CIT. Journal of Computing and Information Technology, Vol. 34, No. 2, June 2026, 99–116
doi: 10.20532/cit.2026.1006371

Multi-Object Tracking in Crowded Scenes by Integrating Image Motion Compensation Mechanism and Improved ByteTrack Algorithm

Huanting Feng

School of Artificial Intelligence and Big Data, Chengdu College of Arts and Sciences, Chengdu, China

This study proposes a multi-object tracking method in dense crowd scenes that integrates the image motion compensation mechanism and improved ByteTrack algorithm. This method aims to improve detection accuracy, enhance the model's perception of small-scale targets and occluded targets, and significantly alleviate the performance bottleneck of multi-object tracking in dense scenes. Subsequently, it performs secondary matching on the low-scoring frames corrected by camera offset compensation and introduces Kalman filtering with adaptive noise adjustment to improve the robustness of trajectory prediction. Finally, through the comprehensive evaluation module, the tracking indicators are output. The results show that the F1 score in different scene complexities is overall greater than 85%. When the intersection ratio threshold is 0.5, the average accuracy is 95.2%. In practical application tests, the average recall rate of the research method exceeds 85% across different detection numbers for each picture. Its overall processing delay time under different population densities is less than 100 ms, and its overall missed detection rate in different target size intervals is less than 10%. Experimental results indicate that the proposed method exhibits competent small target detection capability and provides stable visual analysis under the tested conditions.

ACM CCS (2012) Classification: Computing methodologies → Machine learning → Machine learning algorithms → Spectral methods

Keywords: crowded scenes, multi-object tracking method, image motion compensation, improved bytetrack algorithm, camera offset compensation

1. Introduction

With the rapid development of computer vision technology, Multi-Object Tracking (MOT) in dense scenes has become a cutting-edge topic that has attracted much attention in artificial intelligence. In the real world, the monitoring and analysis of densely populated places such as stations, shopping malls, and stadiums, as well as the monitoring of vehicles in road traffic, require algorithms that can accurately identify, locate, and track a large number of targets at the same time [1]. In crowded environments, challenges include severe occlusion between targets, small-scale target presentation, and high appearance similarity among individuals. These factors together constitute the main difficulty of MOT in dense scenes. At the same time, complex environmental factors such as lighting changes, weather conditions, and camera movements in real scenes further increase the complexity of the problem [2-3]. The ByteTrack algorithm can use low-confidence detection frames to associate targets through two matches, thereby effectively reducing missed target detection and trajectory interruption due to occlusion, and significantly reducing the number of Identity Document (ID) switching times [4]. The Image Motion Compensation (IMC) mechanism can make the motion model focus more on the target's own motion, improve the prediction effect under severe camera motion or low frame rate, and can be flexibly integrated as a pre-processing module to work in conjunction with other improvement strategies [5]. Therefore, to address occlusion and small target detection in dense scenes and improve the discrimination ability of similar targets, this study innovatively proposes an MOT method

for crowded scenes that integrates IMC and the Improved ByteTrack Algorithm (IBTA). This study deeply integrates IMC and trajectory prediction to improve the robustness and accuracy of Multi-Target Tracking (MTT) in crowded scenes. It first corrects global and local displacements caused by camera motion through multi-level IMC to improve detection stability under dynamic backgrounds. Subsequently, in the ByteTrack framework, Camera Offset Compensation (COC), Kalman filter adaptive noise adjustment, and Graph Neural Network (GNN) matching are introduced to enhance trajectory coherence and occlusion processing capabilities. Finally, through a dedicated evaluation module, optimized tracking indicators are output to achieve more stable and accurate visual analysis. It is expected that the research method can meet the design needs of MOT research in crowded scenes and provide technical support for the realization of intelligent and reliable visual perception systems.

2. Literature review

With the rise of the metaverse concept, accurate real-time MOT and tracking technology will become a bridge connecting the physical world and the digital world. Therefore, the research on MOT methods in dense crowd scenes is of great significance. Aiming at the problems of low contrast, small intra-class differences, and many false negatives in infrared target detection and tracking in dense urban traffic, Zha *et al.* proposed a comprehensive infrared MTT method for urban traffic. This study introduced a variety of image enhancement technologies, backbone network replacement, and infrared feature extraction modules, with good tracking accuracy [6]. Liang *et al.* proposed a deep learning method that integrates a spatiotemporal attention mechanism to address that ship detection is susceptible to interference and small target identification is difficult in complex ocean environments. This method could effectively improve the detection accuracy and robustness in multi-target and occlusion scenarios [7]. Scarrica *et al.* proposed a hybrid strategy that integrates optical flow and traditional deep learning architecture to address high resource consumption and difficulty in real-time application of deep learning models in MTT. This method could significantly improve processing efficiency while maintaining high accuracy [8]. Rafique *et al.* proposed a method based on maximum entropy scaling, superpixel segmentation, and multi-feature fusion to address the insufficient MOT accuracy in scene understanding. This method had the advantages of efficient calculation and accurate recognition [9]. Shevchenko *et al.* proposed a cognitive artificial intelligence method that integrates a spatiotemporal attention mechanism to address the accuracy and real-time issues of MOT in video analysis. This method had the advantages of high precision and real-time performance [10].

Many industry scholars have conducted in-depth research and applications on IMC and the IBTA. Qingxia *et al.* proposed a state vector enhanced ByteTrack method to address the identity switching and trajectory prediction errors in MTT of newborn lambs. By integrating confidence information and trajectory reconstruction, it optimized occlusion processing and trajectory accuracy, and had stable and reliable tracking performance [11]. Jang *et al.* proposed a deep learning method that integrates ByteTrack tracking and trajectory similarity analysis to address privacy protection and anomaly detection in vehicle monitoring. This method could effectively analyze traffic flow anomalies while ensuring high detection accuracy and tracking stability [12]. Jia *et al.* proposed a method to integrate improved YOLOv8 and optimized ByteTrack to address that fish target detection and tracking in complex underwater environments are susceptible to interference from occlusion and non-linear motion. This method could effectively improve detection accuracy and tracking stability while reducing the model size [13]. Wang *et al.* proposed a lightweight dual-branch network based on an improved atmospheric scattering model to address poor dehazing effect in real scenes. It designed a network structure that jointly estimates the transformation map and the compensation map, which had the advantages of a lightweight model and strong pertinence [14]. Klein *et al.* proposed a research method using the Fermi level as a universal descriptor to address the difficulty in selective control of the compensation mechanism in chemical substitution. This method could effectively predict and regulate specific compensation mechanisms [15].

To sum up, the existing research has a good effect on the research of the MOT method in dense crowd scenes. However, it still lacks an effective mechanism for dealing with small target detection and occlusion processing. The ByteTrack algorithm can be integrated into a variety of detectors as a plug-and-play module, allowing researchers to introduce further improvements such as attention mechanisms, better motion models, or lightweight designs. IMC can make the positional relationship between low-scoring frames and predicted trajectories more realistic in crowded scenes, thereby improving the success rate of the second association and reducing missed detections and trajectory fragmentation caused by occlusion. Therefore, this study proposes an MOT method integrating IMC and IBTA for crowded scenes, aiming to provide theoretical support for technological innovation in fields such as social public security, urban management, and business intelligence.

3. Research methodology

3.1. MOT method for crowded scenes based on IBTA

Research on MOT in crowded scenes has broad application prospects in many fields such as intelligent monitoring, autonomous driving, smart city management, and large-scale event security [16]. However, most current target detection and recognition algorithms have inherent contradictions between handling fine-grained classification and precise positioning. In an environment where target categories are highly similar, the classification task and the positioning task restrict each other, resulting in a decline in overall recognition performance [17]. The ByteTrack algorithm can reappear even after the target is briefly occluded, and can re-associate it to the original trajectory through the matching of low-scoring detection frames and predicted trajectories. Moreover, the introduction of the COC module or improved state vector can also make motion prediction more realistic [18]. This study's detection boxes are divided into high- and low-score sets based on confidence thresholds, as shown in Equation (1).

$$\begin{cases} D_h = \{d_i \mid s(d_i) > \tau_h\} \\ D_l = \{d_i \mid \tau_l \le s(d_i) \le \tau_h\} \end{cases} \tag{1}$$

In Equation (1), d_i is the i-th detection frame. $s(d_i)$ is the confidence score of the detection frame. τ_h and τ_l are high and low confidence thresholds. D_h and D_l are sets of high and low score detection frames. Next, this study estimates the position and error covariance of the target in the next frame in the crowded scene by describing the prediction step in ByteTrack [19], as shown in Equation (2).

$$\begin{cases} \hat{x}_{k|k-1} = K\hat{x}_{k-1|k-1} \\ P_{k|k-1} = KP_{k-1|k-1}K^T + Q \end{cases} \tag{2}$$

In Equation (2), $\hat{x}_{k|k-1}$ is the prior state estimate of time k. $\hat{x}_{k-1|k-1}$ is the posterior state estimate at time $k-1$. $P_{k|k-1}$ and $P_{k-1|k-1}$ are the prior and posterior error covariance matrices. K is the state transition matrix. Q is the process noise covariance. Subsequently, to implement improved matching rules, this study uses GNN to calculate the matching score of detection frames and trajectories in crowded scenes [20], as shown in Equation (3).

$$\begin{cases} S_{\text{match}}(d_i, t_j) = \text{GNN}\left(\boldsymbol{\psi}\left(d_i\right), \boldsymbol{\psi}\left(t_j\right), M\right) \\ A = \text{Softmax}\left(\dfrac{\boldsymbol{\psi}\left(d_i\right) \cdot \boldsymbol{\psi}\left(t_j\right)^T}{\sqrt{d_o}}\right) \end{cases} \tag{3}$$

In Equation (3), $S_{\text{match}}(d_i, t_j)$ is the matching score, and $\boldsymbol{\psi}(d_i)$ is the feature vector of the detection frame. $\boldsymbol{\psi}(t_j)$ is the feature vector of trajectory t_j. M is the adjacency matrix. d_o is the feature dimension. GNN is the GNN function. The GNN-based matching score leverages graph structures to model relationships between detections and trajectories, enhancing robustness in crowded scenes. To enhance the persistence of temporal features, this study defines the workflow of the memory module, as shown in Equation (4).

$$\begin{cases} \mathcal{M}_t = \mathcal{M}_{t-1} \cup \{F_{\text{hist}}(d_i)\} \\ F_{\text{retrieve}} = \arg\min_{F \in \mathcal{M}_t} \left\| F - F_{\text{current}} \right\| \end{cases} \tag{4}$$

In Equation (4), $\mathcal{M}_t$ is the feature memory bank of time t. $F_{\text{hist}}(d_i)$ is the historical feature of the detection frame. F_{retrieve} is the retrieved matching feature. F_{curent} is the feature of the current frame detection frame. The memory module retains historical features to improve temporal consistency under occlusion. Next, this study focuses on appearance information to enhance the robustness to changes in lighting and other conditions in densely populated scenes, as shown in Equation (5).

$$\text{Sim}_{\text{app}}(d_i, t_j) = \frac{\boldsymbol{f}_{\text{app}}(d_i) \cdot \boldsymbol{f}_{\text{app}}(t_j)}{\left\|\boldsymbol{f}_{\text{app}}(d_i)\right\| \left\|\boldsymbol{f}_{\text{app}}(t_j)\right\|} \tag{5}$$

In Equation (5), $\text{Sim}_{\text{app}}(d_i, t_j)$ is the appearance similarity. $\boldsymbol{f}_{\text{app}}(d_i)$ and $\boldsymbol{f}_{\text{app}}(t_j)$ are the appearance feature vectors of the detection frame and trajectory. Cosine similarity is adopted for appearance matching due to its invariance to feature magnitude and robustness to illumination changes. This study also achieves adaptive noise adjustment by dynamically adjusting the noise covariance, as shown in Equation (6).

$$\begin{cases} Q_{\text{adapt}} = Q_0 \cdot \exp(-\alpha \cdot \Delta t) \\ \Delta t = t_k - t_{\text{last}} \end{cases} \tag{6}$$

In Equation (6), Q_{adapt} is the noise covariance of the adaptive process. Q_0 is the initial noise covariance. α is the attenuation coefficient. Δt is the interval between the current time t_k and the last successful update time t_{last}. Adaptive noise covariance adjusts based on the time since last update, improving Kalman filter stability in dynamic scenes. Subsequently, this study fuses the previous moment trajectory and the current detection frame through weighted average, as shown in Equation (7).

$$\begin{cases} t_j^t = \beta \cdot t_j^{t-1} + (1-\beta) \cdot d_i \\ \beta = sigmoid(\text{Sim}_{\text{app}} + \lambda S_{\text{match}}) \end{cases} \tag{7}$$

In Equation (7), t_j^t is the updated trajectory state at time t. β is the fusion weight. Weighted fusion balances prior predictions with current observations to refine trajectory states. To sum up, the improved ByteTrack's flowchart is shown in Figure 1.

In Figure 1, the IBTA is based on the traditional framework and uses the COC module to perform motion correction on the unmatched prediction frame before the second matching. The purpose is to improve the matching success rate and trajectory coherence of low-scoring detection frames in camera-moving scenes, thereby enhancing the tracking robustness and accuracy of the algorithm in complex dynamic environments. Finally, this study comprehensively

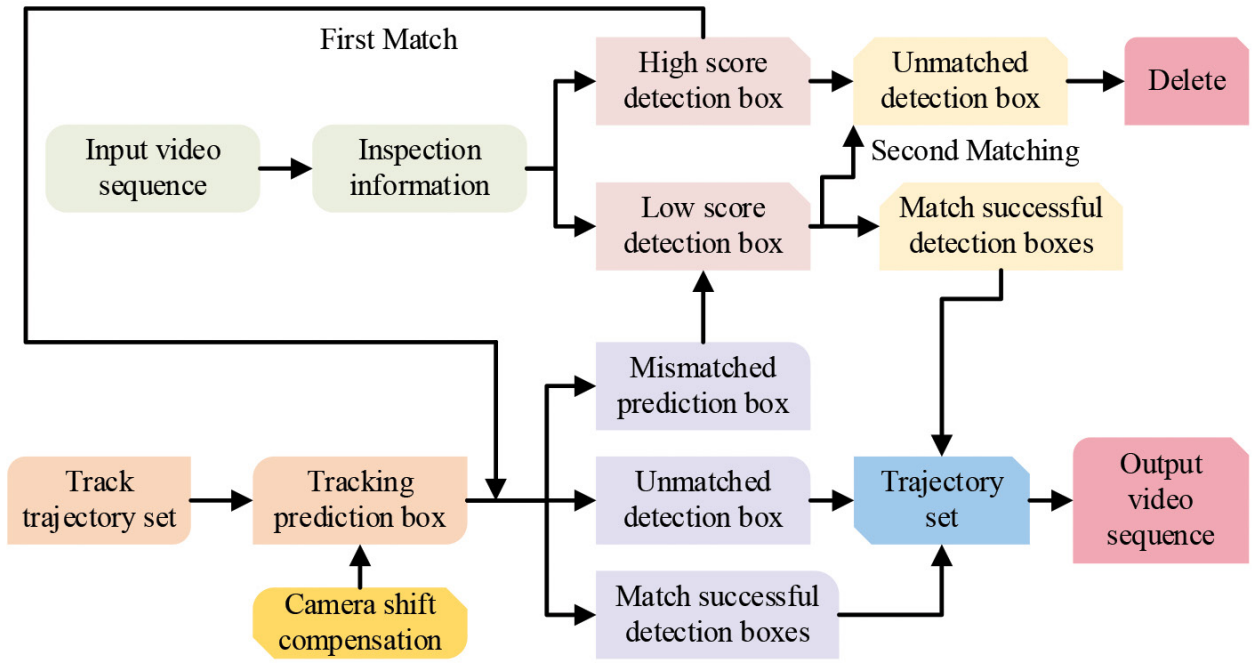


Figure 1. Improved byte track algorithm.

evaluates the detection error and correlation error in crowded scenes, as shown in Equation (8).

$$\mathrm{MOTA}_{\mathrm{enhanced}} = 1 - \frac{\sum_t (FN_t + FP_t + IDS_t)}{\sum_t GT_t} - \lambda \cdot \mathrm{IDF1} \tag{8}$$

In Equation (8), $\mathrm{MOTA}_{\mathrm{enhanced}}$ is the enhanced MTT accuracy. FN_t is the number of missed detections at time t. FP_t is the number of false detections. IDS_t is the number of ID switching times. GT_t is the real target number. IDF1 is the identity F1 score. λ is a trade-off parameter. Taken together, the MOT method framework for crowded scenes based on the IBTA is shown in Figure 2.

In Figure 2, this method is systematically improved on the ByteTrack framework for crowded scenes. It first introduces COC to correct low-scoring frames, and then designs a trajectory management module including Kalman filter adaptive noise adjustment to effectively solve challenges such as severe occlusion and small targets. The entire recognition process significantly enhances trajectory coherence and matching accuracy, and finally outputs improved MTT accuracy and other indicators through a dedicated evaluation module. Among them, IBTA implementation details include: (1) A detector backbone network and training strategy: It is clearly stated that the detector used is a detector based on YOLOv8, and its backbone network is CSPDarknet. The pre-trained weights are initialized on the COCO dataset

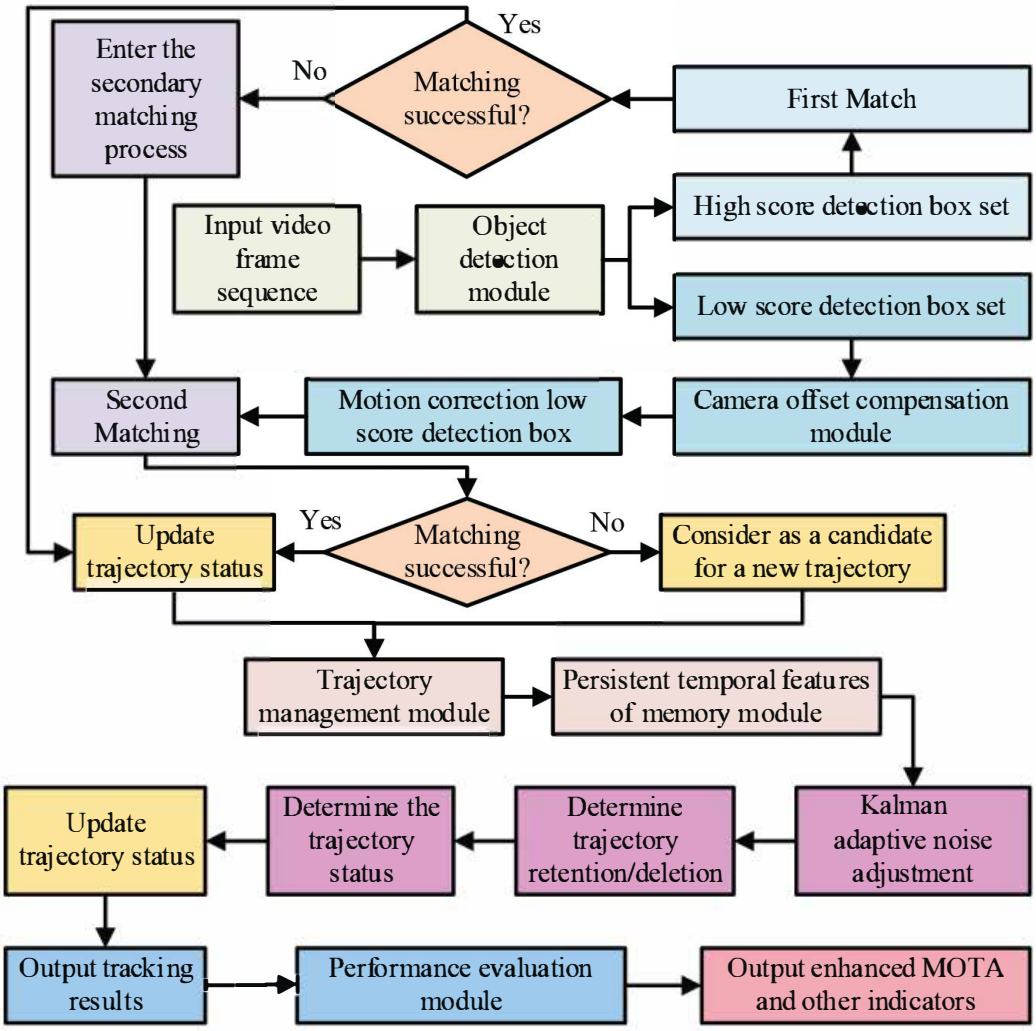


Figure 2. MOT method for crowded scenes based on IBTA.

and fine-tuned on the target dataset. The Adam optimizer is used during training, while the initial learning rate is 1e-3, and the cosine annealing strategy is adopted. (2) Key hyperparameter values: The parameters in Equations (1) and (6) are specified. The high confidence threshold is set to 0.6, and the low confidence threshold is set to 0.1. The trajectory retention time is set to 30 frames. The initial process noise covariance in Equation (6) is set to a diagonal matrix related to the target speed uncertainty, and the attenuation coefficient is set to 0.95. (3) GNN architecture details: Equation (3) illustrates the specific implementation of GNN. A two-layer Graph Attention Network (GAT) is used. The first layer maps the dimension of node features (appearance features and motion feature splicing of detection frames and trajectories) to 128, and the second layer outputs the matching score. The edges in the graph connect each detection box to all trajectory nodes, and the attention mechanism is used to calculate the association weight.

3.2. Fusion and improved IMC MOT method for dense crowd scenes

Although the IBTA can effectively alleviate the occlusion problem through the association strategy of high and low score detection frames, in crowded scenes, individual movement trajectories often appear random, sudden, and nonlinear. It is still prone to large deviations between the predicted trajectory and the real position, leading to ID switching or trajectory breakage [21-22]. IMC aligns the preceding and following frames to the same coordinate system by estimating the global motion transformation model between adjacent frames. This makes motion similarity measurement in ByteTrack more accurate, reducing mismatches and trajectory interruptions caused by dynamic changes in the background [23-24]. Among them, this study uses the image shift model to model the displacement source of image shift in crowded scenes, as shown in Equation (9).

$$\begin{cases} d_g(t) = A \cdot \sin(2\pi ft + \phi) + B \cdot \eta(t) \\ d_l(x,y,t) = C \cdot \nabla I(x,y,t) \cdot v_{\text{target}} \end{cases} \tag{9}$$

In Equation (9), $d_g(t)$ is the global displacement vector at time t. A is the amplitude, f is the vibration frequency, and ϕ is the phase offset. B is the noise coefficient, and $\eta(t)$ is Gaussian white noise. $d_l(x,y,t)$ is the local displacement vector at position (x,y) and time t. C is the proportional constant, $\nabla(x,y,t)$ is the image intensity gradient, and v_{target} is the target movement speed. The displacement model in Equation (9) decouples global camera motion from local target motion, based on the assumption of additive motion components. Next, this study estimates the overall motion vector between image frames in crowded scenes through the optical flow equation, as shown in Equation (10).

$$\begin{cases} \dfrac{\partial I}{\partial x}u + \dfrac{\partial I}{\partial y}v + \dfrac{\partial I}{\partial t} = 0 \\ \min_{u,v} \iint \left(u^2 + v^2\right) dxdy \end{cases} \tag{10}$$

In Equation (10), u and v are the horizontal and vertical displacements of optical flow. $\left(\dfrac{\partial I}{\partial x}, \dfrac{\partial I}{\partial y}\right)$ is the image space gradient. $\dfrac{\partial I}{\partial t}$ is the time gradient. Subsequently, this study quantifies the matching error of local feature points in consecutive frames, as shown in Equation (11).

$$E_{\text{match}} = \sum_{i=1}^{N} \left\| \mathbf{F}(p_i,t) - \mathbf{F}(q_i,t+1) \right\|^2 + \lambda \sum_{i \neq j} \left| d(p_i,q_j) \right| \tag{11}$$

In Equation (11), E_{match} is the sum of matching errors, N is the number of feature points, and p_i is the position in time frame t. q_i is the matching position of the feature point in frame $t + 1$. $F(q_i,t + 1)$ is the feature vector at position p_i and time t. $d(p_i,q_j)$ is the distance function between positions p_i and q_j. The feature matching error in Equation (11) quantifies alignment accuracy between consecutive frames, supporting robust motion estimation for IMC. This study also achieves multi-scale feature fusion through a weighted feature pyramid network structure, as shown in Equation (12).

$$\begin{cases} F_{\text{fused}} = \sum_{s=1}^{S} w_s \cdot \text{Resize}(F_s) \\ w_s = \exp(\alpha_s) \Big/ \sum_{s'=1}^{S} \exp(\alpha_{s'}) \end{cases} \tag{12}$$

In Equation (12), F_{fused} is the fused feature map, S is the number of scales, and F_s is the feature map of the s-th scale. Resize(F_s) is the resize operation, w_s is the weight of the scale, and α_s is

the learnable parameter. Subsequently, to learn long-distance dependencies in images and handle complex image motion patterns in crowded scenes, this study introduces the self-attention mechanism [25] in Transformer, as shown in Equation (13).

$$\begin{cases} \text{Attention}(Q,K,V) = \text{Softmax}\left(\dfrac{QK^T}{\sqrt{d_k}}\right)V \\ Q = F_{\text{fused}}W_Q, \quad K = F_{\text{fused}}W_K, \quad V = F_{\text{fused}}W_V \end{cases} \tag{13}$$

In Equation (13), Q, K, and V are query, key, and value matrices. W_Q, W_K, and W_V are learnable weight matrices. d_k is the dimension of the key vector. Softmax is the normalization function. Next, this study defines the expression of affine transformation as shown in Equation (14).

$$\begin{cases} \begin{bmatrix} x' \\ y' \end{bmatrix} = \mathbf{M}\begin{bmatrix} x \\ y \end{bmatrix} + \mathbf{T} \\ \mathbf{M} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}, \quad \mathbf{T} = \begin{bmatrix} t_x \\ t_y \end{bmatrix} \end{cases} \tag{14}$$

In Equation (14), (x,y) is the original image coordinate. (x',y') is the coordinate after compensation. **M** is a 2×2 affine matrix. **T** is the translation vector. To sum up, the improved version of IMC is shown in Figure 3.

In Figure 3, the improved IMC is a multi-level processing system. It extracts features from the input video, accurately estimates motion by integrating optical flow analysis. Multi-scale feature fusion (Equation 12) and self-attention (Equation 13) are integrated to refine motion compensation, *etc.*, then calculates the affine transformation matrix, and finally realizes the adjustment and reprojection of the image prediction frame and outputs stable video frames. Finally, this study evaluates the performance of the MOT method, as shown in Equation (15).

$$\begin{cases} L = L_{\text{det}} + \lambda L_{\text{comp}} \\ L_{\text{det}} = \sum_{i=1}^{N}\left[\text{CE}(p_i,\hat{p}_i) + \text{GIoU}(b_i,\hat{b}_i)\right] \\ L_{\text{comp}} = \left\| I_{\text{comp}} - I_{\text{gt}} \right\|^2 \end{cases} \tag{15}$$

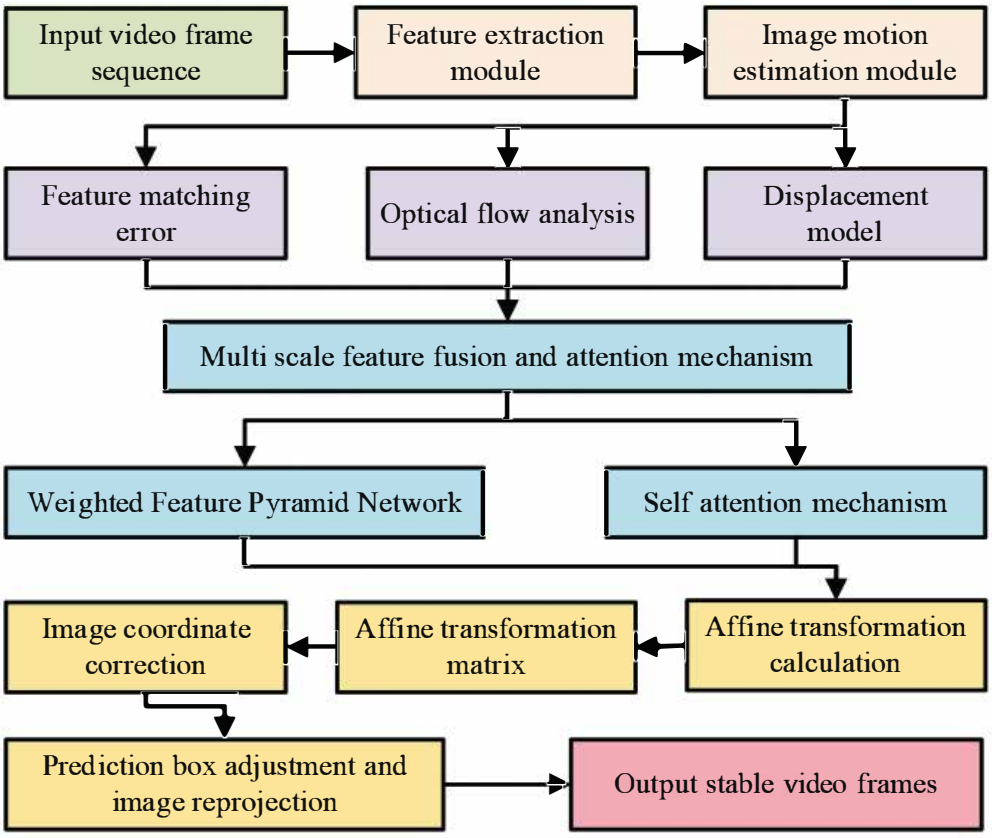


Figure 3. Improved IMC mechanism.

In Equation (15), L is the total loss function. L_{det} is the target detection loss. $CE(p_i,\hat{p}_i)$ is the classification cross entropy loss. $GIoU(b_i,\hat{b}_i)$ is the generalized intersection and union loss. L_{comp} is compensation for losses. I_{comp} is the compensated image. I_{gt} is a real image. Overall, the MOT method for crowded scenes that combines IMC and the IBTA is shown in Figure 4.

In Figure 4, the research method creatively deeply integrates IMC into the improved ByteTrack framework. It first stabilizes the video input through multi-scale feature fusion and motion compensation, and then uses a detector to obtain the target. In the data association stage, the high-scoring detection frames are first matched, and then the low-scoring frames corrected by COC are matched twice to effectively restore the occluded targets. At the same time, it introduces adaptive noise adjustment Kalman filter to improve the robustness of trajectory prediction, and finally outputs tracking indicators through a comprehensive evaluation module. Among them, the integration steps of IMC and IBTA are as follows: First, input preprocessing is performed, and the improved IMC module is used to estimate the affine transformation matrix of the current frame relative to the previous frame on the input video frame. Then, target detection is performed, and the stable frames are input into the YOLOv8 detector to obtain a set of detection boxes and their confidence scores. Afterwards, the detection boxes are compensated, and the coordinates of the low-scoring detection boxes are corrected to obtain a compensated detection box set. Subsequently, trajectory prediction and matching are performed, and Kalman filtering is used to predict the position of existing trajectories at each time step. The high score detection boxes are matched with the predicted trajectories based on IoU and appearance similarity (Equation 5). The unmatched predicted trajectory is matched with the compensated low-scoring detection frame for a second time. The rules are the same as the first matching, and the occluded target is restored. Trajectory management is performed, and the successfully matched trajectory status is updated (Equation 7). New trajectories are initialized for unmatched detection frames, trajectories that have not been matched for a long time are deleted, and finally, the target ID, bounding box, and tracking indicators of the current frame are output.

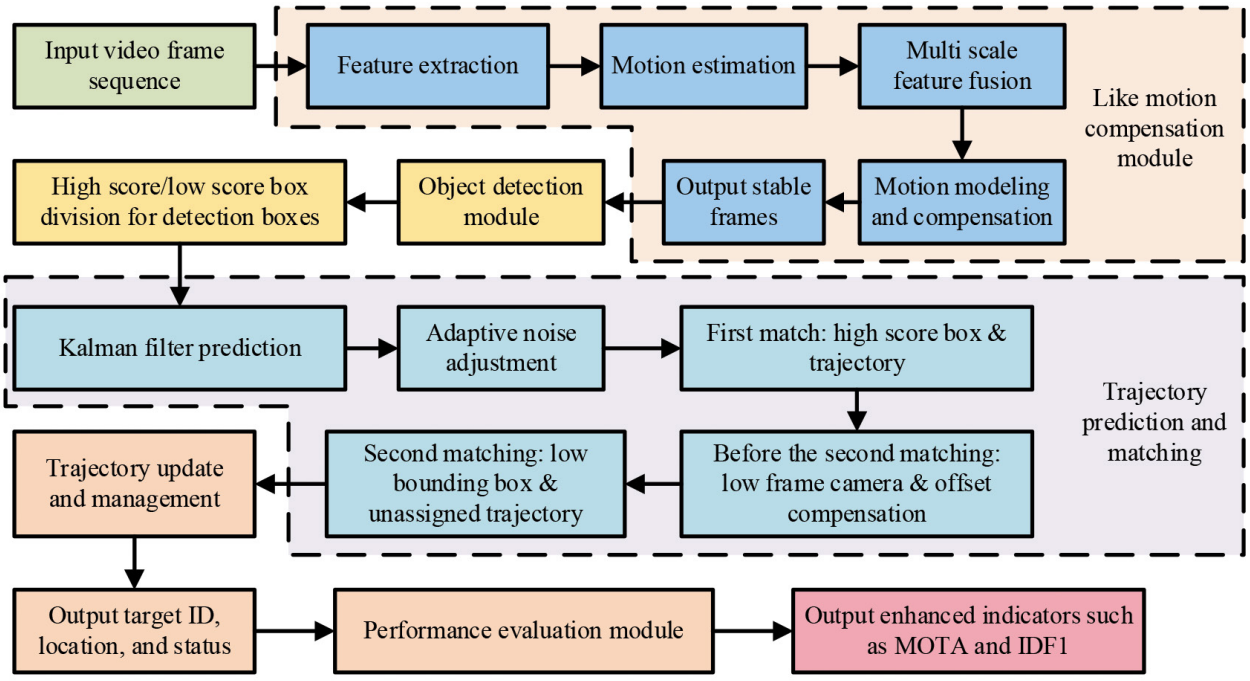


Figure 4. MOT method for crowded scenes integrating IMC and IBTA.

Although previous studies have discussed IMC, COC, and adaptive noise adjustment separately, this study systematically integrates them into the ByteTrack MTT framework for the first time and performs collaborative optimization for challenges such as target occlusion, small target detection, and camera shake in crowded scenes. The novelty of the proposed method is mainly reflected in: (1) A multi-level motion correction mechanism combines global and local IMC to improve detection stability under dynamic backgrounds. (2) Camera-motion-aware secondary matching: It introduces COC before the second association of low-score detection boxes in ByteTrack to enhance the re-identification success rate of occluded targets. (3) Adaptive trajectory prediction: It enhances the robustness of trajectory prediction through noise-adaptive Kalman filtering. (4) End-to-end system integration unifies the above modules into a trainable and evaluable framework, achieving full pipeline optimization from video stabilization to trajectory output.

Therefore, this study is not merely an engineering assembly of existing components, but a structured solution addressing the performance bottlenecks of MTT systems in dense scenes.

4. Results and discussion

4.1. Performance test of MOT method in crowded scenes

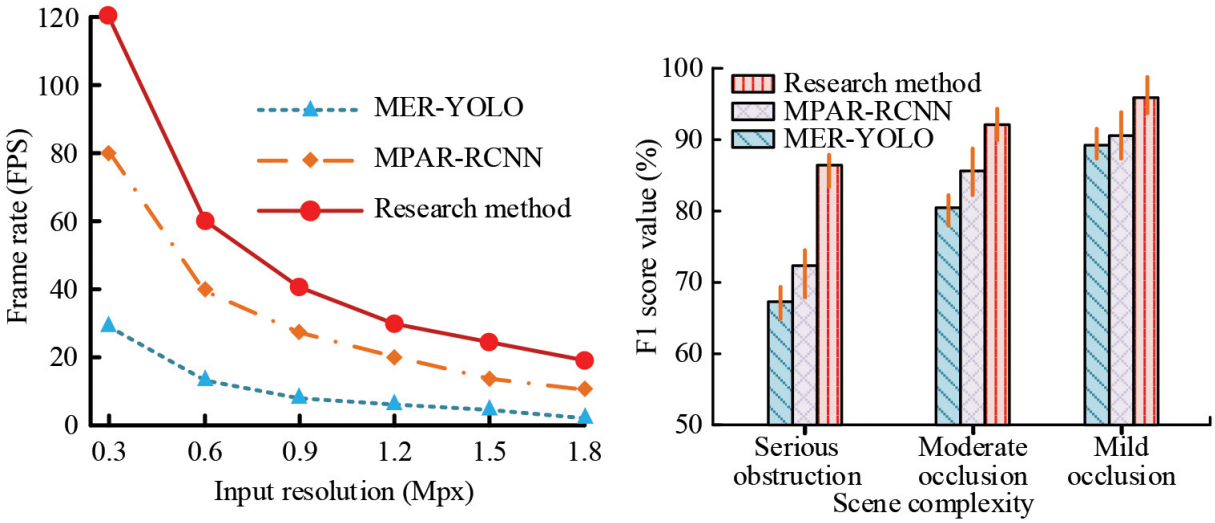
To verify the performance of the MOT method in dense crowd scenes that integrates IMC and the IBTA, this study constructs a simulation platform. The experimental configuration is shown in Table 1.

In Table 1, the experimental dataset uses WiderPerson. The research method is compared with the Multi-Person Attribute Recognition Region-based Convolutional Neural Network (MPAR-RCNN) and the MobileViT-Enhanced Repulsion YOLO (MER-YOLO). The processing speed of the three methods under different image input resolutions and the F1 score comparison under different scene complexities are shown in Figure 5.

In Figure 5(a), the research method maintains the highest FPS across resolutions. The research method has the smallest downward trend, and when the input resolution increases from 0.3Mpx to 1.8Mpx, the processing time drops from 120FPS to 20FPS. The other two methods have significantly smaller frame rates

Table 1. Test environment and specific configuration.

Name	**Configuration**
GPU	NVIDIA RTX 3080/4090
CPU	AMD Ryzen 7/9 series
Memory	32GB or more
Deep learning framework	PyTorch
Image and video processing	OpenCV
Graphical interface construction	PyQt5
Optimizer	Adam or SGD



(a) Processing speed at different input resolutions

(b) F1 score values in different scenarios

Figure 5. F1 score values of processing speed and scene complexity for different input resolutions.

at different input resolutions. In Figure 5(b), the F1 score of the research method in different scene complexities is overall greater than 85%, and it is 86.5%, 92.3%, and 95.8% in severe, moderate, and mild occlusion scenes. The F1 score of the comparison method in various occlusion scenarios is smaller than that of the research method. The research method can better balance accuracy and speed in dense scenes, and has better stability and anti-interference ability when facing complex scenes. Performance tests show that the MOT method for crowded scenes that combines IMC and the IBTA has good recognition accuracy and can effectively reduce missed detections and false detections in extreme environments with highly dense targets and severe mutual occlusion.

4.2. Practical application effect of MOT method in dense crowd scenes

On the basis of verifying the performance of the MOT method in dense crowd scenes that integrates IMC and the IBTA, this study will further verify the practical application effect of the research method. This study builds a multi-target real-time perception and risk assessment platform for dense scenes and uses the MovingDroneCrowd dataset for testing. The main scenes of this dataset include dense crowds such as business districts and scenic spots shot by dynamic drones, and focus on the video data set from the perspective of dynamic drones, including different flight heights, angles, and lighting changes. The research method is compared with MPAR-RCNN and MER-YOLO. The proportion of trajectories and MTT accuracy of the three methods under different scene densities are shown in Figure 6.

In Figure 6, as the scene density increases and occlusion intensifies, the proportion of the number of trajectories and the MTT accuracy of the three methods show a downward trend. In Figure 6(a), in the density of extremely large scenes and small scenes, the number of trajectories of the research method accounts for 85.2% and 98.4%, and the MTT accuracy is 82.1% and 97.2%. In Figures 6(b) and 6(c), the number of trajectories of MPAR-RCNN and MER-YOLO in small scene density accounts for 95.3% and 93.8%, and the MTT accuracy is 94.1% and 92.5%. Therefore, the proportion of the number of trajectories and the MTT accuracy of the two comparison methods are smaller under all scene densities. Overall, the research method can maintain higher stability under occlusion and density changes, thereby significantly improving the reliability of early warning. The average accuracy of the three methods at different IoU thresholds, and the Average Recall Rate (ARR) of different detection numbers for each picture are shown in Figure 7.

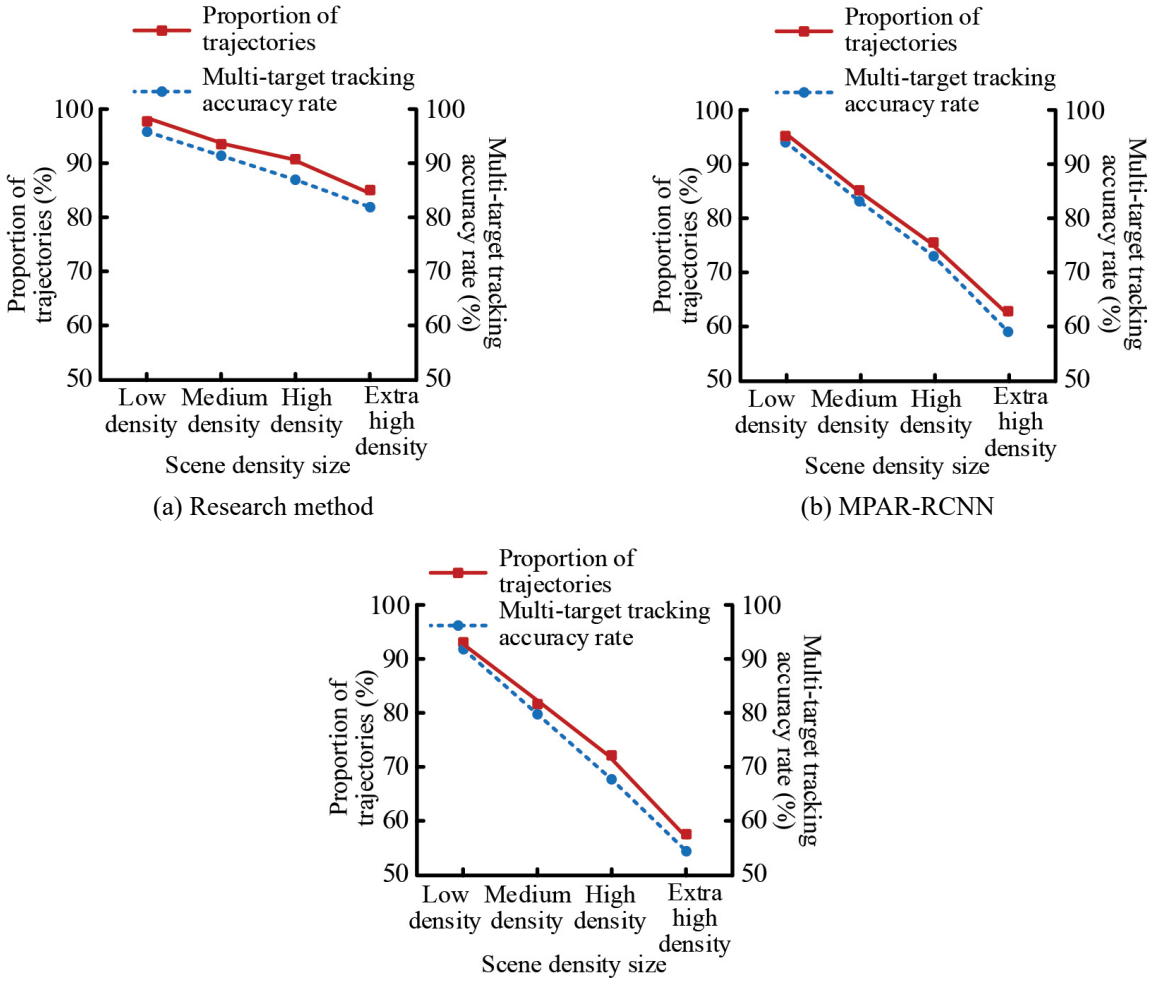
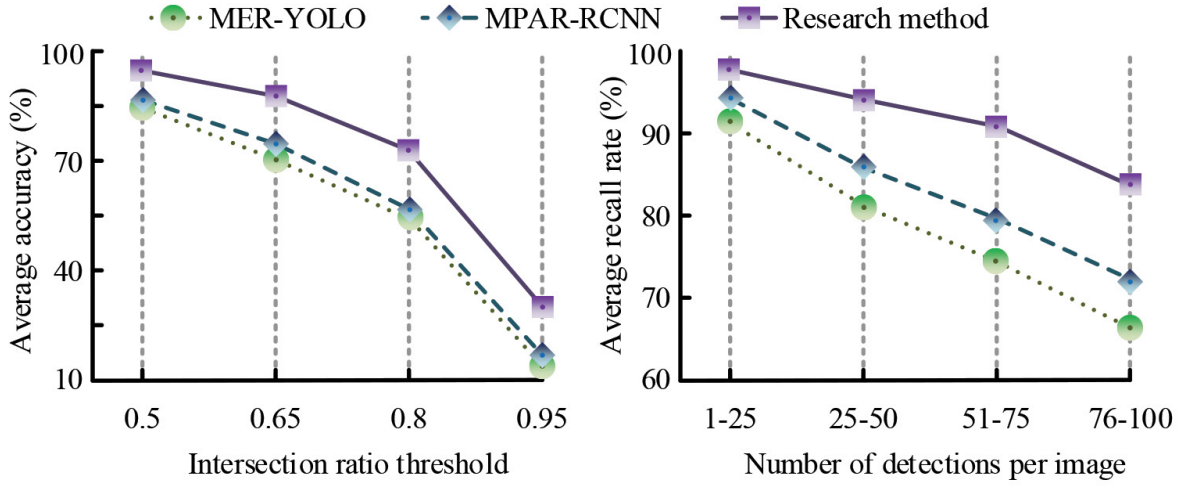


Figure 6. Proportion of trajectory count and MTT accuracy under different scene densities.



(a) Average accuracy of different intersection to union ratio tresholds

(b) Average recall rate of different detection numbers for each image

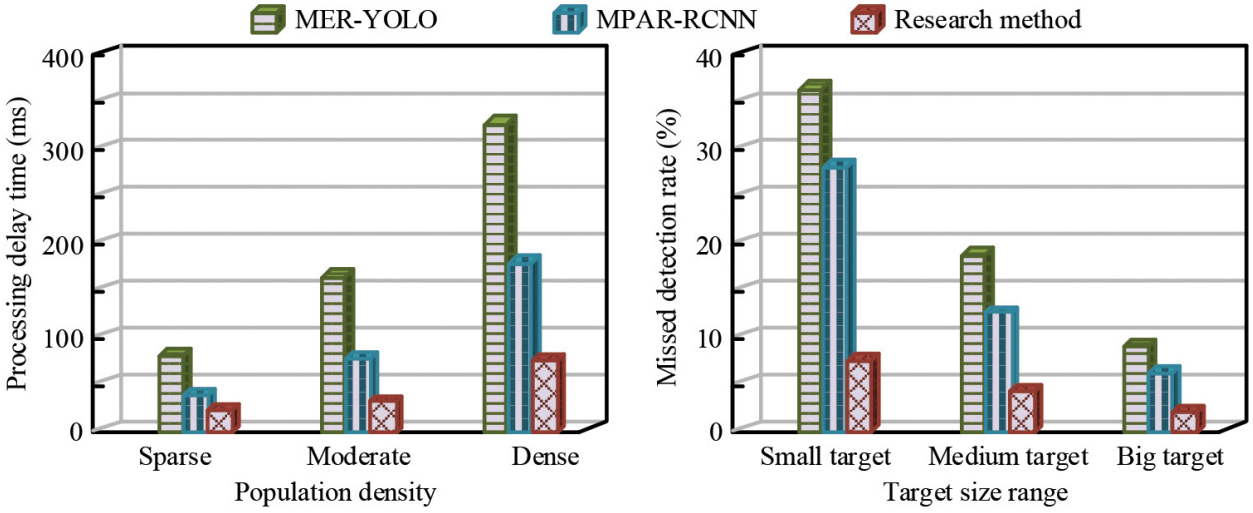
Figure 7. The average accuracy of different IoU thresholds and the ARR of different detection numbers.

In Figure 7(a), the average accuracy of the three methods decreases as the IoU threshold increases. When the IoU thresholds are 0.5 and 0.95, the average accuracy of the research method is 95.2% and 30.5%. MPAR-RCNN and MER-YOLO have average accuracies of 18.2 and 12.7% when the IoU threshold is 0.95. In Figure 7(b), the ARR of the research method is greater than 85% under different detection numbers per picture. The ARR is 98.5% when the number of detections per picture is 1–25, and 85.7% when the number of detections per picture is 76-100 times. The ARR of the comparison methods under different detection numbers per image is significantly smaller than that of the research method. The research method can better reduce the false overlap or omission of adjacent target frames in dense crowd scenes with large amounts of occlusion. Figure 8 shows the processing delay times of the three methods at different population densities and the missed detection rates in different target size intervals.

In Figure 8(a), the overall processing delay time of the research method under different population densities is less than 100 ms. The processing delay time is 28 ms in sparse crowds, 42 ms in medium-density crowds, and 68 ms in dense crowds. The processing latency of the comparison method is significantly greater than that of the study method at various population densities. In Figure 8(b), the overall missed detection rate of the research method in different target size intervals is less than 10%. The missed detection rate is 8.5%, 4.2%, and 2.1% when the target size is small target, medium target, and large target. Other methods have significantly greater missed detection rates in various target size intervals. Research methods can detect potential risks more promptly and issue early warnings, and are more robust and adaptable when dealing with targets of different scales. In practical application tests, the research method has good multi-scale detection capabilities in MOT tasks in dense crowd scenes, can effectively extract features and distinguish overlapping targets, and alleviate the problem of missed detection in dense scenes.

4.3. Comparative Experiments with Mainstream Benchmark Methods and Datasets

To rigorously evaluate the effectiveness of the IMC-IBTA method and make its results meet the standards of the MOT community, the study is conducted on widely used public benchmark datasets and compared with multiple mainstream advanced methods. This verifies the



(a) Processing delay time for different population densities

(b) Missed detection rates in different target size ranges

Figure 8. Processing delay time for different population densities and missed detection rates for different target size intervals.

comprehensive performance of IMC-IBTA under standard evaluation indicators, especially its ability to handle occlusion, small targets, and identity switching problems in crowded scenes. The study first evaluates tracking performance on the MOT17 and MOT20 challenge datasets. These two datasets contain rich pedestrian scenes with different densities, lighting, and occlusion levels, and are authoritative benchmarks for evaluating MOT algorithms. The study compares IMC-IBTA with the original ByteTrack, BoT-SORT, and OC-SORT. These methods are outstanding and widely recognized baselines in the MOT field in recent years. The evaluation uses standard indicators of MTT, including Multi-Target Tracking Accuracy (MOTA), Multi-Target Tracking Precision (MOTP), Identity F1 score (IDF1), High-Order Tracking Accuracy (HOTA), and the number of IDs. The comparison results are shown in Table 2.

In Table 2, on the more challenging MOT20 dataset, all methods exhibit lower metrics compared to MOT17, reflecting the difficulty of high-density scenes. On MOT17, IMC-IBTA outperforms all compared methods in three core metrics: MOTA, IDF1, and HOTA. The significant improvement in IDF1 (78.8%) and the reduction in the number of IDs (1224) directly demonstrate that integrating IMC and improved data association strategies effectively reduces identity switches and enhances trajectory coherence. On MOT20, IMC-IBTA also maintains leading performance across all major metrics, particularly reducing IDs by approximately 14% compared to the original ByteTrack, indicating stronger robustness in extremely crowded and severely occluded environments. To further verify the basic ability of the method in dense crowd detection, the study conducts comparative experiments on the CrowdHuman validation set. CrowdHuman focuses on highly occluded pedestrian detection. Its annotations include visible boxes and full boxes. It is an important dataset for evaluating the performance of dense target detection and correlation pre-order. The study compares the Average Precision (AP) of the detector used by

Table 2. Comparison of tracking performance on MOT17 and MOT20 test sets.

Method	Dataset	MOTA (%)	MOTP (%)	IDF1 (%)	HOTA (%)	IDs
ByteTrack	MOT17	76.6	78.2	75.2	59.8	1597
BoT-SORT	MOT17	77.8	78.9	76.5	61.2	1456
OC-SORT	MOT17	78.0	78.5	77.1	62.0	1389
IMC-IBTA	MOT17	79.5	79.8	78.8	63.5	1224
ByteTrack	MOT20	61.7	75.9	64.2	48.1	4478
BoT-SORT	MOT20	62.6	76.2	65.0	49.0	4321
OC-SORT	MOT20	63.1	75.8	65.8	49.8	4105
IMC-IBTA	MOT20	64.8	76.5	67.5	51.3	3850

IMC-IBTA with YOLOX (the original ByteTrack detector) and some other advanced detectors on this dataset. In addition, the study also reports related performance metrics on the CrowdHuman tracking segmentation (Track) subset. The results are shown in Table 3.

In Table 3, the customized detection backbone network achieves 88.9% AP and 46.8% mMR on the CrowdHuman validation set. The detection performance is better than YOLOX-L, which is the ByteTrack baseline, laying a solid foundation for subsequent high-quality tracking correlation. On the CrowdHuman-Track subset, the Detection Accuracy (DetA) and Association Accuracy (AssA) of IMC-IBTA reach the maximum values, which are 83.1% and 72.8%. This shows that the research method can effectively distinguish multiple highly overlapping targets even when faced with extremely dense and occluded static images in CrowdHuman. Compared with methods that use the same detector but different tracking strategies, the improvement of IMC-IBTA on AssA is particularly obvious, verifying the synergistic effect of the improved ByteTrack framework and IMC mechanism in complex correlation tasks. In summary, through system comparative experiments on MOT17, MOT20, CrowdHuman, and other standard benchmarks, the IMC-IBTA method is superior to ByteTrack, BoT-SORT, and OC-SORT in multiple core tracking and detection indicators and is in line with the evaluation specifications of the MOT community. It strongly proves the effectiveness of this method in improving the accuracy, stability, and identity preservation of target tracking in dense scenes.

4.4. Ablation experiment analysis

To quantify the contribution of each module to the system performance, this study designs an ablation experiment, successively removing the IMC, COC, adaptive noise adjustment, and GNN matching modules. The experimental results are shown in Table 4.

In Table 4, the IMC module contributes the most to improving tracking stability, and COC and GNN matching play a significant role in reducing ID switching. To further evaluate the practicability of the method, this study compares the computing efficiency of IMC-IBTA, ByteTrack, and BoT-SORT in the same hardware environment. When processing 1080p

Table 3. Detection and association performance comparison on the CrowdHuman dataset.

Method	**Detection Backbone**	**CrowdHuman (Val Set)**	**CrowdHuman-Track**	**DetA (%)**	**AssA (%)**
		AP@0.5:0.95 (Full Box) (%)	**mMR (Full Box) (%)**		
YOLOX-S	YOLOX	84.0	52.1	78.2	65.4
YOLOX-L	YOLOX	87.2	48.5	81.5	68.9
Ours (Detector)	Custom	88.9	46.8	83.1	71.2
ByteTrack	YOLOX-L	87.2	48.5	81.5	69.5
OC-SORT	YOLOX-L	87.2	48.5	82.0	70.1
IMC-IBTA	Custom	88.9	46.8	83.1	72.8

Table 4. Ablation experiment results (on the MOT17 validation set).

Configuration	MOTA (%)	IDF1 (%)	IDs
Complete model	79.5	78.8	1224
Remove IMC	77.1	76.3	1456
Remove COC	78.0	77.5	1320
Remove adaptive noise	78.8	77.9	1288
Remove GNN matching	78.2	77.1	1350

images on RTX 3080, the average inference time of IMC-IBTA is 32 ms/frame, which is slightly higher than ByteTrack (28 ms/frame) but significantly lower than BoT-SORT (45 ms/frame). The number of model parameters increases by about 8%, indicating that the performance improvement does not bring significant computational burden.

4.5. Discussion

The proposed multi-object tracking method, which integrates an image motion compensation mechanism with an improved ByteTrack algorithm, demonstrates outstanding tracking performance across multiple dense-scene datasets. The following discussion elaborates on its theoretical contributions and practical implications.

(1) Theoretical Implications: First, this study is the first to deeply integrate an image motion compensation mechanism into the ByteTrack framework, proposing a multi-level motion correction strategy that significantly enhances the stability of object detection and trajectory association under dynamic backgrounds. Traditional methods often struggle to maintain identity consistency under camera shake or rapid motion. By introducing global and local motion compensation, the proposed approach aligns trajectory predictions more closely with actual motion states, representing a strong theoretical innovation. Second, the improved ByteTrack framework incorporates camera offset compensation and a graph neural network matching module into its two-stage association strategy, thereby enhancing the re-identification capability of occluded targets. This improvement preserves the efficiency of the original ByteTrack while significantly boosting robustness in extremely crowded scenes, offering new insights for future research. Additionally, the adaptive noise-adjusted Kalman filter proposed in this study dynamically adjusts the noise covariance when a target has not been updated for an extended period, improving the stability of trajectory prediction. This mechanism theoretically enriches the application of Kalman filtering in dynamic vision tasks.

(2) Practical Implications: From a practical perspective, the proposed method demonstrates strong potential for engineering deployment in several dimensions. First, in intelligent surveillance systems, multi-object tracking in crowded scenes is fundamental for tasks such as high-risk warning and crowd density analysis. The proposed method maintains an F1 score above 85% under varying occlusion and density conditions, with a processing delay of less than

100 ms, making it suitable for real-time deployment. Second, in autonomous driving and mobile robotic vision systems, camera motion is frequent and target scales vary dramatically. Experimental results on the MovingDroneCrowd dataset show that the method achieves a missed detection rate of less than 8.5% for small targets and higher accuracy for medium and large targets, indicating robust multi-scale perception capabilities. Furthermore, the method's strong performance on public benchmark datasets such as MOT17, MOT20, and CrowdHuman validates its generalization ability and robustness. The significant reduction in identity switches and improvement in IDF1 scores suggest that the method can effectively mitigate identity confusion in real-world deployments, enhancing system reliability and user experience.

In summary, this study not only deepens the theoretical integration of motion compensation and data association mechanisms but also empirically validates the method's efficiency and stability in complex dynamic scenes, highlighting its academic value and broad application prospects.

5. Conclusion

To address the traditional MOT method in dense crowd scenes, such as insufficient feature expression ability and a lack of robustness to complex environmental conditions, this study innovatively proposed an MOT method in dense crowd scenes that integrates IMC and the IBTA. This study innovatively integrated COC and IBTA, and improved the recognition accuracy of target tracking in dense scenes through adaptive noise adjustment and multi-level image motion correction. In the experiment, the processing time of the research method was 20FPS when the input resolution was 1.8Mpx. Its F1 scores were 95.8% and 86.5% in mild and severe occlusion scenes. In the actual application test, the research method accounted for 85.2% of the number of trajectories in the stadium scene, and the MTT accuracy was 82.1%. Its AP was 95.2% when IoU threshold = 0.5 and 30.5% when IoU threshold = 0.95. Its ARR was 98.5% and 85.7% when the number of detections per image was 1–25 and 76–100. Its processing latency was 28 ms in sparse crowds and 68 ms in dense crowds. The missed detection rate was 8.5% when the target size was small and 2.1% when the target was large. The research method has better recognition accuracy and stability in crowded scenes, and can effectively alleviate the problem of missed detection in dense scenes. However, this study still has certain limitations. The experiments were primarily validated on two datasets: WiderPerson and MovingDroneCrowd. The scene coverage of these datasets is limited, lacking real dense crowd scenes under conditions such as low-light at night, extreme weather (*e.g.*, rain, snow, fog), and severe illumination variations. Furthermore, this study did not conduct cross-dataset generalization tests on more challenging or out-of-distribution public benchmark datasets. Therefore, the robustness and general applicability of the proposed method in the above-mentioned undiscovered extreme and complex scenarios are still not fully determined. This may impact its performance in real-world, 24x7 and 24x7 surveillance deployments.

Declaration of Competing Interests

The author declares no conflict of interest.

Funding

This research was Supported by Chengdu College of Arts and Sciences School-level Project "Research on Blind Field Intelligent Inspection Technology in Complex Computer Room Scenarios Based on Multimodal Fusion" (No. WLZD202603).

Data Availability

Data used in this article are available upon request from the author.

References

[1] T. J. Thirumalai *et al.*, "A hybrid optimisation enabled deep learning for object detection and multi-object tracking", *Int. J. of Ad Hoc and Ubiquitous Computing*, vol. 46, no. 3, pp. 150–165, 2024. http://dx.doi.org/10.1504/IJAHUC.2024.140033

[2] M. Dang *et al.*, "Multi-object behaviour recognition based on object detection cascaded image classification in classroom scenes", *Applied Intelligence*, vol. 54, no. 6, pp. 4935–4951, 2024. http://dx.doi.org/10.1007/s10489-024-05409-x

[3] A. Albouchi *et al.*, "Implementation of an improved multi-object detection, tracking, and counting for autonomous driving", *Multimedia Tools and Applications*, vol. 83, no. 18, pp. 53467–53495, 2024. http://dx.doi.org/10.1007/s11042-023-17444-w

[4] X. Tong *et al.*, "YO-BYNet: multi-object tracking of fish based on YOLOv8 and BYTE association method", in *Proc. IET Conf. Proc.* CP850, Stevenage, UK, 2023, pp. 111–116. http://dx.doi.org/10.1049/icp.2023.2925

[5] X. Guo, Y. Zheng, and D. Wang, "PMTrack: Multi-object tracking with motion-aware," in *Proc. Asian Conf. Comput. Vis.*, pp. 3091–3106, 2024. https://dx.doi.org/10.1007/978-981-96-0960-4_26

[6] C. Zha *et al.*, "Infrared multi-target detection and tracking in dense urban traffic scenes", *IET Image Processing*, vol. 18, no. 6, pp. 1613–1628, 2024. http://dx.doi.org/10.1049/ipr2.13053

[7] L. Liang, "Computer vision and deep learning in ship target detection", *Journal of Computing and Electronic Information Management*, vol. 17, no. 3, pp. 10–14, 2025. http://dx.doi.org/10.54097/328xmp36

[8] V. M. Scarrica *et al.*, "A hybrid approach to real-time multi-target tracking", *Neural Computing and Applications*, vol. 36, no. 17, pp. 10055–10066, 2024. http://dx.doi.org/10.1007/s00521-024-09799-4

[9] A. A. Rafique *et al.*, "Maximum entropy scaled super pixels segmentation for multi-object detection and scene recognition via deep belief network", *Multimedia tools and applications*, vol. 82, no. 9, pp. 13401–13430, 2023. http://dx.doi.org/10.1007/s11042-022-13717-y

[10] V. D. Shevchenko *et al.*, "Conceptual structure of a decision support system for situational video analytics for multi-object recognition", *Vestnik of Astrakhan State Technical University. Series: Management, Computer Sciences and Informatics*, no. 1, pp. 69–79, 2025. http://dx.doi.org/10.24143/2072-9502-2025-1-69-79

[11] L. Qingxia *et al.*, "Method for the automatic calculation of newborn lambs activity using Byte-Track algorithm enhanced by state vector", *Transactions of the Chinese Society of Agricultural Engineering*, vol. 40, no. 13, pp. 146–155, 2024. http://dx.doi.org/10.11975/j.issn.1002-6819.202404068

[12] J. Jang and Y. Kwon, "Trajectory similarity-based traffic flow analysis using YOLO+ByteTrack", *Journal of Multimedia Information System*, vol. 12, no. 1, pp. 27–40, 2025. http://dx.doi.org/10.33851/JMIS.2025.12.1.27

[13] Y. Jia *et al.*, "Individual motion feature extraction method for sea bass based on improved YOLOv8 and ByteTrack", *Transactions of the Chinese Society of Agricultural Engineering*, vol. 41, no. 5, pp. 182–190, 2025. http://dx.doi.org/10.11975/j.issn.1002-6819.202409089

[14] X. Wang *et al.*, "Compensation atmospheric scattering model and two-branch network for single image dehazing", *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 8, no. 4, pp. 2880–2896, 2024. http://dx.doi.org/10.1109/TETCI.2024.3386838

[15] A. Klein *et al.*, "The Fermi energy as common parameter to describe charge compensation mechanisms: A path to Fermi level engineering of oxide electroceramics", *Journal of Electroceramics*, vol. 51, no. 3, pp. 147–177, 2023. http://dx.doi.org/10.1007/s10832-023-00324-y

[16] J. Yang, D. Feng, Y. Gao, and C. Liu, "Online multi-object tracking based on record confidence and hierarchical association for cyber-physical social intelligence," *Big Data Min. Anal.*, vol. 8, no. 4, pp. 851-866, 2025. https://dx.doi.org/10.26599/BDMA.2025.9020024

[17] A. C. Aka *et al.*, "An efficient anchor-free localization algorithm for all cluster topologies in a wireless sensor network", *International Journal of Computers Communications & Control*, vol. 18, no. 3, p. 4961, 2023. http://dx.doi.org/10.15837/ijccc.2023.3.4961

[18] H. Kang *et al.*, "Recognition of unsafe behaviors of key position personnel in coal mines based on improved YOLOv7 and ByteTrack", *Journal of Mine Automation*, vol. 50, no. 3, pp. 82–91, 2024. http://dx.doi.org/10.13272/j.issn.1671-251x.2024030015

[19] Y. X. Chen *et al.*, "Simulation of crowd evacuation behaviours at subway stations under panic emotion", *International Journal of Simulation Modelling*, vol. 22, no. 4, pp. 667–678, 2023. http://dx.doi.org/10.2507/IJSIMM22-4-668

[20] S. Bilakeri and K. A. Kotegar, "Learning to track with dynamic message passing neural network for multi-camera multi-object tracking," *IEEE Access*, vol. 12, pp. 63317-63333, 2024. https://dx.doi.org/10.1109/ACCESS.2024.3383138

[21] C. Feng *et al.*, "Exploring separable attention for multi-contrast MR image super-resolution", *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 9, pp. 12251–12262, 2024. http://dx.doi.org/10.1109/TNNLS.2023.3253557

[22] H. Wang *et al.*, "Recursive deformable pyramid network for unsupervised medical image registration", *IEEE Transactions on Medical Imaging*, vol. 43, no. 6, pp. 2229–2240, 2024. http://dx.doi.org/10.1109/TMI.2024.3362968

[23] Y. Liu *et al.*, "UWB-INS fusion positioning based on a two-stage optimization algorithm", *Tehnicki vjesnik–Technical Gazette*, vol. 30, no. 1, pp. 185–190, 2023. http://dx.doi.org/10.17559/TV-20221019035741

[24] W. Wang *et al.*, "An improved Deeplabv3+ model for semantic segmentation of urban environments targeting autonomous driving", *International Journal of Computers Communications & Control*, vol. 18, no. 6, p. 5879, 2023. http://dx.doi.org/10.15837/ijccc.2023.6.5879

[25] L. Gao, "MFE-Transformer: Adaptive English text named entity recognition method based on multi-feature extraction and Transformer", *Computer Science and Information Systems*, vol. 21, no. 4, pp. 1865–1885, 2024. http://dx.doi.org/10.2298/CSIS240418061G

Received: January 2026
Revised: March 2026
Accepted: March 2026

Contact address:
Huanting Feng
School of Artificial Intelligence and Big Data
Chengdu College of Arts and Sciences
Chengdu, 610401
China
e-mail: fenghuanting1980@163.com

Huanting Feng is an Associate Professor at the School of Artificial Intelligence and Big Data at Chengdu College of Arts and Sciences, Chengdu, China. She graduated from Sichuan Normal University with a Bachelor's Degree in Computer Science and Technology in 2002. Her research interests include data mining, computer software and hardware and their applications.