Panamerican Mathematical Journal
ISSN: 1064-9735
Vol 35 No. 1s (2025)

Construction and Study of an Improved PP-Human Multi-Objective Tracking Method

Liang Ma[1,2], Vladimir Y. Mariano[1]

[1]National University, Manila 1008 Philippines

[2]Yantai Vocational College of Culture and Tourism, Yantai 264003 China

Corresponding author: Liang Ma (e-mail:maliang@yvcct.edu.cn).

Article History:

Received: *24-08-2024*

Revised: *27-09-2024*

Accepted: *14-10-2024*

Abstract:

Multiple objective tracking represents a crucial research focus within the realm of computer vision. It has a very broad application prospect in video surveillance, intelligent transportation, robot navigation and positioning and other fields.However, multi-target visual perception is affected by light, weather, occlusion and other factors, and is vulnerable to noise interference, faced with problems such as unstable image enhancement effect, target correlation accuracy and robustness.In this study, the PP-Human framework is applied to multi-target tracking. Through training and refinement of the PP-Human model, integration of a high-accuracy detector, enhancement of pedestrian re-identification (ReID) techniques, and optimization of the data association approach, the model's multi-object detection performance is enhanced, achieving efficient and precise multi-target tracking. Using the extended Market 150 dataset to comprehensively evaluate the performance of the proposed multi-objective tracking method, this paper improves the accuracy of multi-objective tracking by 5.0% and reaches 95.0% of MOTA through a series of optimization measures.To validate the efficacy and robustness of this approach, experimental evaluations were conducted. The refined PP-Human framework demonstrated strong performance in multi-target tracking tasks, establishing a reliable basis for practical applications such as pedestrian analysis, behavior recognition, and flow statistics.

Kewords: Multi-objective tracking,PP-Human,Pedestrian reidentification (ReID),Data association rules,Target detection accuracy,Market-1501 dataset.

I. INTRODUCTION

Multi-objective tracking, a vital research domain in the realm of computer vision, has emerged as a cornerstone for a myriad of practical application scenarios[1]. This complex task involves the continuous identification and localization of multiple objects in a sequence of images or videos, thereby enabling a deeper understanding of the visual world. Its significance is particularly pronounced in areas such as video surveillance, where it can enhance security by enabling the monitoring and tracking of multiple individuals or objects simultaneously[2].In the context of intelligent transportation, multi-object tracking plays a pivotal role in enabling autonomous vehicles to perceive and react to their surroundings, ensuring safety on the roads. In the field of robotics, it is instrumental for navigation and positioning, allowing robots to interact with and adapt to their environment effectively.

However, the path to achieving optimal performance in multi-objective tracking is fraught with numerous challenges. One significant obstacle is inter-target occlusion, where one object can block the view of another, leading to tracking failures. This is exacerbated in crowded scenes, where objects frequently intersect and overlap, making it difficult to maintain the association between targets and their tracks.Illumination changes, another critical factor, can significantly affect the appearance of objects, altering their color and contrast, and thereby disrupting the tracking process. The cluttered background, with its abundance of visual distractions, can also pose a significant challenge, as it can cause the tracker to lose focus on the target of interest.Moreover, the dynamic interactions between objects, such as collision or group behavior, can introduce additional complexities, necessitating the tracker to possess a high level of understanding and anticipation of the scene. Lastly, detection errors, which are inherent in most object detection systems, can lead to false positives or false negatives, further deteriorating the tracking performance.

Addressing these challenges requires the development of sophisticated algorithms and models that can effectively handle these complexities in real-time. Researchers have proposed various strategies, such as deep learning-based approaches, to improve the robustness and accuracy of multi-objective tracking systems[3]. Despite the progress made, this area of research remains a fertile ground for innovation and exploration, with the ultimate goal of achieving seamless and reliable multi-object tracking in complex and dynamic environments.

In recent years, on the basis of in-depth research, researchers actively innovate and put forward several multi-target tracking technology methods. Chunjiang Li et al., [4] proposed a multi-target tracking algorithm based on Transformer cross-spatial feature correlation, using the advantages of Transformer structure to extract global features and multi-head attention mechanism to improve the extraction ability of multi-target feature. Wang Wenyuan et al., [5] proposed a detection method based on multiple running targets based on ReInspect algorithm. By modifying the feature label information within the LSTM network, preprocessing the loss function, and employing confidence segmentation to match detection results post-tracking, we addressed the issues of redundant detection and target occlusion for the same object.Wu Mengqi, [6] using DeepSORT single hypothesis tracking matching framework for target trajectory prediction and preliminary matching, in the basic cascade matching stage increased the target block judgment stage, to fight the identity of occlusion number conversion problem, effectively reduce the identity number of the occlusion conversion, improve the accuracy of the tracking. Ning Qing et al., [7] proposed an improved Quasi-Dense multi-objective tracking algorithm. By integrating the concepts of attention mechanism and adaptability, the model enhances its capacity to detect and track targets that undergo significant scale variations. Fengwei Yu et al., [8] introduces POI, a high-performance detection and appearance feature-based method for multiple object tracking. POI leverages accurate detections and discriminative appearance features to achieve robust tracking results. Philipp Bergmann et al., [9] proposed a simple and efficient multi-object tracking framework that relies on basic components like a cost matrix and the Hungarian algorithm, demonstrating that complexity is not always necessary for effective tracking. Zhang et al.,[10] proposed an innovative online multi-target intelligent tracking algorithm, which uses deep, long and short-term memory network (DLSTM) to cope with the challenges of uncertain movement and observation noise. Their MTIT-DLSTM algorithm curates the tracking task in the time series by defining a target tuple set, effectively improving the accuracy and robustness of multi-target tracking. This research result provides new ideas and methods for the field of multi-target tracking. Jin Jianan et al.,[11] proposed an effective multi-target tracking algorithm, aiming to improve the safety and accuracy of unmanned vehicles in navigation and driving on campus. By improving the tracking algorithm, the research team has successfully realized the stable tracking of multiple forward targets in the campus environment, providing important technical support for the application of unmanned vehicles in complex scenarios.

Despite the presence of uncertain factors in multi-target tracking scenarios, such as occlusion among multiple targets, interference from similar targets, and blurring due to rapid motion, significant improvements in multi-target tracking accuracy have not yet been achieved. Building upon the enhanced

PP-Human multi-object tracking approach, this study introduces a more sophisticated feature extraction network, reinforces pedestrian re-identification (ReID), and optimizes the data association strategy. Experimental assessments confirm the efficacy and robustness of this method.

Compared with existing multi-target tracking methods, this method achieves better performance on enhanced Market 1501 datasets[12], providing an efficient and accurate solution for multi-target tracking visual applications.

The main work contents of this paper are summarized as follows:

1) The focus of this paper is to conduct a comprehensive optimization and refinement of the original model's detection and tracking components. The detection phase, being the initial step in the process, is crucial for accurately identifying and localizing objects. This aim to improve the algorithm's ability to discern and isolate objects from complex backgrounds, reducing false positives and improving overall precision. Concurrently, this paper are refining the tracking aspect to ensure seamless continuity in the identification of objects across frames, even in the presence of occlusions or rapid movements.

2) recognizing the need for enhanced adaptability in varying environments, we introduce a ReID (Re-Identification) appearance feature extraction module. This module is designed to capture and analyze the unique visual characteristics of each object, enabling the system to maintain consistency in identifying objects even when they reappear after being out of sight or in similar-looking objects. This addition is particularly beneficial in scenarios with multiple, closely interacting individuals, or in environments with changing lighting or camera angles.

3) This paper are not only upgrading the algorithm's ability to detect and recognize objects but also optimizing the data association strategy. This strategy plays a pivotal role in connecting the dots between frames, matching objects across time to generate accurate trajectories. By refining this strategy, This paper aim to minimize the likelihood of mismatches, thereby improving the overall tracking accuracy and the integrity of the generated trajectories.

Throughout this research, This paper will draw on extensive data sets, incorporating diverse scenarios and conditions, to rigorously test and validate the effectiveness of our proposed improvements. This paper also plan to utilize quantitative metrics, such as the Multiple Object Tracking Accuracy (MOTA) and the Identity F1 score, to objectively evaluate the performance of the enhanced ByteTrack algorithm.

This paper's central objective is to push the boundaries of the ByteTrack algorithm, equipping it with enhanced detection, tracking, and re-identification capabilities. These enhancements are expected to significantly improve its performance in real-world applications, ranging from surveillance systems to autonomous driving, and from crowd analysis to sports analytics.

II. MULTI-OBJECTIVE TRACKING ALGORITHM

Multi-target tracking is an important research hotspot in the field of computer vision. In order to improve the overall efficiency of multi-target tracking, the academia has proposed many research schemes, such as FairMOT, Deep-SORT algorithm.

Over the years, the computer vision community has dedicated significant efforts to enhance the overall efficiency and accuracy of multi-target tracking. One such breakthrough is the development of the FairMOT algorithm. FairMOT, which stands for Fairness, Accuracy, and Robustness in Multi-Object Tracking, introduces a novel approach that integrates object detection, tracking, and re-identification into a single, end-to-end deep learning model. By doing so, it significantly reduces computational complexity while maintaining high accuracy, thereby improving the overall efficiency of the tracking process.

Another notable research scheme that has gained traction is the Deep-SORT algorithm. Deep-SORT, an abbreviation for Deep Learning-based Sequential Tracking and Re-identification, builds upon the traditionalSORT algorithm by incorporating deep learning for feature extraction and re-identification. This innovation allows the algorithm to handle challenges such as occlusions, target appearance changes, and track switches more effectively, thereby enhancing the robustness of multi-target tracking.

Both FairMOT and Deep-SORT, while distinct in their methodologies, share a common goal of pushing the boundaries of multi-target tracking. They have not only set new benchmarks in performance but have also paved the way for further research and innovation in the field. For instance, researchers are now exploring ways to integrate these algorithms with other technologies, such as graph neural networks and reinforcement learning, to create even more sophisticated and adaptive tracking systems.

FairMOT Algorithm [13], using an end-to-end framework, simultaneously achieves object detection and object feature embedding learning. The loss function in FairMOT typically comprises three components: detection loss for targets, embedding learning loss, and center point loss, which can be expressed as in:

$$L = L\_det + \lambda 1 * L\_emb + \lambda 2 * L\_cen \tag{1}$$

where, L_det represents the target detection loss, L_emb denotes the embedding learning loss, L_cen is the center point loss, and $\lambda 1$ and $\lambda 2$ are weighting coefficients that serve to balance these three loss terms.

Deep-SORT Algorithm [14], is an algorithm for multi-object tracking that uses two cost functions in the data association process, the Mahalanobis distance and the cosine distance. Considering the two cost functions, Mahalanobis distance and cosine distance comprehensively, we can more accurately measure the similarity between targets and make target correlation, as in

$$\mathrm{C}_{i,j} = \lambda d^{(1)}(i,j) + (1-\lambda) d^{(2)}(i,j) \tag{2}$$

where, $C_{i,j}$ represents the similarity or distance between target i and target j, $d^{(1)}(i,j)$ represents the Mahalanobis distance, $d^{(2)}(i,j)$ represents the cosine distance, and λ is a weight parameter used to balance the importance of the two distance measures.

III. PP-HUMAN

PP-Human consolidates essential functions including target detection, tracking, and key point detection, aiming to enhance the precision and efficiency of portrait analysis. It has demonstrated excellent capabilities in the three key fields of human body attribute identification, behavior identification and traffic flow statistics, and provided strong technical support for intelligent security, smart city and other fields.

In terms of human body attribute identification, PP-Human can accurately identify complex characteristics such as gender, age and clothing color, which greatly enriches the dimension of data analysis. In terms of behavior recognition, it can analyze and understand various human movements in real time, such as running, jumping, waving, etc., providing the possibility for behavior analysis and early warning. In addition, its traffic flow statistics function can accurately calculate the number of pedestrians and vehicles in a specific area, which provides strong data support for traffic management and planning.

The flexibility of PP-Human is also worth mentioning. It can adapt to different types of input, such as single graph, single or multi-channel video, and meet the needs of various application scenarios. PP-Human can maintain stable performance and show strong environmental adaptability in complex background interference in light conditions and cross-lens scenes.

As shown in Figure 1, PP-Human's multi-target tracking model scheme uses advanced PP-YOLOE [15] as a detector, and its high precision and rapid response ability in target detection are widely recognized in the industry. While ByteTrack[16], as a tracker, ensures the accurate tracking of multiple targets in a complex environment with its powerful target association ability and robustness.

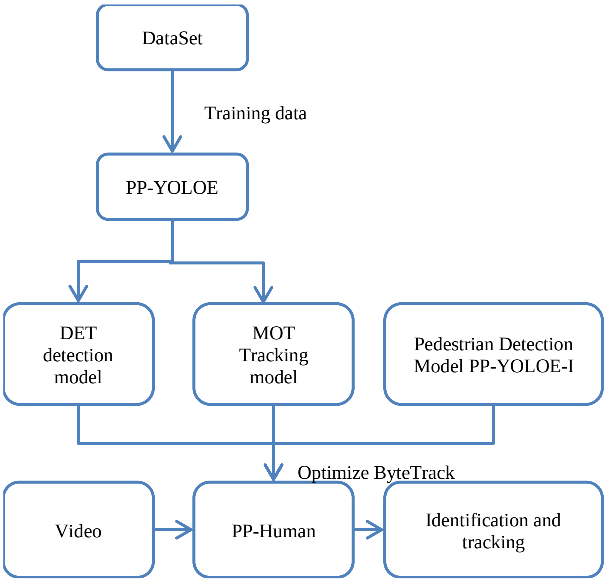


FIGURE 1. PP-Human Technology Roadmap.This display is the comprehensive application process of PP-Human technology. Starting with the initial "DataSet" and "training data", these rich data sources provide a solid foundation for subsequent pedestrian detection and tracking. Then, by using "PP-YOLOE" for pedestrian detection, we were able to accurately identify and locate the pedestrians in the video. Subsequently, the "MOT Tracking" module tracks the detected pedestrians and optimizes the tracks through "Optimize ByteTrack" to ensure the continuity and accuracy of the tracking. In this process, PP-Human technology has shown a strong performance advantage, enabling effectively responding to pedestrian tracking challenges in a variety of complex scenarios.

A. OPTIMIZE AND IMPROVE THE BYTETRACK ALGORITHM MODEL

In the model Backbone and Neck parts, replace the CSP 2 module at the end with the Transformer coding block TEB, using the residual connection between the sublayers, the attention mechanism ACmix module[17] is added; In the Head part of the model, a new branch uses the Coupled Head structure to predict small targets, as shown in Figure 2.

Drawing inspiration from the Transformer architecture's success in natural language processing, the TEB introduces a distinctive methodology for information processing in computer vision tasks. By leveraging self-attention mechanisms, the model is empowered to more effectively capture long-range dependencies and contextual information.To further enhance the flow of information, a residual connection is introduced between the sublayers of the TEB. This architectural tweak allows the model to learn more robust features while mitigating the risk of gradient vanishing.

In addition to the TEB integration, the Head part of the model undergoes a significant transformation as well. A new branch is appended to the existing structure, employing a Coupled Head structure specifically designed for the detection of small targets. This innovative design acknowledges the challenges associated with small object detection, where minute details and low contrast often lead to inferior performance. By dedicating a separate branch to small targets, the model can now focus on enhancing localization accuracy and feature representation for these intricate objects.

The architectural changes are visually depicted in Figure 2, where the transition from CSP 2 to TEB in the Backbone and Neck regions is clearly illustrated. The addition of the attention-driven ACmix module and the Coupled Head branch in the Head part showcases the intricate fusion of traditional and cutting-edge techniques in the pursuit of improved detection capabilities.

Through these modifications, the model aims to not only enhance its ability to discern complex scenes but also to excel in the challenging task of detecting small objects, thereby broadening its applicability in real-world scenarios such as autonomous driving, surveillance, and medical imaging. The integration of Transformer-based blocks and attention mechanisms, along with the tailored design for small target detection, positions this revised model at the forefront of advanced computer vision research.

1) SUBSTITUTION OF THE CSP 2_1 MODULE WITH THE TRANSFORMER ENCODING BLOCK (TEB)

Step 1, Identify and map the CSP 2_1 module within the Backbone and Neck sections of the model. The CSP 2_1 module comprises a sequence of convolutional, normalization, and activation function layers.

Step 2, Replace the CSP 2_1 module with the TEB. Each TEB is composed of two sub-layers: a multi-head attention layer and a Fully Connected Layer (MLP).

Step 3, In multi-head attention layer, the model will learn dependencies between different positions in the input sequence to capture information in different subspaces through multiple parallel attention heads.

Step 4, Fully Connected Layer (MLP) was used to further extract and integrate the features. Incorporating residual connections between the sublayers aids in mitigating the issue of gradient vanishing during the training of deep neural networks and facilitates faster model convergence.

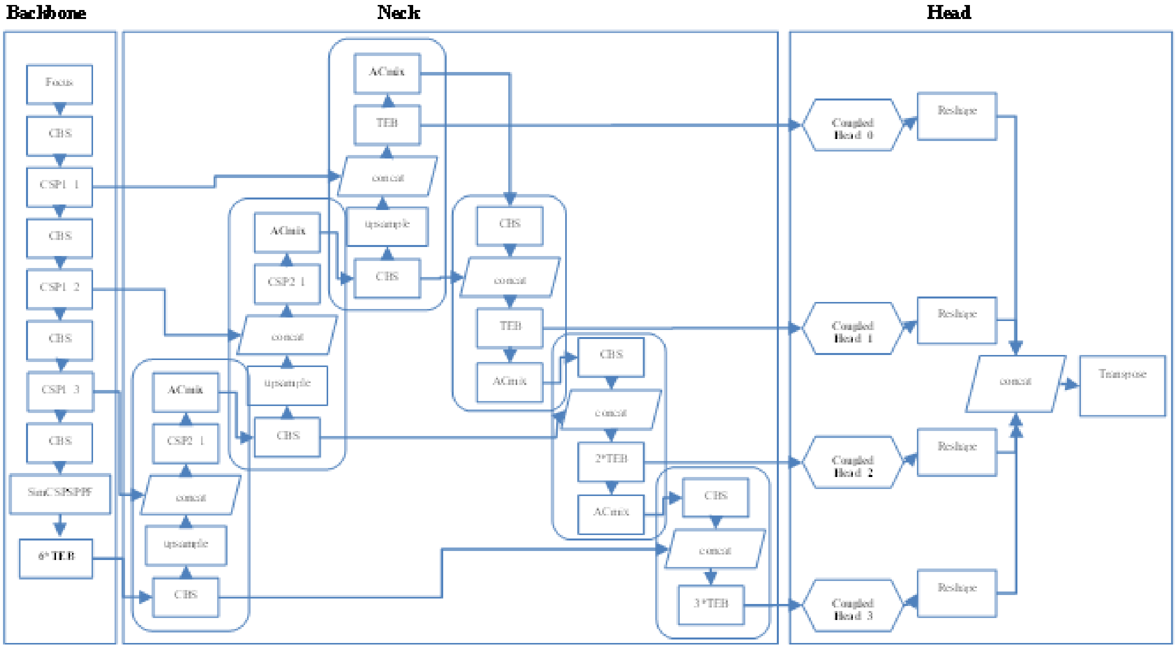


FIGURE 2. Schematic Representation of the Algorithm Model Enhancement in ByteTrack.Schematic diagram of ByteTrack algorithm model improvement. These improvements of ByteTrack algorithm model not only improve the accuracy of target detection, but also enhance the stability and robustness of the model. By introducing the Transformer coding block TEB, residual connection, and attention mechanism ACmix module, the model can better adapt to a variety of complex scenarios, providing more reliable technical support for practical applications.

2) INCORPORATE THE ACMIX ATTENTION MECHANISM MODULE IN THE NECK SECTION

In the intricate architecture of the neural network, the Neck section plays a pivotal role in information processing and fusion. This is the point where integrating the ACmix module can markedly boost the model's performance. The ACmix module, representing a novel attention mechanism, is engineered to allow the model to selectively focus on vital features while reducing the impact of irrelevant noise. The advent of attention mechanisms in deep learning has transformed how models interpret and handle information.They allow the model to dynamically allocate computational resources based on the relevance of features, thereby improving its ability to discern critical patterns. The ACmix module takes this a step further by introducing a novel approach to feature weighting.

The implementation of the ACmix module involves a two-fold process. Firstly, it computes the correlations between the feature maps generated by different layers or branches of the network. These correlations provide a measure of how interdependent or redundant the features are. Secondly, it reweights the feature maps based on these computed correlations. By doing so, it encourages the model to give more emphasis to those features that exhibit higher uniqueness and relevance to the target task, while downplaying features that are highly correlated or less informative.

3) NEW BRANCHES THAT ARE USED TO PREDICT SMALL TARGETS

A novel branch has been introduced in the Head section of the model, specifically for the underlying high-resolution feature map. This branch is solely devoted to predicting small targets, which is pivotal in enhancing the model's detection capabilities for such targets.For the new branches, the Coupled Head structure is used. Coupled Head Structure involves sharing layers and parameters to utilize feature information more efficiently.The crux of this advancement lies in the adoption of the Coupled Head Structure. This architectural design philosophy advocates for the strategic sharing of layers and parameters across different branches. By doing so, it fosters a more efficient utilization of the rich feature information that is extracted during the computational process. The rationale behind this lies in the understanding that shared knowledge can often provide a more comprehensive and detailed comprehension of the input data, especially when dealing with the complexities of small target detection.

4) MODIFY THE LOSS FUNCTION FOR PREDICTING THE FOREGROUND BACKGROUND BRANCH

In the original three Decoupled Head structures, the branch of the predicted foreground background is found, and its Loss function is changed from BCELoss to FocalLoss. FocalLoss is a loss function tailored for dense target detection tasks, which efficiently addresses the issue of category imbalance by directing the model's attention more towards samples that are challenging to classify.By using FocalLoss, the model has better performance when predicting the foreground background, especially when handling targets in complex background and occlusion situations.

B. RE-IDENTIFICATION OF PERSONS （ReID）

The ReID[18] appearance feature extraction module is added to the original model to adapt to the complex and changeable scene, and evaluates the similarity of the two frames before and after in the way of depth measurement learning to complete the matching task. The loss function Margin Loss[19] is introduced, as in

$$L = -\frac{1}{N}\sum_{i=1}^{N} 1g \frac{e^{s(cos(\theta_{y_i}+m))}}{e^{s(cos(\theta_{y_i}+m))} + \sum_{j=1x,j\neq y_i}^{n} e^{scon\theta_j}} \qquad (3)$$

where, N represents the batch size, n denotes the number of classes, θ_{y_i} signifies the angle between the weight W_{y_i} and feature x_i, m is the extra angle penalty bias term, and s represents the feature x_i Hscaled following l_2 regularization. Margin Loss formulates the loss function by examining the angular statistics between features and weights, enhancing the model's discriminative power and stabilizing the training process, all while incurring minimal additional computational cost.

First of all, ByteTrack incorporates the ReID (Re-Identification) technology into the original method of detecting association. The ReID technology is primarily used to identify and track specific pedestrians in video sequences. By extracting the appearance features of pedestrians and measuring their distance, ByteTrack demonstrates the capability to identify targets more precisely and recover them even after they have been occluded or temporarily vanished.The addition of this technology greatly improves the robustness and accuracy of ByteTrack in complex scenarios.

Secondly, ByteTrack combines the appearance model with the motion model Unscented Kalman Filter (UKF)[20]. UKF is a powerful state estimation technique that effectively handles the motion uncertainty of targets in dynamic scenes. By combining the appearance model with UKF, ByteTrack is able to more accurately predict the movement trajectory of the target, and maintain stable tracking during the target displacement or deformation.

In addition, For ReID feature extraction, ByteTrack used the unsupervised method Cluster Contrast ReID. This method is able to learn effective feature representation without annotated data, thus reducing the dependence on large amounts of annotated data. By extracting useful ReID features from large amounts of unannotated data, ByteTrack further improves the generalization ability and performance in practical applications.

And then, ByteTrack also employs two data augmentation techniques: Mosaic and MixUp. These two methods enhance the model's generalization capacity by randomly combining and blending the original data to produce a more varied set of training samples.

And that, After adding the ReID feature, the Multi-Object Tracking (MOT)[21] algorithm adopts the State-Dependent Exponential (SDE)[22] type algorithm. This algorithm is able to dynamically adjust the parameters of its motion model according to its state, in order to more precisely depict the motion pattern of the target. By combining with the ReID features, the SDE algorithm enables more stable and accurate multi-objective tracking in complex scenarios.

The last, The original ByteTrack model is developed based on the Pytorch architecture, enabling the model to make full use of Pytorch's powerful deep learning function and flexible programming interface. In order to achieve interconversion between different frameworks and meet online and real-time requirements, ByteTrack utilizes the Open Neural Network Exchange (ONNX)[23] format for model conversion. The model compression acceleration tools PocketFlow[24], TVM[25] and TensorRT[26] are used to reduce the computational complexity and memory footprint of the model while maintaining the performance, so as to meet the application scenarios with high real-time requirements.

C. ENHANCEMENT OF THE DATA ASSOCIATION STRATEGY

This paper introduces a combined data association approach that relies on spatiotemporal consistency and target interaction. This approach takes into account the target's position changes across successive frames, its speed, and its interaction with other targets, enabling more precise target matching and trajectory generation.Adopt the VJ-IMP[27] algorithm instead of the Hungarian algorithm; when setting the matching cost matrix, abandon the calculation method of simple weighted sum of appearance measure and motion measure, fuse the cosine distance by integrating spatial similarity and appearance similarity, creates a loss matrix combining motion and appearance information, as in

$$\ell_{i,j} = \begin{cases} 0.25 \cdot A_{i,j}^{cos} + 0.25 \cdot M_{i,j}^{iou}, & (A_{i,j}^{cos} < \varphi_{emb}) \wedge (M_{i,j}^{iou} < \varphi_{emb}) \\ 0.5 \cdot M_{i,j}^{cos}, & (A_{i,j}^{cos} < \varphi_{emb}) \wedge (M_{i,j}^{iou} > \varphi_{emb}) \\ 0.5 \cdot A_{i,j}^{cos}, & (A_{i,j}^{cos} < \varphi_{emb}) \wedge (M_{i,j}^{iou} < \varphi_{emb}) \\ 1, & otherwise \end{cases} \quad (4)$$

where, is the element at the i th row and j th column of the matching cost matrix; it signifies the IoU (Intersection over Union) distance between the i th predicted bounding box and the j th detected bounding box of the trajectory segment. This represents the motion loss. The cosine distance between the appearance description i of the trajectory segment and the new detection description j denotes the appearance loss;is close to the threshold and is set to 0.5, which is used to discard trajectory segments and detection pairs with unlikely matches;is the appearance threshold, also set to 0.5, and it serves to distinguish between positive and negative associations of the trajectory segment's appearance state and the detected embedding vector. For targets that exhibit both high appearance and IoU similarity, a lower penalty is assigned. If the appearance similarity exceeds the threshold but the IoU similarity does not, the appearance loss is taken as the determinant of the overall loss, and vice versa. In all other cases, the loss is set to 1. The elements of the matching cost matrix are updated according to this rule.

1) DATA PREPROCESSING

Feature extraction: For each detected target, its appearance features (color, texture, etc.) and motion features (speed, acceleration, etc.) are extracted.

Target representation: Create a feature vector containing appearance and motion features.

2) SIMILARITY CALCULATION

Appearance similarity: Use the cosine similarity or other similarity measures to calculate the appearance similarity between targets.

Motion similarity: evaluate the similarity by considering motion characteristics, namely the target's position, speed, and acceleration.

Spatial similarity: By evaluating the relative position and relationship between objects in space to judge the degree of similarity between them, the spatial distance is included in the similarity calculation, which improves the accuracy of data association.

3) BUILD A MATCHING COST MATRIX

For each pair of targets, consisting of the detection in the current frame and the trajectory from the previous frame, a matching cost is determined as a weighted combination of appearance similarity, motion similarity, and spatial similarity, as in

$$cost = W_1 * (1-S_1) + W_2 * (1-S_2) + W_3 * (1-S_3\ ) \quad (5)$$

where, S_1 is the appearance similarity, S_2 is the motion similarity, S_3 is the spatial similarity; W_1, W_2 and W_3 are the weight parameters.

IV. EXPERIMENTAL EVALUATION

This paper used mean Average Precision (mAP)[28] to measure detection performance. The mAP (mean Average Precision) represents the average value of AP (Average Precision) across various categories, as in

$$AP = \int_0^1 p(r)dr \quad (6)$$

$$mAP = \frac{1}{N}\sum_{I=1}^{N} AP_i \quad (7)$$

where, p and r denote precision and recall, respectively, while AP represents the area beneath the curve.

This paper uses Multiple Object Tracking Precision (MOTP)[29] and Multiple Object Tracking Accuracy (MOTA)[29] to measure the ability of continuous tracking targets;

MOTP measures the total estimated error between the successfully matched object and the assumed position across all frames. This error is determined by aggregating the position deviations of all matched pairs and then dividing by the overall number of matches. This metric serves as an indication of the tracker's capability to accurately estimate the target's position, as in

$$MOTP = \frac{\sum_{\mathrm{i},t} d_t^i}{\sum_t c_t} \quad (8)$$

where, C_t denotes the count of matched target positions and the hypothesized position in frame t ; while $d_{\mathrm{t}}^{\mathrm{i}}$ signifies the distance between the target in frame t and its corresponding matched target position, referred to as the matching error.

The MOTA (Multi-Object Tracking Accuracy) metric is capable of reflecting the precision regarding false alarms, under-reporting, and ID switching that occur during the tracking process;

$$MOTA = 1 - \frac{\sum_t m_t + fp_t + mme_t}{\sum_t g_t} \tag{9}$$

where, m_t signifies the count of targets that remain undetected at time t ; while fp_t denotes the number of instances where objects are incorrectly classified as targets. Furthermore; mme_t represents the count of targets that are incorrectly matched to other tracks. Consequently, MOTA (Multi-Object Tracking Accuracy) can be interpreted as a comprehensive evaluation metric that is derived from these three error ratios.

$$\overline{m} = \frac{\sum_t m_t}{\sum_t g_t} \tag{10}$$

where, Is a measure of the proportion of target objects that fail to be detected to all target objects that should be detected. This proportion reflects the omission of the tracking algorithm when detecting the target.

$$\overline{fp} = \frac{\sum_t m_t}{\sum_t g_t} \tag{11}$$

where, the ratio of false positives fp measures the frequency of falsely identifying non-target objects as targets, indicating the tracker's tendency to make mistaken detections.

$$\overline{mme} = \frac{\sum_t m_t}{\sum_t g_t} \tag{12}$$

where, the ratio of mismatches mme reflects how often the tracker assigns detected objects to the wrong trajectories.

V. EXPERIMENTAL DESIGN AND RESULTS

A. EXPERIMENTAL ENVIRONMENT

The configuration of this experiment is: 4*GPU Tesla V100; CPU 32 Cores; Video Mem 128GB; RAM 256GB; Disk 1TB.

B. ENHANCED DATASETS

The Market 1501 dataset[12], adding a self-built photo extension, was used as the training and test dataset. Building upon the original dataset, an additional 1000 individuals were incorporated, resulting in a total of 2501 pedestrians. Out of these, 1651 pedestrian annotations were allocated for training purposes, while 850 were reserved for the test set, as show in Figure 3.

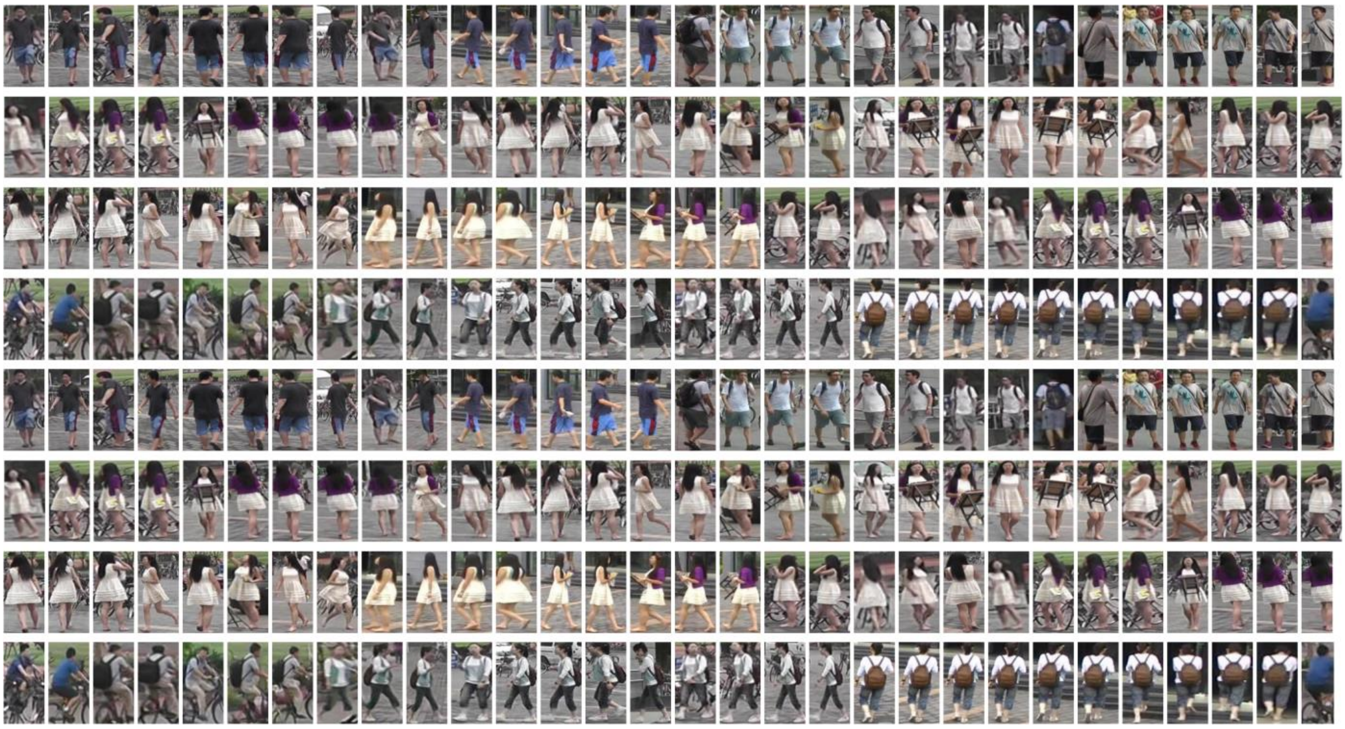


FIGURE 3. Add self-built photo extension Market 1501 dataset. Source of Market 1501 dataset: https://drive.google.com/file/d/0B8-rUzbwVRk0c054eEozWG9COHM/view?resourcekey=0-8nyl7K9_x37HlQm34MmrYQ

C. *ABLATION EXPERIMENT*

During the course of this research, a thorough series of ablation studies were conducted with meticulous attention to detail, aiming to investigate the individual impact of each refinement element on the algorithm's performance. These studies involved an exhaustive analysis of the implications of detection optimization, tracking optimization, the incorporation of a ReID-based appearance feature extraction module, as well as the augmentation of the data association strategy's impact on the overall system performance.By adopting this approach, we aim to gain a more precise comprehension of the enhancements brought about by each modification, and concurrently validate their adaptability and efficacy in intricate and dynamically changing environments.

1) THE INITIAL BYTETRACK ALGORITHM

The nascent ByteTrack algorithm underwent evaluation on a standard benchmark dataset, meticulously documenting its performance parameters. These included the evaluation metrics of MOTA (Multi-Object Tracking Accuracy), MOTP (Multi-Object Tracking Precision), False Negatives (FN), False Positives (FP), ID Switches (IDSW), and Frames Per Second (FPS).The algorithm demonstrated an initial MOTA score of 90.0%, coupled with an equivalent MOTP of 90.0%. It registered 1500 instances each for FN and FP, an IDSW count of 500, and a processing speed of 40.0 FPS. These figures establish a baseline for gauging the efficacy of subsequent optimization efforts, as illustrated in Table 1

TABLE I
THE INITIAL BYTETRACK ALGORITHM PERFORMANCE.

Method	MOTA/%	MOTP/%	FN	FP	IDSW	FPS
Original ByteTrack	90.0	90.0	1500	1500	500	40.0

2) DETECTION OPTIMIZATION AND TRACKING OPTIMIZATION

The enhanced ByteTrack algorithm was subjected to separate evaluations for its optimized detection and tracking components, each assessed in relation to the efficacy of the original ByteTrack algorithm. This comparative analysis, as illustrated in Table 2, highlights the disparities in detection precision and tracking stability between the base and the refined algorithm, thus highlighting the efficacy of the suggested enhancements.

TABLE II THE PERFORMANCE COMPARISON OF THE IMPROVED ALGORITHM.

Method	MOTA/%	MOTP/%	FN	FP	IDSW	FPS
Original ByteTrack	90.0	90.0	1500	1500	500	40.0
Detection Optimization	92.0	91.0	1300	1400	480	39.0
Tracking Optimization	91.5	91.5	1350	1350	450	39.5
Detection & Tracking Opt.	93.5	92.0	1150	1200	400	38.0

Result display. Detection Optimization MOTA Increased by 2%, MOTP increased by 1%, and the number of FN and FP has decreased. Tracking Optimization Similar results have been achieved. When the detection and tracking parts were optimized simultaneously, the performance improvement was more obvious, with MOTA increasing to 93.5%, and the number of missed and missed tests decreased significantly, indicating that the optimized algorithm could detect and track targets more accurately.

3) ADD THE REID MODULE

On the basis of detection and tracking optimization, ReID module is added to enhance the tracking ability of the algorithm. Compared with the algorithm performance before and after the addition of ReID module, the adaptability improvement in complex and changeable scenarios is observed, as show in Table 3

TABLE III COMPARISON OF THE ALGORITHM PERFORMANCE BEFORE AND AFTER ADDING THE REID MODULE.

Method	MOTA/%	MOTP/%	FN	FP	IDSW	FPS
Before ReID	93.5	92.0	1150	1200	400	38.0
After ReID	94.5	92.5	1120	1100	250	37.5

The results indicate that MOTA has improved by an additional 1.0% to reach 94.5%, while MOTP has also increased by 0.5% to 92.5%. Simultaneously, there was a reduction of 28 in the number of missed detections and a decrease of 100 in the number of false positives. More notably, the number of identity switches decreased significantly by 150, suggesting that the ReID module is highly effective in preserving the continuity of target tracking. However, the integration of the ReID module also introduces a minor performance trade-off, causing a slight reduction in the frames processed per second to 37.5 frames..

4) DATA ASSOCIATION POLICY OPTIMIZATION

On the basis of integrating the ReID module and optimizing the detection and tracking, the data association strategy is further optimized to evaluate the improvement of the target matching accuracy of the optimized data association strategy compared with the original strategy, as shown in Table 4

TABLE IV THE ECTS OF DATA ASSOCIATION STRATEGY OPTIMIZATION.

Method	MOTA/%	MOTP/%	FN	FP	IDSW	FPS
Before Association Opt.	94.5	92.5	1120	1100	250	37.5
After Association Opt.	95.0	92.9	1107	1026	180	37.0

The results reveal that the refined data association strategy led to a further increase in MOTA by 0.5%, bringing it to 95.0%, and MOTP to 92.9%. Additionally, there was a further reduction of 15 in the number of missed detections and a decrease of 74 in the number of false positives. The number of identity switches was also significantly reduced by 70. These findings suggest that the optimized data association strategy enhances the precision of target matching and the consistency of trajectory formation. Nevertheless, this improvement also entails a minor performance cost, causing the frame rate to drop to 37.0 frames per second.

The ablation experiments progressively demonstrated the enhancements in performance achieved through detection optimization, tracking optimization, the incorporation of the ReID module, and the refinement of the data association strategy. The ultimate experimental outcomes revealed that these improvement strategies can notably boost the precision and continuity of multi-target tracking, while still preserving a level of real-time performance.

D. COMPREHENSIVE EXPERIMENTAL COMPARISON

A comparison was made between the performances of JDE, Deep-SORT, FairMOT, and MOT algorithms, assessing their tracking accuracy, precision, MOTP, FN, FP, IDSW, and real-time capabilities. The experimental outcomes indicate that the proposed method exhibits notable advantages in crucial metrics like MOTA and MOTP, delivering high tracking precision and real-time performance, as show in Table 5

TABLE V PERFORMANCE COMPARISON.

Method	MOTA/%	MOTP/%	FN	FP	IDSW	FPS
JDE[30]	90.5	91.5	1270	1230	420	32.5
Deep-ORT[14]	87.5	89.5	1509	1000	620	47.0
FairMOT [13]	92.6	93.0	1230	1180	390	36.2
Ours	95.0	92.9	1107	1026	370	37.0

MOTA Analysis: The experimental results reveal that our algorithm has attained the highest score of 95.0% in the MOTA metric, notably surpassing the other algorithms. This underscores the significant advantages of our algorithm in tracking accuracy and its effectiveness in minimizing errors such as false alarms, missed detections, and identity switches.

MOTP Analysis: Regarding the MOTP metric, our algorithm achieved a score of 92.9%, which is marginally lower than FairMOT's 93.0% but higher than JDE and Deep-SORT. This demonstrates that our algorithm performs well in predicting the accuracy of target positions, although there is still potential for improvement.

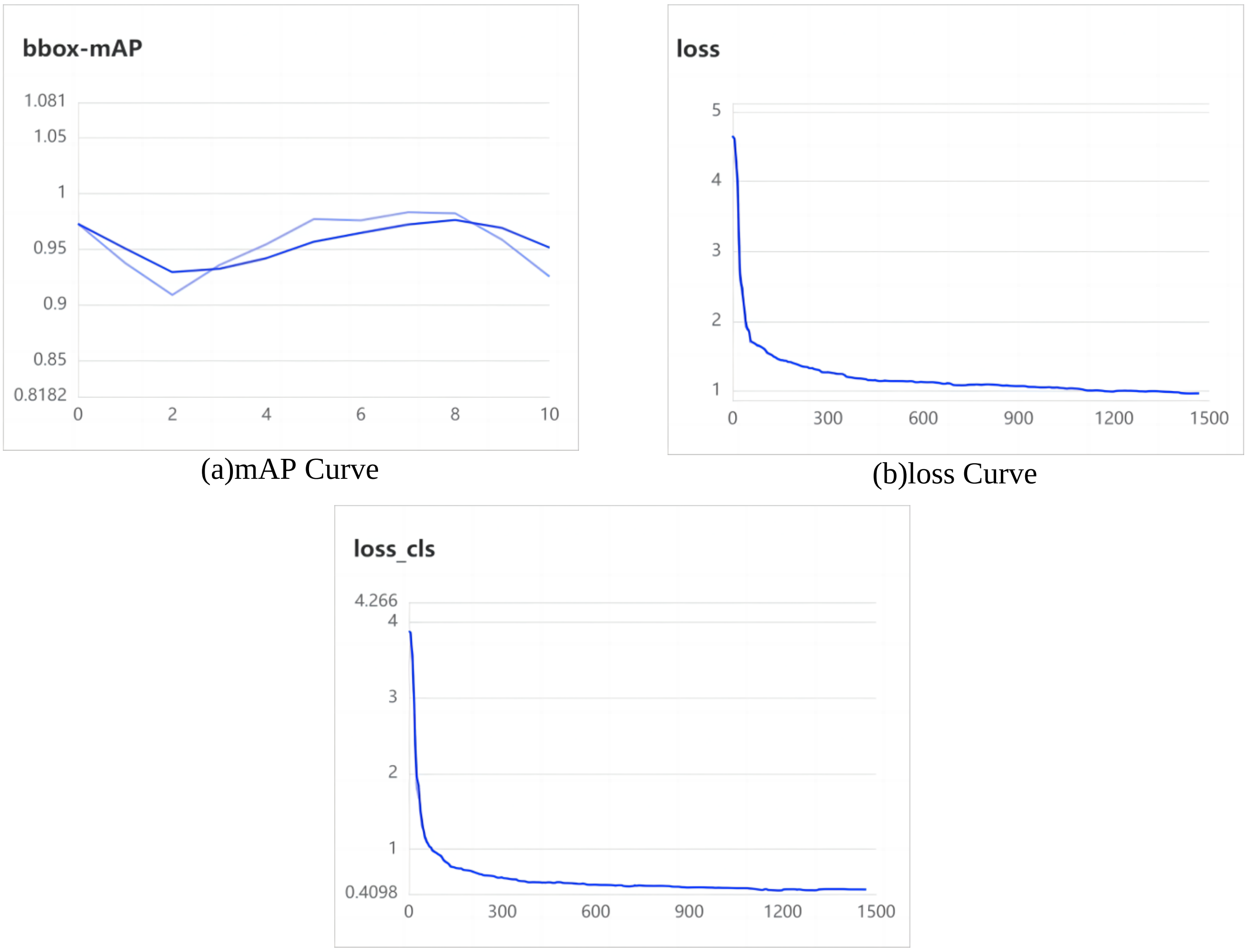
FN and FP Analysis: Our algorithm outperforms the others in both FN and FP metrics, suggesting that it has achieved favorable results in reducing false negatives and false positives. This is primarily due to the optimization strategies employed during the target detection and tracking processes.

IDSW Analysis: In the IDSW metric, our algorithm scored 370, which is lower than the scores of the other algorithms. This indicates that our algorithm has some advantages in mitigating identity switching errors, although further optimization is required to further reduce their occurrence.

FPS Analysis: With respect to real-time performance, our algorithm's FPS is 37.0, which is slightly higher than FairMOT and JDE but lower than Deep-SORT. This highlights the need for our algorithm to continue focusing on optimizing real-time performance while sustaining high tracking accuracy.

E. VISUALIZATION OF CORE INDICATORS

To enhance our understanding of the model's training performance, we have incorporated visual representations for key performance indicators during the training process. This encompasses the mAP (mean average precision) trajectory, the overall loss curve, and the classification loss (loss_cls) curve. By utilizing these visual graphs, we can efficiently observe the model's learning process and make informed adjustments to our training approach, as show in Figure 4.



(a)mAP Curve

(b)loss Curve

(c)loss_cls Curve

FIGURE 4. Visualization curves depicting various metrics related to model training, including (a) bounding box mean Average Precision (bbox-mAP), (b) loss, and (c) loss with classification score (loss_cs).

VI. EXPERIMENTAL RESULTS SHOW

By utilizing tracking demonstrations with images from the custom-created photo-enhanced Market 1501 dataset, the system successfully accomplished accurate identification and continuous tracking of numerous targets, As show in Figure 5.

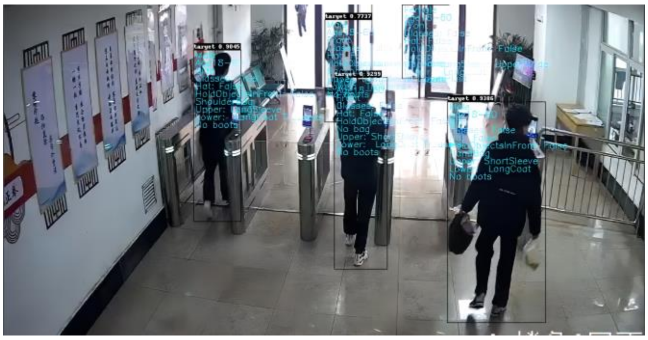


FIGURE 5. Illustrates the experimental results of multi-target tracking in a typical complex scenario. This scenario features the presence of multiple targets and significant light contrast. The proposed method, as detailed in this paper, demonstrates in accurately identifying and tracking these targets. Its advantages in multi-target tracking are evident, as it not only identifies and consistently tracks the target objects but also provides stable and reliable information on target positions and motion trajectories. even in the most intricate settings, the method responds effectively, exhibiting superior performance and outcomes.

VII. CONCLUSION

In the field of multi-object tracking, a significant challenge is the precise and consistent identification of multiple targets within dynamic environments. This paper addresses this challenge by introducing a novel multi-object tracking method that builds upon the enhanced PP-Human model. By refining the ByteTrack algorithm, integrating a ReID appearance feature extraction module, and optimizing the data association strategy, this innovative approach improves the accuracy and robustness of multi-object tracking.

The ByteTrack algorithm, a cornerstone in the field, has been carefully optimized to bolster its ability to differentiate and track numerous objects concurrently. Through the integration of the ReID feature extraction module, our method excels at extracting more distinctive features, enhancing object identification accuracy even in situations involving occlusions or similar appearances. Additionally, the enhanced data association strategy reduces the likelihood of incorrect associations, thereby fortifying the robustness of the tracking process. Experimental assessments conducted on the Market 1501 dataset, complemented by our own compiled images, have underscored the notable superiority of our proposed method. Significant improvements in key performance metrics, namely Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP), affirm the efficacy and superiority of our approach.

Despite these promising results, there is still room for further refinement. In complex scenarios, such as crowded spaces or environments with rapidly changing lighting conditions, the tracking performance could be enhanced by incorporating more contextual information or leveraging more advanced feature extraction networks, such as Transformers or ConvNeXt. Moreover, real-time performance, a crucial aspect for practical applications, can be optimized by streamlining the network architecture or adopting more efficient computation strategies, such as parallel processing or quantization.Another promising avenue for future exploration lies in the integration of multi-modal information. By synergistically combining visual cues with auditory, textual, or other sensory data, the tracking system could potentially achieve a more comprehensive understanding of the environment, thereby enhancing its performance in complex and diverse scenarios.

As deep learning technology continues to evolve at an unprecedented pace, it presents boundless opportunities for advancing multi-object tracking algorithms. By leveraging advancements in areas such as attention mechanisms, graph neural networks, or even meta-learning, we can anticipate the development of more sophisticated and adaptable tracking systems capable of tackling ever more complex tasks.

In conclusion, our study has made a significant stride in enhancing multi-object tracking, but the quest for perfection in this domain remains an open and exciting challenge. With continued research and innovation, we envision a future where tracking systems can seamlessly navigate and understand the intricate dynamics of the real world.

References

[1] YANG Jie,CAO Xiaoshu,YAO Jun,KANG Zhewen,CHANG Jianxia & WANG Yimin.(2024).Geographical big data and data mining:A new opportunity for "water-energy-food" nexus analysis. Journal of Geographical Sciences(02),203-228. doi:CNKI:SUN:ZGDE.0.2024-02-001.

[2] Frank Ngeni,Judith Mwakalonge & Saidi Siuhi.(2024).Solving traffic data occlusion problems in computer vision algorithms using DeepSORT and quantum computing. Journal of Traffic and Transportation Engineering(English Edition)(01),1-15. doi:CNKI:SUN:JTTE.0.2024-01-001.

[3] Bodi MA,Zhenbao LIU,Feihong JIANG,Wen ZHAO,Qingqing DANG,Xiao WANG... & Lina WANG.(2023).Reinforcement learning based UAV formation control in GPS-denied environment. Chinese Journal of Aeronautics(11),281-296. doi:CNKI:SUN:HKXS.0.2023-11-017.

[4] Chunjiang Li, Guangzhu Chen, Rongsong Gou, Zaizuo Tang,Detector–tracker integration framework and attention mechanism for multi–object tracking,Neurocomputing,Volume 464,2021,Pages 450-461,ISSN 0925-2312,DOI:10.1016/j.neucom.2021.08.107.

[5] Wang Wenyuan, Jin Hong, Song Wenjing & Wang Yiwei. (2022). Multi-target tracking based on the ReInspect algorithm. Journal of Metrology (04), 470-474. doi:CNKI:SUN:JLXB.0.2022-04-006.

[6] Wu Mengqi & Liu Junqing. (2020). Problem study of a single-hypothesis multi-target tracking method based on DeepSORT. Information and communication (11), 40-42. doi:CNKI:SUN:HBYD.0.2020-11-013.

[7] Ning Qing, Bao Hong, Pan Weiguo, et al. Improved Quasi-Dense multi-objective tracking algorithm [J]. Sensors and Microsystems, 2023,42(6):124-128.DOI:10.13873/J.1000-9787 (2023) 06-0124-05.

[8] FENGWEI YU, WENBO LI, QUANQUAN LI, et al. POI: Multiple Object Tracking with High Performance Detection and Appearance Feature[C]. //Computer vision -- ECCV 2016 Workshops : Part II /Springer, 2016:36-42.

[9] PHILIPP BERGMANN, TIM MEINHARDT, LAURA LEAL-TAIXÉ. Tracking Without Bells and Whistles[C]. //2019 IEEE/CVF International Conference on Computer Vision: IEEE/CVF International Conference on Computer Vision (ICCV 2019), 27 Oct.-2 Nov. 2019, Seoul, Korea.Institute of Electrical and Electronics Engineers, 2019:941-951.

[10] Zhang, Yongquan, Zhenyun Shi, Hongbing Ji, et al. "Online multi-target intelligent tracking using a deep long-short term memory network." Chinese Journal of Aeronautics (English Edition), 2023, 36(9): 313-329. DOI: 10.1016/j.cja.2023.02.006.

[11] Jin Jia Nan, Wu Yanfeng, Shi Zhenning, et al. Research on multi-target tracking algorithm in front of unmanned vehicles for complex campus environment [J]. Modern Manufacturing Engineering, 2023 (8): 44-52. DOI:10.16731/j.cnki.1671-3133.2023.08.006.

[12] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang and Q. Tian, "Scalable Person Re-identification: A Benchmark," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 1116-1124, doi: 10.1109/ICCV.2015.133.

[13] Zhang, Y., Wang, C., Wang, X., Zeng, W., & Liu, W. (2020). FairMOT: On the Fairness of Detection and Re-identification in Multiple Object Tracking. International Journal of Computer Vision, 129, 3069 - 3087.

[14] Wojke, N., Bewley, A., & Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. 2017 IEEE International Conference on Image Processing (ICIP), 3645-3649.

[15] Xu, S., Wang, X., Lv, W., Chang, Q., Cui, C., Deng, K., ... & Lai, B. (2022). PP-YOLOE: An evolved version of YOLO. arXiv preprint arXiv:2203.16250.

[16] Zhang, Y., Sun, P., Jiang, Y., Yu, D., Yuan, Z., Luo, P., Liu, W., & Wang, X. (2021). ByteTrack: Multi-Object Tracking by Associating Every Detection Box. European Conference on Computer Vision.

[17] Pan, X., Ge, C., Lu, R., Song, S., Chen, G., Huang, Z., & Huang, G. (2021). On the Integration of Self-Attention and Convolution. arXiv preprint arXiv:2111.14556

[18] MING, ZHANGQIANG, ZHU, MIN, WANG, XIANGKUN, et al. Deep learning-based person re-identification methods: A survey and outlook of recent works[J]. 2022,119(Mar.):104394.1-104394.18. DOI:10.1016/j.imavis.2022.104394.

[19] DENG J K， GUO J， XUE N N， et al. ArcFace：additive angular margin loss for deep face recognition［C］// Proceedings of the 2019 IEEE/CVF onference on Computer Vision and Pattern Recognition. Piscataway：IEEE， 2019：4685-4694.

[20] ERIC A. WAN, RUDOLPH VAN DER MERWE. The unscented kalman filter for nonlinear estimation[C]. //2000 Adaptive Systems for Signal Processing Communications, and Control Symposium (AS-SPCC 2000). 2000:153-158.

[21] Saleh, F. , Aliakbarian, S. , Rezatofighi, H. , Salzmann, M. , & Gould, S. . (2020). Probabilistic tracklet scoring and inpainting for multiple object tracking.

[22] Thomas Rückstie, Felder, M. , & Jürgen Schmidhuber. (2008). State-Dependent Exploration for Policy Gradient Methods. Machine Learning and Knowledge Discovery in Databases, European Conference, ECML/PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part II. DBLP.

[23] BRADDOCK GASKILL. ONNX: the Open Neural Network Exchange Format[J]. Linux journal,2018(Apr. TN.285):157-161.

[24] Jiang, Y., Zhang, G., You, J. et al. PocketFlow is a data-and-knowledge-driven structure-based molecular generative model. Nat Mach Intell 6, 326–337 (2024). https://doi.org/10.1038/s42256-024-00808-8

[25] Chen T , Moreau T , Jiang Z ,et al.TVM: An Automated End-to-End Optimizing Compiler for Deep Learning[J]. 2018.DOI:10.48550/arXiv.1802.04799.

[26] Zhou Lijun, Liu Yu, Bai Lu, et al. Deep learning inference using TensorRT[J]. Journal of Applied Optics, 2020, 41(2): 337-341. DOI: 10.5768/JAO202041.0202007.

[27] Chen, Z. , Tian, M. , Bo, Y. , & Ling, X. . (2019). Infrared small target detection and tracking algorithm based on new closed-loop control particle filter. Proceedings of the Institution of Mechanical Engineers, 233(4), 1435-1456.

[28] LI, KEHUANG, HUANG, ZHEN, CHENG, YOU-CHI, et al. A maximal figure-of-merit learning approach to maximizing mean average precision with deep neural network based classifiers[C]. //2014 IEEE International Conference on Acoustics, Speech and Signal Processing: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4-9 May 2014, Florence, Italy.Institute of Electrical and Electronics Engineers, 2014:4503-4507.

[29] Bernardin, K. & Stiefelhagen, R. Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. Image and Video Processing, 2008(1):1-10, 2008.

[30] ZHONGDAO WANG, LIANG ZHENG, YIXUAN LIU, et al. Towards Real-Time Multi-Object Tracking[C]. //Computer vision -- ECCV 2020 : Part XI /Springer, 2020:107-122.

AVAILABILITY OF DATA AND MATERIALS

The data supporting the findings of the article is available within the article.

FUNDING

This research did not receive any financial support.

COMPETING INTEREST STATEMENT

The authors declared that there is no competing interest.

Liang Ma, PhD candidate, National University of the Philippines, currently working with Yantai Vocational College of Culture and Tourism, China. Research interests include computer vision, digital image processing, and machine learning.

Vladimir Y. Mariano, received the B.S. degree in statistics and the M.S. degree in computer science from the University of the Philippines Los Banos, and the Ph.D. degree in computer science and engineering from The Pennsylvania State University. His research interests include computer vision, digital image processing, and machine learning.