based intelligent storage management system for hazardous chemicals," in *Industrial Safety and Environmental Protection*, vol. 48. Wuhan, China: Wuhan Safety & Environmental Protection Research Institute Co., Ltd., China Steel Group, 2022, pp. 32–36.

[4] D. Shi, H. Mi, E. G. Collins, and J. Wu, "An indoor low-cost and high-accuracy localization approach for AGVs," *IEEE Access*, vol. 8, pp. 50085–50090, 2020, doi: 10.1109/ACCESS.2020.2980364.

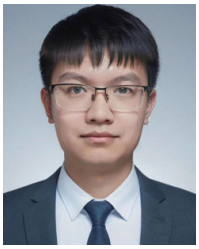
[5] Y. Liu, Z. Hou, Y. Tan, H. Liu, and C. Song, "Research on multi-AGVs path planning and coordination mechanism," *IEEE Access*, vol. 8, pp. 213345–213356, 2020, doi: 10.1109/ACCESS.2020.3039959.

[6] J. Yang, S. Wu, L. Gou, H. Yu, C. Lin, J. Wang, P. Wang, M. Li, and X. Li, "SCD: A stacked carton dataset for detection and segmentation," *Sensors*, vol. 22, no. 10, p. 3617, May 2022, doi: 10.3390/s22103617.

[7] W. Zou and B. Liu, "Robot box depalletizing method based on image instance segmentation," *Comput. Eng. Appl.*, pp. 1–9, 2023.

[8] F. Zafari, A. Gkelias, and K. K. Leung, "A survey of indoor localization systems and technologies," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2568–2599, 3rd Quart., 2019, doi: 10.1109/COMST.2019.2911558.

[9] Y. Zhang, "Smart logistics warehouse management system based on RFID and Internet of Things technology," in *Proc. 2nd Int. Conf. Netw., Commun. Inf. Technol. (NetCIT)*, Manchester, U.K., Dec. 2022, pp. 489–492, doi: 10.1109/NetCIT57419.2022.00121.
[10] H. Xiong and J. Cheng, "Investigation of short-range high precision 3D localization via UWB radio," in *Proc. IEEE Global Commun. Conf.*, Austin, TX, USA, Dec. 2014, pp. 4090–4095, doi: 10.1109/GLOCOM.2014.7037448.
[11] Y. Zhuang, C. Zhang, J. Huai, Y. Li, L. Chen, and R. Chen, "Bluetooth localization technology: Principles, applications, and future trends," *IEEE Internet Things J.*, vol. 9, no. 23, pp. 23506–23524, Dec. 2022, doi: 10.1109/JIOT.2022.3203414.
[12] M. Maheepala, M. A. Joordens, and A. Z. Kouzani, "A low-power connected 3-D indoor positioning device," *IEEE Internet Things J.*, vol. 9, no. 11, pp. 9002–9011, Jun. 2022, doi: 10.1109/JIOT.2021.3118991.
[13] G. Lasser and C. F. Mecklenbräuker, "Self-interference noise limitations of RFID readers," in *Proc. IEEE Int. Conf. RFID*, San Diego, CA, USA, Apr. 2015, pp. 145–150, doi: 10.1109/RFID.2015.7113085.
[14] V. Niculescu, D. Palossi, M. Magno, and L. Benini, "Energy-efficient, precise UWB-based 3-D localization of sensor nodes with a nano-UAV," *IEEE Internet Things J.*, vol. 10, no. 7, pp. 5760–5777, Apr. 2023, doi: 10.1109/JIOT.2022.3166651.
[15] S. Gao, H. Wang, H. Wang, and G. Wang, "A low-cost wireless location system based on zigbee technology," in *Proc. Int. Conf. Transp., Mech., Electr. Eng. (TMEE)*, Changchun, Dec. 2011, pp. 1665–1668, doi: 10.1109/TMEE.2011.6199531.
[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[17] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[18] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.
[19] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, 2023, pp. 7464–7475, doi: 10.1109/CVPR52729.2023.00721.
[20] K. Saleh, S. Szénási, and Z. Vámossy, "Occlusion handling in generic object detection: A review," in *Proc. IEEE 19th World Symp. Appl. Mach. Intell. Informat. (SAMI)*, Herl'any, Slovakia, Jan. 2021, pp. 000477–000484, doi: 10.1109/SAMI50585.2021.9378657.
[21] A. Wang, Y. Sun, A. Kortylewski, and A. Yuille, "Robust object detection under occlusion with context-aware CompositionalNets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 12642–12651.
[22] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3645–3649.
[23] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9626–9635.
[24] S. Xu, X. Wang, W. Lv, Q. Chang, C. Cui, K. Deng, G. Wang, Q. Dang, S. Wei, Y. Du, and B. Lai, "PP-YOLOE: An evolved version of YOLO," 2022, *arXiv:2203.16250*.
[25] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7132–7141.
[26] D. Misra, "Mish: A self regularized non-monotonic activation function," 2019, *arXiv:1908.08681*.
[27] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "TOOD: Task-aligned one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3490–3499.
[28] X. Li, W. Wang, and L. Wu, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21002–21012.
[29] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Jun. 2000.
[30] T. Y. Lin, M. Maire, and S. Belongie, "Microsoft COCO: Common objects in context," in *Proc. Comput. Vis.-ECCV 13th Eur. Conf.*, Zurich, Switzerland, Berlin, Germany: Springer, Sep. 2014, pp. 740–755.
[31] P. Dendorfer, A. Ošep, A. Milan, K. Schindler, D. Cremers, I. Reid, S. Roth, and L. Leal-Taixé, "MOTChallenge: A benchmark for single-camera multiple target tracking," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 845–881, Apr. 2021.
[32] T. Li, J. Liu, and W. Zhang, "UAV-human: A large benchmark for human behavior understanding with unmanned aerial vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 16266–16275.



CHEN LIANG was born in Guangxi, China, in 1998. He received the degree from North China Electric Power University. He is currently pursuing the master's degree with the Shenyang Institute of Automation, Chinese Academy of Sciences. His main research interest includes computer vision.

WEI YANG received the degree from the Department of Electronic Engineering, Fudan University, in 1991, and the master's degree in pattern recognition and intelligent systems from the Shenyang Institute of Automation, Chinese Academy of Sciences, in 1997. He is currently a Master's Tutor and a Researcher with the Shenyang Institute of Automation, Chinese Academy of Sciences. His main research interest includes public health big data analysis.

LONGLONG PANG was born in Anhui, China, in 1993. He received the degree from Lanzhou Jiaotong University. He is currently pursuing the Ph.D. degree with the Shenyang Institute of Automation, Chinese Academy of Sciences. His main research interests include public health big data analysis and machine learning.

ZHUOZHANG ZOU was born in Changchun, Jilin, China, in 1998. He received the master's degree from the Shenyang Automation Research Institute, in 2023. His main research interest includes natural language processing.

QUANGAO LIU was born in Yibin, Sichuan, in 2000. He received the degree from Southwest Jiaotong University. He is currently pursuing the master's degree with the Shenyang Institute of Automation, Chinese Academy of Sciences. His main research interest includes public health big data analysis.