IEEE Open Journal of the
Industrial Electronics Society

Received 14 February 2024; revised 27 March 2024; accepted 9 April 2024. Date of publication 15 April 2024;
date of current version 22 April 2024. The review of this article was arranged by Associate Editor Paulo Leitão.

Digital Object Identifier 10.1109/OJIES.2024.3388632

Improved YOLOX-DeepSORT for Multitarget Detection and Tracking of Automated Port RTG

ZHENGTAO YU [1], XUEQIN ZHENG [2], JUN YANG [3] (Fellow, IEEE), AND JINYA SU [1] (Member, IEEE)

[1]School of Automation, Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Southeast University, Nanjing 210096, China
[2]Meituan, Beijing 100102, China
[3]Department of Aeronautical and Automotive Engineering, Loughborough University, LE11 3TU Loughborough, U.K.

CORRESPONDING AUTHORS: JINYA SU; JUN YANG (e-mail: sucas@seu.edu.cn e-mail: j.yang3@lboro.ac.uk)

The work of Jinya Su was supported in part by the National Natural Science Foundation of China under Grant 62303110 and in part by the Start-Up Research Fund of Southeast University under Grant RF1028623226.

ABSTRACT Rubber tire gantry (RTG) plays a pivotal role in facilitating efficient container handling within port operations. Conventional RTG, highly depending on human operations, is inefficient, labor-intensive, and also poses safety issues in adverse environments. This article introduces a multitarget detection and tracking (MTDT) algorithm specifically tailored for automated port RTG operations. The approach seamlessly integrates enhanced YOLOX for object detection and improved DeepSORT for object tracking to enhance the MTDT performance in the complex port settings. In particular, Light-YOLOX, an upgraded version of YOLOX incorporating separable convolution and attention mechanism, is introduced to improve real-time capability and small target detection. Subsequently, OSNet-DeepSORT, an enhanced version of DeepSORT, is proposed to mitigate ID switching challenges arising from unreliable data communication or occlusion in real port scenarios. The effectiveness of the proposed method is validated in various real-life port operations. Ablation studies and comparative experiments against typical MTDT algorithms demonstrate noteworthy enhancements in key performance metrics, encompassing small target detection, tracking accuracy, ID switching frequency, and real-time performance.

INDEX TERMS DeepSORT, multitarget tracking, rubber tire gantry (RTG), target detection, YOLOX.

I. INTRODUCTION

As global trade continues to advance, ports emerge as pivotal nodes within the global maritime container transportation network, requiring improved efficiency and operational cost optimization [1]. To improve the efficiency of conventional ports and strengthen their competitiveness, the automation transformation of rubber tire gantry (RTG) operations becomes imperative, thus elevating the overall automation level in port operations. An automated and unmanned RTG transformation solution can reduce port operating costs while ensuring efficient operations and operational safety. Within port environments, along with the RTGs, there exist internal and external container transporters, on-site safety officers, and various small vehicles that navigate the port area. Consequently, for the autonomous operation and automatic navigation of RTG, real-time detection and tracking of the surrounding targets become indispensable. This ensures that the tracking outcomes are seamlessly integrated into the decision-making processes of the RTG system. Consequently, multitarget detection and tracking (MTDT) in port scenes is of great significance. The related studies about target detection and target tracking are reviewed as follows.

A. TARGET DETECTION

State-of-the-art (SOTA) target detection algorithms predominantly leverage deep learning networks, which can be broadly categorized into two-stage algorithms and one-stage algorithms based on their detection processes [2]. The two-stage algorithms involve first generating candidate bounding boxes from input images, and then, using convolutional neural networks (CNNs) to extract target features for subsequent target detection tasks. Well-known representatives of this category include R-CNN [3] and its enhanced versions such as Fast R-CNN [4] and Faster R-CNN [5]. One-stage algorithms treat

target detection as a regression problem, bypassing the step of pregenerating candidate bounding boxes. These algorithms rely solely on CNNs to detect targets in input images. Typical one-stage algorithms include Yolo v2-v5 [6], [7], [8], [9], FSSD [10], RSSD [11], using the anchor box mechanism to improve model positioning accuracy, and anchor-free methods such as CornerNet [12], CenterNet [13], FCOS [14], and YOLOX [15], allowing anchor-free methods to be integrated into first-level algorithms for improved performance.

B. MULTIPLE TARGET TRACKING

SOTA deep-learning-based multitarget tracking algorithms are primarily categorized into two classes: Detection-based tracking (DBT) and joint detection tracking (JDT) [16], which are elaborated as as follows.

1) DBT PARADIGM TRACKING

DBT paradigm tracking algorithms employ a separate target detection network before conducting multitarget tracking. This approach currently dominates the landscape of multitarget tracking with representatives including SORT [17], DeepSORT [18], and ByteTrack [19]. SORT has a simple structure with a low computational burden. However, in cases of occlusion, its tracking results may exhibit instability, leading to more noticeable ID switches. DeepSORT improves the SORT algorithm by adding a feature extraction network to capture target's appearance features. ByteTrack introduces a low-confidence target association matching, further enhancing tracking performance in occluded scenarios. However, this algorithm is highly dependent on target detection results, since it abandons the feature extraction network.

2) JDT PARADIGM TRACKING

JDT paradigm algorithms achieve target detection, feature extraction, and data association operations within a single network, allowing effective information sharing among different stages. CenterTrack [20] is a multitarget tracking algorithm utilizing an anchor-free approach. By transforming the multitarget tracking problem into a tracking problem based on the target's center point, it eliminates the need for separately extracting appearance features, thereby reducing model complexity and enabling real-time operation. However, this design results in a higher frequency of ID switches during long-term tracking. FairMOT [21] combines target detection and feature extraction via a deep layer aggregation network. It can improve tracking performance, particularly in densely populated target scenarios.

It should be also noted that many existing MTDT solutions often rely on additional sensors such as lasers or RGBD cameras. For example, the Mask R-CNN and LMB-based target detection and tracking scheme [22] is based on infrared images. The authors in [23], [24], and [25] rely on devices like lidar for target detection. It is worth noting that these lasers or specialized camera sensors may incur additional cost and also tend to yield suboptimal data quality in the face of weather variations within the port environment, such as fog, night conditions, or intense brightness.

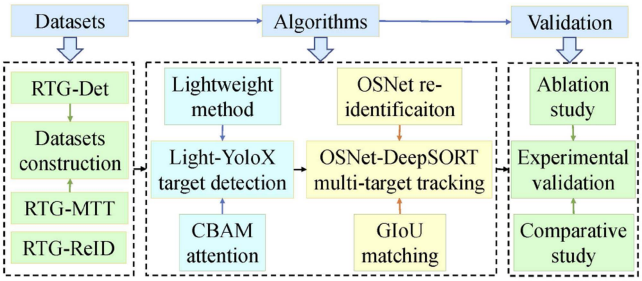


FIGURE 1. Overall structure of the proposed method.

This study focuses primarily on addressing the challenges of MTDT for automated RTG operations within port environments. The main challenges include scale variations induced by target motion and shape variations arising from target occlusion. The overall structure of the study is displayed in Fig. 1, which includes datasets construction, target detection network, multitarget tracking network, and experimental validation. To the best of our knowledge, this is the first study that integrates enhanced YOLOX and improved DeepSORT to address MTDT problems in port scenarios specifically for automated RTG operations. The key contributions are outlined as follows.

1) MTDT dataset specifically tailored for automated RTG operations within port scenes is produced.[1]
2) Real-time MTDT algorithm is proposed by seamlessly integrating enhanced YOLOX and improved DeepSORT, which allows for fast, high-precision target detection and tracking within complex port environments.
3) Ablation studies and comparative experiments against SOTA MTDT algorithms are performed, showing enhanced performance in critical metrics such as small target detection, tracking accuracy, ID switching frequency, and frames per second (FPS).

II. DATASETS CONSTRUCTION

Only public datasets are insufficient to train models that meet the specific operational requirements of RTGs in port environments. Therefore, it is critical to create datasets specific to port scene. These datasets will be used for training and validating the target detection network, training the reidentification network within the multitarget tracking network, and validating and testing the multitarget tracking network. The constructed datasets and their corresponding purposes are displayed in Table 1.

A. DATASET COLLECTION

Most of the data were collected by two Hanwha XNF-8010RVM industrial cameras, which were located at a height

[1]The datasets in this study will be shared openly upon article publication.

TABLE 1. Datasets in This Study and Their Purposes

RTG-Det	Target detection training and evaluation
RTG-MTT	Multiple target tracking evaluation
RTG-ReID	Re-identification network training

FIGURE 2. RTG platform and camera installation in this study. The red box shows the location where the AXIS P1448-LE or Hanwha XNO-6080R cameras are mounted, with lane lines in both directions. The yellow box shows the Hanwha XNF-8010RVM camera mounting location.

FIGURE 3. Sample images containing person, car, ITK, and ETK. (a) Person. (b) Car. (c) ITK. (d) ETK.

TABLE 2. Number of Different Targets in the RTG-Det Dataset

Period	ETK	ITK	Car	Person	total
daytime	1527	1157	640	417	3741
night	396	349	37	51	833

of about 3 m on both sides of the RTG, and some of the data were collected by AXIS P1448-LE and Hanwha XNO-6080R industrial cameras mounted along the lane lines' direction as shown in Fig. 2.

The captured video resolutions are of 1920 × 1080 with 30 FPS. Due to the operational needs of a fully automated RTG, each lane line camera is installed with the adjusted focus and aperture, and camera calibration [26] is performed in the field. A total of 23 video clips of varying durations ranging from a few tens of seconds to more than 20 min were captured, 5 of which were captured at night, 2 of which were captured at nightfall, and the remaining 18 of which were captured during daytime with sufficient sunlight illumination. The datasets were produced in such a way that data with no detection targets in the video frames were discarded. The sample images are illustrated in Fig. 3.

B. TARGET DETECTION DATASET

Based on the video clips captured in Section II-A, video frames containing the targets of interest are saved by every ten frames. The compiled training set comprises a total of 4574 sample images, covering four categories of targets including security officer (person), small cars (car), internal terminal tractors (ITK), and external terminal tractors (ETK). The proportions of different targets under daytime or night before data augmentation are shown in Table 2. The collected sample images were manually labeled using the LabelImg software tool to produce a dataset of PASCAL VOC format [27]. The ratio of images in the training, validation and test sets without image enhancement is about 8:1:1. The ratio of images of small target images (less than 0.1 of the image) to normal target images in the dataset is 1:1.2. To enhance model robustness, data augmentation techniques are then applied, which include horizontal flipping, image rotation, random Gaussian noise addition, and random brightness adjustments. After data augmentation, the dataset (named RTG-Det dataset) comprises 15 237 sample images.

C. MULTIPLE TARGET TRACKING DATASETS

Multiple target tracking-related datasets in this article include the RTG-ReID dataset for reidentification network and the RTG-MTT dataset for the multitarget tracking algorithm. These images are manually annotated following the format of the Market-1501 pedestrian reidentification dataset [28]. The RTG-ReID dataset created in this article comprises a total of 3 345 sample images. The training and test sets were divided by a ratio of 1:1, i.e., 1769 images for each set. The RTG-MTT dataset, with a total number of 5956 sample images, uses the same format as the MOT17 dataset, and the annotations include marking regions of interest, assigning unique ID values to each target, and specifying the category of each target.

III. TARGET DETECTION BY LIGHT-YOLOX

Different from the anchor-based detection and simplified label assignment strategy employed in earlier generations of YOLO algorithms, Ge et al. [15] integrate new technologies to achieve higher detection accuracy and faster detection speed. Therefore, building upon the YOLOX framework, this study introduces specific enhancements to better align with the real-time and robust requirements of RTG target detection in port scenes, which are detailed as follows including network lightweight method and attention mechanism.

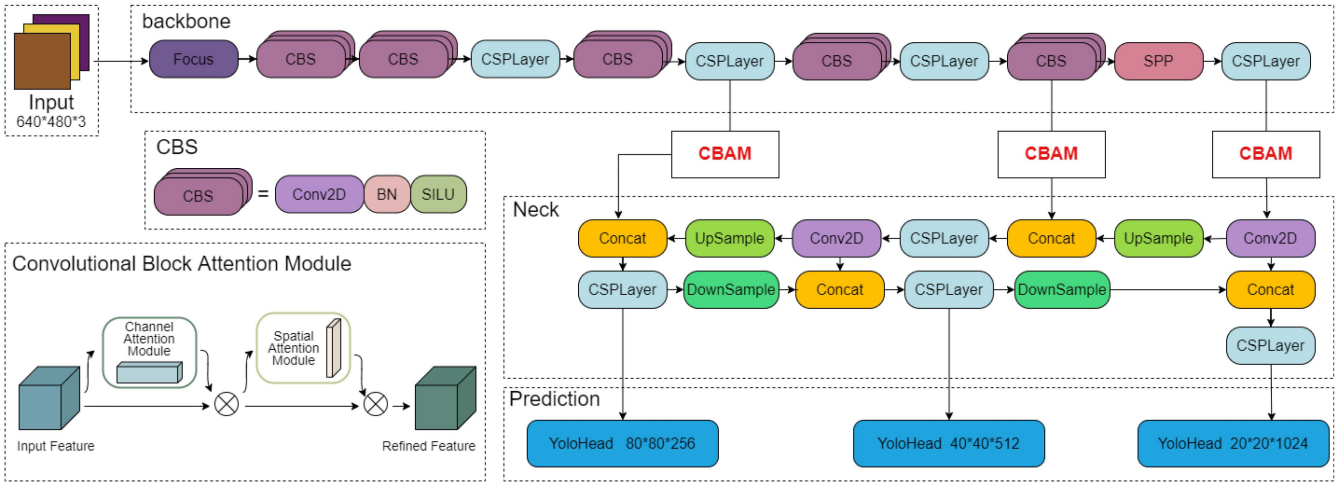


FIGURE 4. Target detection network structure with added CBAM module.

A. NETWORK LIGHTWEIGHT METHOD

To fulfill the real-time operational requirements imposed by computation resource-constrained devices in port environments, the original network's standard convolutions are replaced with depthwise separable convolutions. This strategic tradeoff introduces a marginal reduction in accuracy, but concurrently reduces the model's size and parameter count, thereby augmenting the model's inference speed.

Depthwise separable convolutions were initially proposed in MobileNet [29]. This approach changes the convolution operation of the convolution kernel, leading to a decrease in the model parameters and computational workload. The parameter ratio between standard convolution and depth-separable convolution is shown in (1), while the corresponding computation quantity ratio is expressed in (2). M denotes the number of channels of the input tensor, D_k denotes the size of the convolution kernel, N denotes the number of channels of the output feature tensor, and D_o denotes the size of the output tensor.

$$\frac{D_k \times D_k \times M + M \times N}{D_k \times D_k \times M \times N} = \frac{1}{N} + \frac{1}{D_k^2} \tag{1}$$

$$\frac{D_k \times D_k \times M + M \times N) \times D_o \times D_o}{D_k \times D_k \times M \times N \times D_o \times D_o} = \frac{1}{N} + \frac{1}{D_k^2}. \tag{2}$$

Typically utilizing a 3×3 convolution kernel for both standard and depthwise separable convolutions, and considering a significant number of feature tensors N, the parameters of depthwise separable convolution are approximately 1/9 of those in standard convolution. This highlights the substantial reduction in both parameter count and computational overhead achieved by depth-separable convolutions compared to standard convolutions.

B. ATTENTION MECHANISM

Moreover, to increase model's detection ability for small targets and model performance under various light conditions, an attention mechanism is added to the Neck part of the YOLOX network to enhance network's feature extraction ability.

Convolution block attention module (CBAM) [30] is a lightweight and versatile attention module with structural diagram in the lower left corner of the Fig. 4. This module consists of two independent submodules: the channel attention module and the spatial attention module. They are connected in series and can infer attention weights from both the channel and spatial dimensions, respectively. First, the feature map Fea is input into the channel attention module, which extracts the channel dimension attention feature map $M_c(\text{Fea})$. This attention feature map is then element-wise multiplied with the input feature map Fea, resulting in a feature map Fea$'$ with channel attention weights. Second, Fea$'$ is input into the spatial attention module, which extracts the spatial dimension attention feature map $M_s(\text{Fea}')$. This attention feature map is then element-wise multiplied with the feature map Fea$'$ with channel attention weights, resulting in a feature map Fea$''$ with both channel and spatial attention weights. Equations (3) and (4) show the element-wise multiplication calculation. This feature map has a stronger feature representation ability. In the formula, $\otimes$ represents element-wise multiplication, Fea is the input feature map, $M_c(\text{Fea})$ is the channel attention feature map output by the channel attention module, Fea$'$ is the feature map with channel attention weights, $M_s(\text{Fea}')$ is the spatial attention map output by the spatial attention module, and Fea$''$ is the feature map with both channel and spatial attention weights output by CBAM.

$$\text{Fea}' = M_c(\text{Fea}) \otimes \text{Fea} \tag{3}$$

$$\text{Fea}'' = M_s(\text{Fea}') \otimes \text{Fea}'. \tag{4}$$

In summary, the CBAM module is added to the three connections between the backbone part and the neck part of YOLOX to improve the feature extraction capability of the network for large, medium, and small targets, and the improved network structure diagram is shown in Fig. 4.

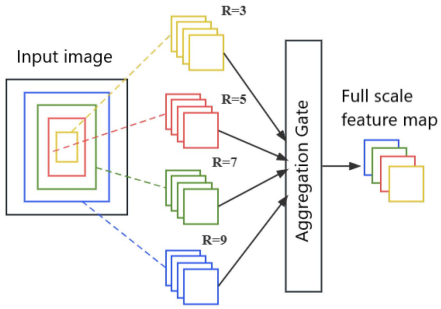


FIGURE 5. Overall structure of the OSNet residual block.

IV. MULTITARGET TRACKING BY OSNET-DEEPSORT

The DeepSORT algorithm, a detection-based multitarget tracking approach, utilizes the outputs of a single-object detection network as input for multitarget tracking. This article introduces improvements to the DeepSORT algorithm to effectively address the challenges associated with target ID switches during tracking, thereby augmenting the model's overall tracking performance.

A. REIDENTIFICATION NETWORK IMPROVEMENTS

In multitarget tracking algorithms, the main purpose of incorporating reidentification technology is to extract features from the targets and measure the distance between these features to determine if targets in different frames belong to the same entity. Although the DeepSORT algorithm adds a deep feature extraction network based on the SORT algorithm, there are still a number of limitations. For example, the number of layers of the feature extraction network is small, and therefore, its feature extraction ability is limited; in addition, it does not make full use of the feature information between different channels.

In this study, OSNet [31], a lightweight reidentification network capable of learning full-scale features, is adopted. It is composed of multiple residual blocks, where each block incorporates convolutional feature streams with varying receptive field sizes to capture spatial features at different scales. The structure of the residual block is depicted in Fig. 5. In this illustration, "R" signifies the size of the receptive field and "AG" represents the universal aggregation gate. The AG dynamically merges spatial features from different scales based on weight information. Initially, full-scale residual blocks are used to extract image features in various receptive fields. Subsequently, features from distinct scales are individually input into the AG universal aggregation gate, which assigns varying weights to features of different scales and amalgamates them to generate full-scale feature maps. Finally, the obtained full-scale feature maps undergo measurement to derive the ultimate reidentification results.

B. MATCHING METHOD

The original DeepSORT encounters limitations in cascade matching due to inherent challenges in the tracking process, including illumination variations, alterations in viewpoints, and occlusions. These factors can induce substantial alterations in the appearance of the same between frames, causing similarity scores between the appearance features of consecutive frames to drop below a predetermined threshold, thereby leading to the failure of cascade matching. Consequently, the successful matching of detection boxes and tracked trajectories through IoU after cascade matching failure becomes critical. In the context of automated RTG operations, video transmission can experience degradation due to network fluctuations, resulting in a reduction in the number of video frames transmitted. This situation can lead to short-distance movement of the same target between frames received by the perceptual model. As a result, slight movement of detection boxes, despite their close proximity, can cause the detection box from the subsequent frame to lack overlap with the trajectory prediction box established from the preceding frame. This absence of overlap results in an IoU value of 0, rendering the IoU matching of the DeepSORT algorithm incapable of achieving successful matching. Consequently, this contributes to the issue of identity switching.

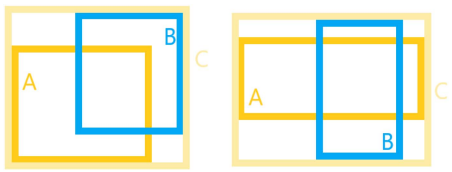


FIGURE 6. Two representations of the three rectangular boxes of GIoU.

To mitigate these issues, this article introduces the GIoU [32] into the DeepSORT algorithm, replacing the original IoU matching. This modification helps alleviate ID switching problems caused by occlusions and dropped video frames, thereby enhancing the tracking performance. The GIoU calculation process is expressed in (5), where A and B represent two rectangular boxes, $\mathrm{IoU}(A, B)$ denotes the IoU value between the two boxes, C represents the area of the minimum bounding box that contains both rectangular boxes, and $A \cup B$ represents the union of the two rectangular boxes. The relationship between A, B, and C is as shown in the Fig. 6.

$$\mathrm{GIoU}(A, B) = \mathrm{IoU}(A, B) - \frac{|C - (A \cup B)|}{|C|}. \tag{5}$$

GIoU is designed to consider not only the overlap between two rectangular boxes but also the nonoverlapping parts, allowing for a better measurement of the spatial relationship between the boxes. As a result, in situations where occlusion occurs between targets or video frames are dropped, setting an appropriate GIoU threshold can resolve the issue of failed IoU association matching due to nonintersecting rectangles.

V. EXPERIMENT VALIDATION

In this section, we perform a comprehensive evaluation of the proposed algorithm under different settings and compare

TABLE 3. Computation Environment for the Algorithms

Operating System	Ubuntu 20.04.5 LTS
CPU	Intel i9-9900K@3.60GHz
Memory	64G
Graphics Card	NVIDIA GeForce RTX3080Ti
Deep Learning Framework	Pytorch 1.13.0
Programming Language	Python 3.8
CUDA	Cuda11.1

TABLE 4. Results of Ablation Study on Light-YOLOX for Target Detection

Group	Accuracy(%)	Recall(%)	mAP (%)	FPS
(a)	90.93%	93.02%	93.13%	43
(b)	89.59%	92.57%	91.67%	**56**
(c)	**93.24%**	**95.81%**	**95.54%**	52

against several SOTA MTDT algorithms. The improved versions of YOLOX and DeepSORT are referred to as Light-YOLOX and OSNet-DeepSORT, respectively. The information about the experimental environment is shown in Table 3.

Regarding the evaluation of algorithm performance, four metrics are adopted for the improved Light-YOLOX detection algorithm, which include precision, recall, mean average precision (mAP), and FPS. For the improved OSNet-DeepSORT multitarget tracking algorithm, the selected evaluation metrics include multitarget tracking accuracy (MTTA), multitarget tracking precision (MTTP), and number of target ID switching (IDS) and FPS.

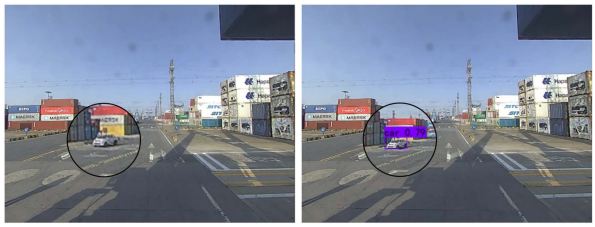
A. DETECTION PART

For the training of the target detection model, input image size is 640 × 480, the number of iterations is 500, and the batch size is 16. For algorithm optimization, the stochastic gradient descent method is adopted with an initial learning rate of 0.001, momentum of 0.94, and weight decay of 0.0005. To verify the effectiveness of the improved target detection algorithm, both ablation experiments (i.e., against the baseline YOLOX) and comparison experiments (i.e., against other popular detection networks) are conducted.

1) ABLATION STUDY

To verify the effect of different model modifications, this part uses the RTG-Det dataset to perform an ablation study on the improved Light-YOLOX model. Group (a) is the original YOLOX algorithm as the benchmark model; Group (b) is based on YOLOX (a), but replaces the standard convolution by a depth-separable convolution; and Group (c) builds on Group (b) and introduces the CBAM module. The results of ablation study are shown in Table 4, with the following observations.

1) By replacing the standard convolution with depth-separable convolution, the model accuracy, recall, and mAP are slightly reduced, but the detection speed FPS is significantly improved from 43 to 56.
2) Adding the CBAM module can significantly improve the model detection performance with an improvement of 3.65%, 3.24%, and 3.87% in accuracy, recall, and mAP, respectively. Although the addition of the CBAM module also brings additional model parameters and computation load, as a result, the FPS drops slightly.
3) Overall speaking, Light-YOLOX outperforms baseline YOLOX in terms of accuracy (3.56%), recall (3.61%), mAP (3.92%), and FPS (9) and fulfills the requirement of RTG real-time target detection in ports.



(a) (b)

FIGURE 7. YOLOX versus Light-YOLOX for small target detection. (a) YOLOX test results. (b) Light-YOLOX test results.

TABLE 5. Light-YOLOX Against SOTA Target Detection Algorithms

Model	Accuracy(%)	Recall(%)	mAP (%)	FPS
Faster R-CNN	92.21%	93.68%	93.67%	16
Yolo V3	86.96%	90.47%	90.78%	31
Yolo V5	89.37%	92.16%	92.57%	38
CenterNet	88.45%	91.59%	91.68%	34
YoloX	90.93%	93.02%	93.13%	43
Light-YoloX	**94.49%**	**96.63%**	**97.05%**	**52**

The results of the sample target detection under the default YOLOX and the improved Light-YOLOX are shown in Fig. 7. It can be seen that the original YOLOX does not detect the presence of the car, but the improved Light-YOLOX succeeds because of its improved detection performance for small targets with enhanced feature extraction capability.

2) COMPARATIVE STUDY

The improved Light-YOLOX is also compared against some commonly used target detection models including Faster R-CNN, CenterNet, Yolo V3, Yolo V5, and YOLOX. The comparative results are shown in Table 5 with the following observations.

1) The two-stage detection algorithm Faster R-CNN outperforms Yolo V3, Yolo V5, YOLOX, and CenterNet in terms of accuracy, recall, and mAP but with the price of significantly high computation load (e.g., FPS of 16).
2) The improved Light-YOLOX proposed in this study significantly outperforms the Faster R-CNN not only in FPS, but also in accuracy (by 2.28%), recall (by 2.95%), and mAP (by 3.38%).

The sample target detection results of the proposed Light-YOLOX are shown in Fig. 8, which includes different port scenarios such as strong light, low light, occluded targets, and small targets. It follows from Fig. 8(a) and (b) that Light-YOLOX can accurately identify the target under varying light conditions. It follows from Fig. 8(c) that the improved Light-YOLOX also works well in the presence of mutual occlusion

TABLE 6. Comparative Results of Different Model Combinations for Multitarget Tracking

Group	Model combination	MTTA (%)	MTTP (%)	IDS (times)	FPS (frame/second)
1	YoloX + DeepSort	79.58%	83.23%	298	32
2	YoloX + OSNet-DeepSORT	82.53%	85.97%	251	28
3	**Light-YoloX + OSNet-DeepSORT**	**86.14%**	**89.92%**	**227**	**37**

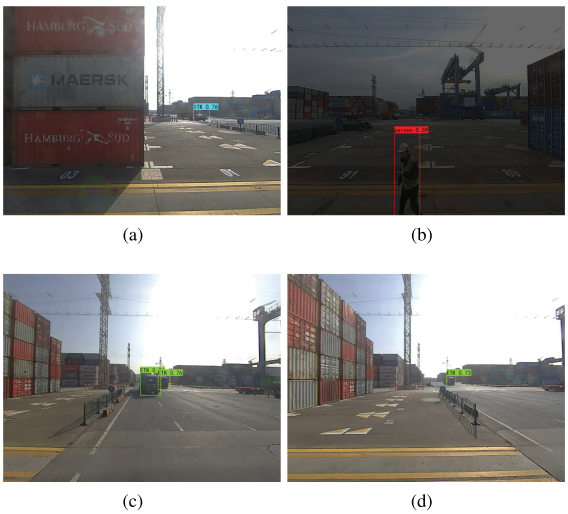


FIGURE 8. Target detection results of the proposed Light-YOLOX under different port scenes. (a) Strong light. (b) Low light. (c) Occluded objects. (d) Small targets.

between the targets. It follows from Fig. 8(d) that the Light-YOLOX can also correctly identify small targets far from the RTG.

B. TRACKING PART

1) ABLATION STUDY

This part presents the results for target tracking model evaluation on the self-made RTG-MTT dataset. In particular, three sets of experiments are conducted for performance comparison. Group 1 uses the original YOLOX as the detector, combined with the original DeepSORT for multitarget tracking. Group 2 uses the original YOLOX as the detector, combined with the improved OSNet-DeepSORT for multitarget tracking to verify the improved effect of the OSNet-DeepSORT algorithm. Group 3 uses the Light-YOLOX algorithm as the detector, combined with OSNet-DeepSORT for multitarget tracking, which is to verify the effectiveness of the proposed algorithm combination. The comparative results are summarized in Table 6 with the following observations.

1) MTTA and MTTP of Group 2 increase by 2.95% and 2.74% compared to Group 1, and the number of ID switches drops 47 times. This is because OSNet-DeepSORT algorithm's OSNet for feature extraction and GIoU for data association matching can improve model tracking accuracy and reduce the number of ID switches. However, the FPS of Group 2 decreases by 4, since the OSNet is more complex than the wide residual network structure in the original DeepSORT.
2) MTTA and MTTP of Group 3 increase by 3.61% and 3.95% compared to Group 2, the number of ID switching also drops 24 times, and FPS increases by 9. This shows that the tracking accuracy DBT paradigm based multitarget tracking depends to a certain extent on detector performance, and the proposed Light-YOLOX algorithm further improves the tracking accuracy of the OSNet-DeepSORT algorithm.
3) MTTA and MTTP of Group 3 are increased by 6.56% and 6.69% than Group 1, the number of ID switching drops 71 times, and the FPS increases by 5. This shows that compared with the original YOLOX-DeepSORT algorithm combination, the proposed combination of Light-YOLOX and OSNet-DeepSORT significantly improves tracking accuracy, ID switching times, and FPS, verifying the rationality of the proposed method.

2) COMPARATIVE STUDY

The combination of Light-YOLOX and OSNet-DeepSORT is also compared against the combinations of Light-YOLOX with other tracking algorithms such as SORT, ByteTrack, and FairMOT on the self-made RTG-MTT dataset. Comparative tracking results are shown in Table 7 with the following observations.

1) The proposed algorithm combination (Light-YOLOX + OSNet-DeepSORT) significantly outperforms other three algorithms in terms of MTTA, MTTP, and IDS.
2) While in terms of FPS, Light-YOLOX + SORT is the best, reaching 56 with its simple algorithm structure but with the worst tracking performance.
3) The proposed algorithm reaches 37 FPS, second only to the Light-YOLOX + SORT algorithm, but is sufficient for real-world applications.

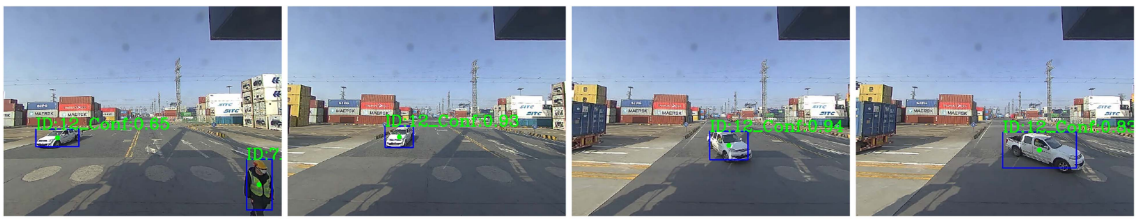
3) SCENARIO TEST RESULTS

This part presents the test results of the proposed multitarget tracking algorithm under different port scenarios. The first scenario involves an RTG traversing an intersection, with a stationary car nearby and a security officer on patrol. In this setting, there is relative motion between the RTG and the tracked target. Tracking results on cars and pedestrians are shown in Fig. 9. Throughout the process, the tracking algorithm can track the target correctly, while keeping the same assigned ID.

The second scenario involves an RTG passing through an intersection, and there are two external container trucks driving at the adjacent intersection. The tracking results on a container truck are shown in Fig. 10, with the following

TABLE 7. Comparative Experimental Results of Different Multitarget Tracking Model Combinations

Group	Multiple target tracking model	MTTA (%)	MTTP (%)	IDS (times)	FPS (frame/second)
1	Light-YoloX + SORT	76.85%	81.59%	347	**56**
2	Light-YoloX + ByteTrack	79.73%	84.16%	282	34
3	FairMOT	78.32%	80.87%	319	30
4	**Light-YoloX + OSNet-DeepSORT**	**86.14%**	**89.92%**	**227**	37



(a) (b) (c) (d)

FIGURE 9. Sample tracking results of the proposed algorithm for cars and people. (a) Frame 152. (b) Frame 184. (c) Frame 217. (d) Frame 256.

(a) (b) (c) (d)

FIGURE 10. Sample tracking results of the proposed algorithm for ETK tracking. (a) Frame 51. (b) Frame 73. (c) Frame 96. (d) Frame 125.

observations. In Fig. 10(a), two external trucks are successfully identified, and in Fig. 10(c), the external truck No. 4 completely blocks the external truck No. 5 but in Fig. 10(d), the No. 4 external truck no longer blocks the No. 5 external truck, the model resumes tracking the external truck No. 5.

Through the aaforementioned comprehensive analysis, the proposed multitarget tracking algorithm integrating the Light-YOLOX and OSNet-DeepSORT algorithm can correctly identify and track the targets of interest around the RTG. Furthermore, it also exhibits the capability to successfully reidentify the target even when occluded, ensuring the continuity of target tracking. However, while our algorithm demonstrates considerable promises in enhancing the efficiency and reliability of a fully automated RTG system in port environments, we acknowledge that its application is also subject to certain limitations. Specifically, our algorithm is tailored for standard operating conditions typically encountered in ports, which include environments with adequate lighting at night and weather conditions that allow the port to operate properly. As a result, under extreme weather conditions such as typhoons, severe rainstorms, heavy snowfall, or dense fog, port operations may deviate significantly from normal conditions, leading to a suspension of RTG activities.

VI. CONCLUSION

This article introduces a DBT approach by integrating enhanced YOLOX and improved DeepSORT for automated RTG MTDT. An enhanced Light-YOLOX method is proposed for target detection, incorporating separable convolution and an attention mechanism to enhance the baseline YOLOX model's real-time capabilities while improving the detection performance for small targets. Regarding multitarget tracking, an OSNet-DeepSORT is presented to address ID switching issues arising from phenomena such as unstable data communication or occlusion in real-liFe port scenes. Comprehensive experiments are conducted, including an ablation study and comparison with SOTA detection and tracking algorithms, in various port scenarios (e.g., cars, pedestrians, and external container chassis) to validate the effectiveness of the proposed algorithm. The comparative results show improved performance in both detection metrics and tracking metrics.

REFERENCES

[1] X. Clark, D. Dollar, and A. Micco, "Port efficiency, maritime transport costs, and bilateral trade," *J. Dev. Econ.*, vol. 75, no. 2, pp. 417–450, 2004, doi: 10.1016/j.jdeveco.2004.06.005.

[2] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," in *Proc. IEEE*, vol. 111, no. 3, pp. 257–276, Mar. 2023, doi: 10.1109/JPROC.2023.3238524.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.

[4] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.
[6] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7263–7271.
[7] K. Ehsani, H. Bagherinezhad, J. Redmon, R. Mottaghi, and A. Farhadi, "Who let the dogs out? Modeling dog behavior from visual data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4051–4060, doi: 10.1109/CVPR.2018.00426.
[8] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Scaled-YOLOv4: Scaling cross stage partial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13029–13038.
[9] F. Zhou, H. Zhao, and Z. Nie, "Safety helmet detection based on YOLOv5," in *Proc. IEEE Int. Conf. Power Electron., Comput. Appl.*, 2021, pp. 6–11, doi: 10.1109/ICPECA51329.2021.9362711.
[10] Q. Wang, H. Zhang, X. Hong, and Q. Zhou, "Small object detection based on modified FSSD and model compression," in *Proc. IEEE 6th Int. Conf. Signal Image Process.*, 2021, pp. 88–92, doi: 10.1109/ICSIP52628.2021.9688896.
[11] S. Zhou and J. Qiu, "RSSD: Object detection via attention regions in SSD detector," in *Proc. 2nd Int. Conf. Saf. Produce Informatization*, 2019, pp. 266–269, doi: 10.1109/IICSPI48186.2019.9095895.
[12] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 734–750.
[13] M. Li, H. Ge, and H. Wang, "IMG-CenterNet: An optimized algorithm based on CenterNet for pedestrian detection," in *Proc. IEEE 6th Inf. Technol. Mechatronics Eng. Conf.*, 2022, pp. 203–208, doi: 10.1109/ITOEC53115.2022.9734594.
[14] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.
[15] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.
[16] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards real-time multi-object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 107–122.
[17] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. Int. Conf. Image Process.*, 2016, pp. 3464–3468.
[18] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. Int. Conf. Image Process.*, 2017, pp. 3645–3649.
[19] Y. Zhang et al., "ByteTrack: Multi-object tracking by associating every detection box," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 1–21.
[20] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 474–490.
[21] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," *Int. J. Comput. Vis.*, vol. 129, pp. 3069–3087, 2021, doi: 10.1007/s11263-021-01513-4.
[22] Z. Liu et al., "Target detection and tracking algorithm based on improved mask RCNN and LMB," in *Proc. Int. Conf. Control, Autom. Inf. Sci.*, 2021, pp. 1037–1041, doi: 10.1109/ICCAIS52680.2021.9624519.
[23] G. Li et al., "Key supplement: Improving 3-D car detection with pseudo point cloud," *IEEE Sensors J.*, vol. 23, no. 16, pp. 18856–18866, Aug. 2023, doi: 10.1109/JSEN.2023.3292137.
[24] X. Wang, C. Fu, J. He, S. Wang, and J. Wang, "StrongFusionMOT: A multi-object tracking method based on LiDAR-Camera fusion," *IEEE Sensors J.*, vol. 23, no. 11, pp. 11241–11252, Jun. 2023, doi: 10.1109/JSEN.2022.3226490.
[25] Z. Peng, Z. Xiong, Y. Zhao, and L. Zhang, "3D objects detection and tracking using solid-state LiDAR and RGB camera," *IEEE Sensors J.*, vol. 23, no. 13, pp. 14795–14808, Jul. 2023, doi: 10.1109/JSEN.2023.3279500.
[26] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000, doi: 10.1109/34.888718.
[27] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, pp. 303–338, 2010, doi: 10.1007/s11263-009-0275-4.
[28] Y. Lin et al., "Improving person re-identification by attribute and identity learning," *Pattern Recognit.*, vol. 95, pp. 151–161, 2019, doi: 10.1016/j.patcog.2019.06.006.
[29] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*, doi: 10.48550/arXiv.1704.04861.
[30] W. Wang, X. Tan, P. Zhang, and X. Wang, "A CBAM based multiscale transformer fusion approach for remote sensing image change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sensors*, vol. 15, pp. 6817–6825, 2022, doi: 10.1109/JSTARS.2022.3198517.
[31] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3702–3712.
[32] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 658–666.

ZHENGTAO YU is currently working toward the M.Sc. degree in electronic information with the School of Automation, Southeast University, Nanjing, China.

His research interests include vision-based detection and tracking in port environment.

XUEQIN ZHENG received the M.Sc. degree in electronic information from the School of Automation, Southeast University, Nanjing, China, in 2023.

He is currently an Engineer with Meituan, Beijing, China. His research interests includes target detection and tracking.

JUN YANG (Fellow, IEEE) received the B.Sc. degree from the Department of Automatic Control, Northeastern University, Shenyang, China, in 2006, and the Ph.D. degree in automation from the School of Automation, Southeast University, Nanjing, China, in 2011.

He joined the Department of Aeronautical and Automotive Engineering, Loughborough University, Loughborough, U.K., in 2020, as a Senior Lecturer and is currently a Reader. His research interests include disturbance observer, motion control, visual servoing, nonlinear control, and autonomous systems.

Dr. Yang is a Fellow of the American Institute of Aeronautics and Astronautics. He was the recipient of EPSRC New Investigator Award. He serves as an Associate Editor or Technical Editor for IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, IEEE/ASME TRANSACTIONS ON MECHATRONICS, and IEEE OPEN JOURNAL OF INDUSTRIAL ELECTRONICS SOCIETY.

JINYA SU (Member, IEEE) received the Ph.D. degree in autonomous systems from Loughborough University, Loughborough, U.K., in 2016.

He held various positions including Postdoctor, Lecturer, and Senior Lecturer in the U.K. He is currently a Full Professor with the School of Automation, Southeast University, Nanjing, China. He has authored and coauthored more than 80 journal and conference papers. His research interests include perception and control, and their real-world applications in robotics systems.