GroupTrack: A Feature-aware Enhanced Group Tracking Method

Minjin Zhao
School of Aerospace Engineering
Beijing Institute of Technology
Beijing,China
3120240082@bit.edu.cn

Luheng Cui
School of Aerospace Engineering
Beijing Institute of Technology
Beijing,China
3120240064@bit.edu.cn

Shupeng Guo
School of Aerospace Engineering
Beijing Institute of Technology
Beijing,China
gsp7911@163.com

Chujun Liu
School of Aerospace Engineering
Beijing Institute of Technology
Beijing,China
harold0617@163.com

Yuxuan Jiang
School of Aerospace Engineering
Beijing Institute of Technology
Beijing,China
2375029263@qq.com

Yan Ding*
School of Aerospace Engineering
Beijing Institute of Technology
Beijing,China
dingyan@bit.edu.cn

***Abstract*—Group target tracking in sequential images is crucial for applications like public security. This paper proposes GroupTrack, a novel group multi-object tracking method based on temporal awareness and multi-modal association. The framework comprises a Feature-aware Enhanced Group Detection Network (FEGDN) and a Multi-Dimensional Similarity-based association strategy (MDS). FEGDN enhances robustness through multi-scale perception and bidirectional feature fusion, while MDS achieves stable trajectory association by fusing appearance, IoU, and scale consistency. Experiments on the dataset from the "National Big Data and Computational Intelligence Challenge" show that GroupTrack achieves a state-of-the-art average score of 88.18%, significantly outperforming advanced baselines. It secured the first prize in the competition, demonstrating its advancement and practical utility.**

Keywords—Group Multi-Object Tracking, Sequential Images, Feature-aware Enhancement, Multi-Dimensional Similarity, Trajectory Association

I. INTRODUCTION

Group target tracking in sequential images is a critical technology with extensive applications in public security surveillance, abnormal event early warning, and crowd behavior analysis [1, 2]. By leveraging visual perception and multi-source information fusion, it enhances the capability to perceive group activities, thereby providing crucial support for predicting potential security incidents [3].

Despite its significance, traditional multi-object tracking (MOT) algorithms face substantial challenges when dealing with group targets in real-world scenarios [4, 5]. Firstly, algorithmic stability is often compromised by complex environmental factors. Variations in illumination, weather conditions, and drastic changes in crowd density can severely degrade the accuracy of feature extraction and target association [6]. Secondly, temporal variations pose a significant hurdle. The appearance and motion characteristics of group targets (e.g., merging, splitting, occlusion) evolve over time, making it difficult for simple motion models or appearance cues alone to maintain consistent identity association across frames [7, 8]. These limitations restrict the practical deployment of existing tracking systems.

To address the aforementioned challenges, this study aims to propose a group multi-target tracking method named GroupTrack, based on temporal awareness and multi-modal association. The core work of this paper involves designing a detection network that incorporates multi-scale feature fusion and a group target aggregation strategy to frame group targets. Subsequently, by constructing a hybrid similarity association mechanism that integrates appearance, Intersection over Union, and scale consistency, the system achieves stable and robust group multi-object tracking , providing technical support for intelligent decision-making in areas such as public security early warning.

The structure of this paper is as follows: Chapter 2 reviews related work; Chapter 3 elaborates in detail on the design of the feature extraction model based on multi-scale encoding; subsequently, Chapter 4 presents the trajectory update model guided by spatio-temporal information; Chapter 5 showcases the detailed experimental setup and result analysis; finally, Chapter 6 concludes the paper and outlines future work. The overall organizational structure of the paper is illustrated in the Fig. 1.

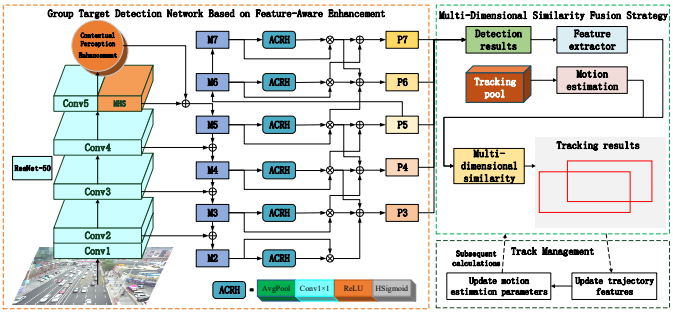

Fig. 1. Overall Architecture of the GroupTrack Algorithm

II. Literature Review

Group Multi-object Tracking aims to detect, associate, and predict the trajectories of multiple targets exhibiting group behavior (such as dense crowds, vehicle fleets, and drone swarms) in video sequences. Its core challenges lie in handling frequent occlusions between targets, appearance similarities, and the dynamic changes in group structures. Existing research paradigms primarily focus on trajectory update strategies and group modeling mechanisms, aiming to enhance the algorithm's robustness and accuracy in complex environments.

Regarding trajectory update and association strategies, most early methods followed the SORT [9] (Simple Online and Real-time Tracking) paradigm, utilizing the Kalman filter to predict target states and combining it with the Hungarian algorithm for cross-frame association. However, the linear motion model struggles to adapt to the non-linear motion patterns of group targets, and the independent matching strategy is prone to causing ID switches when dealing with dense targets. To address this, researchers introduced multi-feature fusion mechanisms to enhance association robustness. For instance, DeepSORT introduced appearance features to assist in matching, effectively mitigating identity confusion issues in occlusion scenarios [10]. ByteTrack enhanced the ability to maintain low-confidence targets by incorporating low-score detection boxes for supplementary updates [11]. Additionally, some work integrated motion interaction models [12] or scene geometric constraints [13], leveraging the cooperative relationships between targets to correct prediction deviations in individual trajectories.

Group modeling and structural perception are the core distinctions between group tracking and traditional tracking. To address the problem of the number of targets within a group and their spatial distribution changing over time, researchers have proposed the Random Finite Set (RFS) framework [14]. This framework models the group state as a random set and estimates the group centroid and the states of internal members through a multi-Bernoulli filter, thereby avoiding explicit data association [15]. In the field of computer vision, Graph Neural Networks (GNNs) are also employed to model the internal structure of groups. By modeling nodes and edges, GNNs capture the spatial proximity and motion consistency between targets, thereby assisting in resolving association ambiguities in dense intersection scenarios [16].

Focusing on the problem of accurately tracking group targets in sequential images, this paper highlights the methods of group-aware detection aggregation and multi-dimensional information fusion for trajectory association. Specifically, this study proposes a group multi-object tracking method named GroupTrack. The method first designs a detection network that incorporates multi-scale feature fusion to enhance target perception robustness in complex environments. Subsequently, it introduces a group target aggregation strategy based on dynamic adjacency judgment to generate compact group target representations. Finally, it constructs a hybrid similarity association mechanism that fuses appearance, distance, and scale consistency to achieve stable cross-frame target association and trajectory updates. This enables stable and efficient group multi-object tracking in complex, temporally varying scenarios, providing reliable technical support for abnormal event early warning.

III. Group Target Detection Network Based on Feature-Aware Enhancement

A. Feature-Aware Enhanced Backbone Network

This paper designs a multi-head perceptual structure, bridging between the fourth and fifth convolutional stages of the ResNet-50 feature extraction network [17], to enhance the feature maps output by the fifth stage. Additionally, a context information perception enhancement module is designed to receive the feature maps output by the fifth stage for spatial feature aggregation, further improving the network's feature extraction capability.

Fig. 2 illustrates the structural diagram of MHS-Conv5. The structure of the multi-head perceptual module is similar to that of Conv5, utilizing three perceptual blocks. Each block consists of two 1×1 convolutional layers and a multi-head perceptual layer. Here, B represents the number of input feature maps in the current training batch, C denotes the number of channels, and W and H represent the width and height of the feature maps, respectively. h denotes the number of heads. In this work, 4 heads are used. R_{hw} denotes the learnable positional encoding parameters. ⊕ and ⊗ represent element-wise addition and matrix multiplication, respectively. W_Q, W_K, and W_V represent three 1×1 pointwise convolutions, which produce the q, k, and v feature groups. The different heads of the q and k features are multiplied respectively, normalized via the Softmax function, and then multiplied with the v features. Subsequently, the head features are reintegrated to obtain the new image features that have undergone global perception.

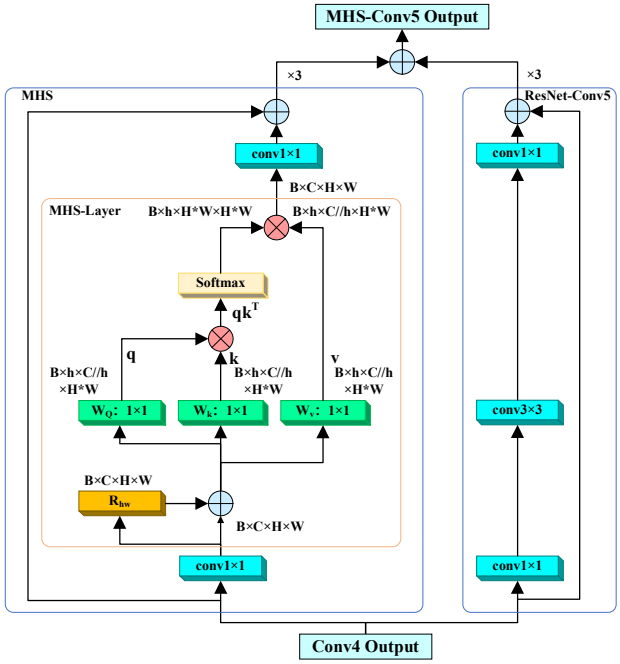

Fig. 2. Structural Diagram of the MHS-Conv5 Module

The Contextual Information Perception Enhancement Module proposed in this paper is shown in Fig. 3. First, multi-

scale adaptive average pooling is performed on the feature map output by MHS-Conv5 to generate contextual features at multiple different scales ($C\times\lambda_1W\times\lambda_1H$, $C\times\lambda_2W\times\lambda_2H$, $C\times\lambda_3W\times\lambda_3H$). In this algorithm, these three scale contextual features are generated by setting specific parameters. Subsequently, each feature is channel-aligned using a 1×1 convolution followed by upsampling. The three feature maps are then concatenated and fed into the Contextual Weight Perception Module. This module employs a 1×1 convolution, ReLU non-linear activation, a 3×3 convolution, and a Sigmoid mapping to generate adaptive contextual weights for each feature map. Finally, the adaptive spatially aggregated feature map is obtained through weighted fusion.

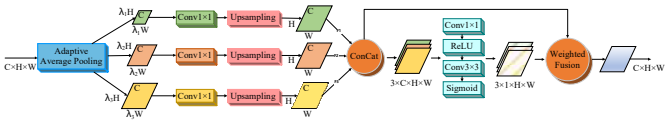

Fig. 3. Structural Diagram of the Contextual Information Perception Enhancement Module

B. *Cross-Scale Feature Bidirectional Fusion Neck Network*

Considering that the single-layer feature map output by the backbone network lacks the capacity for joint feature representation, this paper improves upon the Feature Pyramid Network (FPN) and designs a Cross-Scale Feature Bidirectional Fusion Network.

Based on the output features of the backbone network—Conv2, Conv3, Conv4, MHS-Conv5, and the context-aware enhanced feature—the network utilizes 1×1 convolutions to integrate channels and reconstructs four new lateral transitional features: M2, M3, M4, and M5. M6 and M7 are subsequently derived from the output feature map P5. This feature map is then fed into the constructed ACRH module, which possesses varying receptive fields, to obtain features weighted by scale-aware mechanisms. These weighted features are then dynamically aggregated with features from other levels in a cross-scale manner. Ultimately, five multi-scale feature maps—P3, P4, P5, P6, and P7—are obtained for the object detection head.

The scale information perception module ACRH established in this paper performs adaptive average pooling on the input features, as shown in Fig. 1, to obtain scale weights. These weights are then reduced to a single channel via a 1×1 convolution, nonlinearly activated by a ReLU function, and finally mapped to the [0, 1] interval using the HSigmoid function. The HSigmoid activation function approximates the Sigmoid function. Its value range is [0, 1], the slope is 1/6, and its mathematical expression is:

$$f^{\mathrm{HSigmoid}}(x)=\begin{cases}1,(x>3)\\ x/6+0.5,(3\geq x\geq -3)\\ 0,(x<-3)\end{cases} \tag{1}$$

C. *Group Target Aggregation Strategy Based on Dynamic Adjacency Judgment*

To aggregate discrete individual detection boxes into semantically complete group targets, this paper proposes an aggregation strategy based on dynamic adjacency judgment. The core of this strategy is to perform clustering based on the spatial proximity and scale consistency between individuals. Its pseudo-code is shown in Algorithm 1.

Algorithm 1: Pseudo-code of Group Detection Bounding Box Aggregation

Input: Detection results list $D=\{d_i\}_{i=1}^{N}$, where $d_i=(x_1,y_1,x_2,y_2,s,c)$

Output: Aggregated group bounding box list D'

Group D by target class c .
for $d_a,d_b\in D_c$ **do**:
  $d_{center}=\left\|\left(\frac{x_1^a+x_2^a}{2},\frac{y_1^a+y_2^a}{2}\right)-\left(\frac{x_1^b+x_2^b}{2},\frac{y_1^b+y_2^b}{2}\right)\right\|$
  if $d_{center}<\frac{y_2^a-y_1^a}{2}+\frac{y_2^b-y_1^b}{2}+T_d$ **then**
    $G\leftarrow d_a\cup d_b$
  end
end
for each formed group $G_k=\{d_i\}$ **do**
  $d_{G_i}=\left(min(x_1^i),min(y_1^i),max(x_2^i),max(y_2^i),\frac{1}{|G_k|}\sum s_i,\frac{1}{|G_k|}\sum c_i\right)$
  $D'\leftarrow D'\cup d_{G_i}$
end
Return D'

The algorithm employs dynamic adjacency judgment criteria, allowing the merging threshold to adaptively adjust based on the target's own height. This ensures spatial continuity in the vertical direction and effectively mitigates the problem of missed merges caused by partial occlusions or posture changes between targets. Through this strategy, the individual targets D output by the detection network are organized into group targets (denoted as D'), providing robust input for subsequent tracking association.

IV. Target Association Strategy Based on Multi-Dimensional Similarity Fusion

To achieve stable association between targets in adjacent frames, this study comprehensively considers the target's appearance features, spatial overlap relationship, and size consistency in the tracking strategy, constructing a multi-dimensional similarity measurement method.

A. *Appearance Similarity Modeling Based on ViT Features*

Appearance similarity, serving as a crucial metric for evaluating the visual consistency of targets, plays a decisive role in solving association problems in complex scenarios such as occlusion, intersection, and high target density. In this work, the target region is obtained based on the aggregated detection box positions. A Vision Transformer (ViT) network [18] is introduced as the feature extractor to extract target features, thereby obtaining a high-dimensional appearance feature vector

representation for each target. Finally, cosine similarity is employed to quantify the appearance consistency between targets. The calculation formula is as follows:

$$S_{\text{app}}(i,j)=\frac{F_i^t \cdot F_j^{t+1}}{\left\|F_i^t\right\|_2 \cdot \left\|F_j^{t+1}\right\|_2+\varepsilon} \tag{2}$$

where F_i^t represents the appearance feature of tracked target i in frame t, and F_j^{t+1} represents the appearance feature of detected target j in frame $t+1$. $\|\cdot\|_2$ denotes the L_2 norm of a vector, and ε is a very small constant introduced to prevent division by zero. A larger similarity value indicates that the two targets are more closely distributed in the appearance feature space, and thus are more likely to correspond to the same target.

B. Spatial Overlap Similarity Modeling Based on IoU

Motion continuity of targets across the temporal dimension is also an important prior for effective target association. The spatial position change of the same target is relatively smooth between consecutive frames. Therefore, this paper introduces a spatial similarity metric based on Intersection over Union (IoU) to characterize the consistency of target spatial positions between adjacent frames. Let the bounding box of track target i in frame t be B_i^t, and the bounding box of detection target j in frame $t+1$ be B_j^{t+1}. The IoU between them is defined as:

$$S_{\text{iou}}(i,j)=\frac{\text{Area}\left(B_i^t \cap B_j^{t+1}\right)}{\text{Area}\left(B_i^t \cup B_j^{t+1}\right)} \tag{3}$$

C. Target Size Similarity Modeling

In addition to appearance and spatial information, the scale of the target is also a crucial criterion for target association. The actual physical size of the same target changes relatively smoothly between consecutive frames. Therefore, modeling the target size can effectively suppress erroneous matches between targets with excessively large scale differences. The target size information is described jointly by the width and height of the detection bounding box. Let the bounding box width and height of track target i in frame t be (w_i^t, h_i^t), and those of detection target j in frame $t+1$ be (w_j^{t+1}, h_j^{t+1}). The size similarity is defined as:

$$S_{\text{scale}}(i,j)=1-\frac{1}{2}\left(\frac{|w_i^t-w_j^{t+1}|}{w_i^t+w_j^{t+1}}+\frac{|h_i^t-h_j^{t+1}|}{h_i^t+h_j^{t+1}}\right) \tag{4}$$

D. Multi-dimensional Similarity Fusion

This paper integrates the three aforementioned similarity measures to construct a multi-dimensional similarity metric. Its expression is as follows:

$$S(i,j)=w_1 S_{\text{app}}(i,j)+w_2 S_{\text{iou}}(i,j)+w_3 S_{\text{scale}}(i,j) \tag{5}$$

where w_1, w_2 and w_3 are weighting coefficients used to balance the relative importance of different similarity terms in target association.

To determine the optimal weights in the hybrid similarity metric, a parameter sensitivity analysis is performed. As shown in Fig. 4, a grid search in the 3D parameter space (with the constraint $w_1+w_2+w_3=1$) reveals a performance peak at (0.4, 0.3, 0.3), yielding the highest tracking accuracy of 88.18%. The analysis indicates that the appearance similarity weight w_1 has the most significant impact on performance. Appropriately increasing its value to 0.4 effectively leverages the appearance features of targets. Setting both the spatial similarity weight w_2 and the size similarity weight w_3 to 0.3 strikes a balance between the target motion model and scale variations. This parameter combination demonstrates good robustness and generalization capability across multiple test sequences.

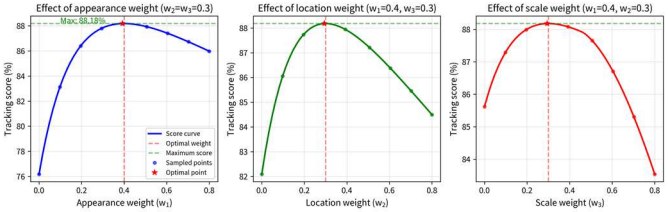

Fig. 4. Visualization Examples of the GroupTrack Algorithm

Furthermore, to address changes in target motion and attitude, this paper employs a Kalman filter to estimate the target's motion state and constructs a trajectory feature pool to store the trajectory's feature information. An exponential moving average mechanism is used during feature update, with the following formula:

$$f_{new}=\alpha \cdot f_{track}+(1-\alpha)\cdot f_{det} \tag{6}$$

V. Experimental Results and Analysis

A. Experimental Details

The dataset used in this experiment was provided by the "Group Target Tracking Based on Sequential Images" challenge in the "National Big Data and Computational Intelligence Challenge." It consists of 9 sequences with 1,150 training images and 4 sequences with 168 test images.

The evaluation criteria primarily focus on the accuracy of target identification and target tracking. The metric for measuring target identification accuracy is IDF1 (Identification F1). This metric evaluates the ability to preserve target identities and is defined as the harmonic mean of Identification Precision (IDP) and Identification Recall (IDR), calculated as follows:

$$IDF1=\frac{2}{\frac{1}{IDP}+\frac{1}{IDF}}=\frac{2IDTP}{2IDTP+IDFP+IDFN} \tag{7}$$

where IDTP refers to the number of positive samples correctly predicted as positive by the model, IDFP refers to the number of negative samples incorrectly predicted as positive by the model, and IDFN refers to the number of positive samples incorrectly predicted as negative by the model.

The metric for measuring target tracking accuracy is MOTA (Multiple Object Tracking Accuracy), which assesses

overall tracking accuracy by comprehensively considering target omissions (False Negatives, FN), target misclassifications (False Positives, FP), and identity switches (ID Switch, IDSW) within the group. The calculation method is as follows:

$$MOTA = 1 - \frac{FN + FP + IDSW}{GT} \tag{8}$$

where GT (Ground Truth) represents the total number of occurrences of tracked group targets in all images (the sum of the number of ground-truth targets in each image).

The final score is the average of the target identification accuracy and the target tracking accuracy, calculated as follows:

$$Score = \frac{IDF1 + MOTA}{2} \tag{9}$$

B. Comparison with State-of-the-Art and Visualization

To comprehensively evaluate the performance of the GroupTrack method proposed in this paper, experiments selected several advanced models that have demonstrated outstanding performance in the field of multi-object tracking in recent years as comparative baselines. These include FairMOT [19], ByteTrack [11], MOTR [20], and MOTRv2 [21]. The quantitative comparison results of the models on the officially provided test set are shown in TABLE I. .

The data in the table show that the GroupTrack proposed in this paper achieves the highest final scores on all four test sequences (ID: 65005, 65006, 65007, 65008), with an average accuracy of 88.18%, significantly outperforming the other comparative methods. Specifically, in the most challenging sequence, 65005, our method achieves an accuracy of 75.51%, substantially ahead of the second-best method, MOTR, which achieved 54.84%, demonstrating the strong robustness of the algorithm in complex scenarios. GroupTrack also maintains leading performance in the remaining sequences, particularly achieving excellent results of 94.23% and 92.92% on sequences 65007 and 65008, respectively, fully validating the effectiveness of the proposed algorithm.

This method ultimately contributed to our team winning the first prize in the "Group Target Tracking Based on Sequential Images" challenge at the "National Big Data and Computational Intelligence Challenge," proving its advancement and practical value in addressing real-world application demands.

TABLE I. COMPARISON WITH STATE-OF-THE-ART

Methods	65005(%)	65006(%)	65007(%)	65008(%)	Ave(%)
fairmot	41.07	88.40	85.34	83.39	74.55
ByteTrack	45.63	89.89	79.25	85.11	74.97
MOTR	54.84	51.31	73.07	77.60	64.20
MOTRv2	45.86	69.42	91.60	66.53	68.35
GroupTrack(ours)	75.51	90.07	94.23	92.92	88.18

The visualization results are shown in Fig. 5. In the left image, the algorithm successfully frames two adjacent pedestrians as a single group target and continuously tracks their movement trajectory in the right image. Even when partial occlusion occurs between the targets, stable bounding boxes and ID labels are maintained, demonstrating the effectiveness of the group aggregation strategy in dense scenarios.

Fig. 5. Visualization Examples of the GroupTrack Algorithm

C. Ablation Study

To verify the effectiveness of each core module within the framework proposed in this paper, we conducted ablation experiments, and the results are shown in TABLE II. . The Feature-aware Enhanced Group Detection Network (FEGDN) significantly improved the robustness of target perception, achieving an accuracy of 96.57% on the most challenging sequence, 65007, thereby providing high-quality input for subsequent association. The Multi-Dimensional Similarity-based target association strategy (MDS) is key to enhancing tracking continuity. Its fusion strategy effectively alleviates target association difficulties, enabling the system to achieve an average accuracy of 86.01% even without FEGDN. Finally, the combination of FEGDN and MDS constitutes the complete GroupTrack, achieving an average accuracy of 88.18%. This

result significantly outperforms any single module, verifying the synergistic necessity of high-quality detection and robust association strategies. It also demonstrates that the proposed method can effectively address the core challenges in group target tracking within sequential images.

TABLE II. ABLATION STUDIE ON THE FEGDN AND MDS MODULES

FEGDN	MDS	65005(%)	65006(%)	65007(%)	65008(%)	Ave(%)
✓		51.04	89.66	96.57	83.99	80.32
	✓	67.18	88.81	94.39	93.64	86.01
✓	✓	75.51	90.07	94.23	92.92	88.18

VI. CONCLUSION

This paper addresses the challenges of tracking stability and association arising from complex environmental changes, dense group targets, and time-varying individual features in sequential images. It proposes a group multi-object tracking method based on temporal awareness and multi-modal association. The core of this method lies in the design of a feature-aware enhanced group detection network, which achieves robust bounding of group targets. Additionally, by constructing a multi-dimensional similarity association mechanism that integrates appearance, Intersection over Union and scale consistency, accurate matching of target states across frames and trajectory updates are accomplished, effectively overcoming the shortcomings of traditional methods in group representation and long-term association.

Experimental results on the dedicated dataset provided by the "National Big Data and Computational Intelligence Challenge" demonstrate that GroupTrack achieves leading performance on key metrics. The final average accuracy reaches 88.18%, significantly surpassing advanced baseline methods such as FairMOT and ByteTrack, with a particularly notable advantage on the most challenging sequence, 65005. This validates the effectiveness of the proposed group detection aggregation strategy and hybrid association mechanism in enhancing overall tracking accuracy and improving identity consistency. The method ultimately secured first prize in the competition, proving its advancement and practical value in real-world applications.

Future work will focus on further optimizing the model's computational and memory efficiency to meet the demands of edge devices and large-scale real-time monitoring scenarios. It will also explore deeper integration of temporal context and richer scene semantic information to enhance automated analysis capabilities for group abnormal behaviors and complex interaction patterns.

REFERENCES

[1] LEAL-TAIXÉ L, MILAN A, REID I, et al. MOTChallenge: A Benchmark for Single-Camera Multiple Target Tracking[J]. International Journal of Computer Vision, 2021, 129(4): 845-881.

[2] WANG X, SUN Z, CHEHRI A, et al. Deep learning and multi-modal fusion for real-time multi-object tracking: Algorithms, challenges, datasets, and comparative study[J]. Information Fusion, 2024, 105: 102247.

[3] YANG C, XU M, LIANG X, et al. A Review of Resolvable Group Target Tracking Technology Based on Random Finite Set Filters[J]. Journal of Signal Processing, 2024, 40(10): 1763-1772.

[4] WOJKE N, BEWLEY A, PAULUS D. Simple Online and Realtime Tracking with a Deep Association Metric[C]// 2017 IEEE International Conference on Image Processing (ICIP), 2017: 3645-3649.

[5] ZHANG Y, SUN P, JIANG Y, et al. ByteTrack: Multi-Object Tracking by Associating Every Detection Box[C]// European Conference on Computer Vision, 2022: 1-21.

[6] KALAKE L, WAN W, HOU L. Analysis Based on Recent Deep Learning Approaches Applied in Real-Time Multi-Object Tracking: A Review[J]. IEEE Access, 2021, 9: 32650-32671.

[7] ZENG F, DONG B, ZHANG Y, et al. MOTR: End-to-End Multiple-Object Tracking with Transformer[C]// European Conference on Computer Vision, 2022: 659-675.

[8] ZHANG Y, WANG C, WANG X, et al. FairMOT: On the Fairness of Detection and Re-identification in Multiple Object Tracking[J]. International Journal of Computer Vision, 2021, 129(11): 3069-3087.

[9] Bewley A, Ge Z, Ott L, et al. Simple Online and Realtime Tracking[C]// 2016 IEEE International Conference on Image Processing (ICIP), 2016: 3464-3468.

[10] Wojke N, Bewley A, Paulus D. Simple Online and Realtime Tracking with a Deep Association Metric[EB/OL]. arXiv, 2017.

[11] Zhang Y, Sun P, Jiang Y, et al. ByteTrack: Multi-Object Tracking by Associating Every Detection Box[EB/OL]. arXiv, 2022.

[12] Wang X, Sun Z, Chehri A, et al. Deep learning and multi-modal fusion for real-time multi-object tracking: Algorithms, challenges, datasets, and comparative study[J]. Information Fusion, 2024, 105: 102247.

[13] Rakai L, Song H, Sun S, et al. Data association in multiple object tracking: A survey of recent techniques[J]. Expert Systems with Applications, 2022, 192: 116300.

[14] Yang C, Xu M, Liang X, et al. A Review of Resolvable Group Target Tracking Technology Based on Random Finite Set Filters[J]. Journal of Signal Processing, 2024, 40(10): 1763-1772.

[15] Xue X, Wu J, Huang S, et al. A Review of UAV Swarm Target Tracking Methods[J]. Tactical Missile Technology, 2025(5): 95-110, 140.

[16] Kalake L, Wan W, Hou L. Analysis Based on Recent Deep Learning Approaches Applied in Real-Time Multi-Object Tracking: A Review[J]. IEEE Access, 2021, 9: 32650-32671.

[17] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[18] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[EB/OL]. arXiv, 2021.

[19] Zhang Y, Wang C, Wang X, et al. FairMOT: On the Fairness of Detection and Re-identification in Multiple Object Tracking[J]. International Journal of Computer Vision, 2021, 129(11): 3069-3087.

[20] Zeng F, Dong B, Zhang Y, et al. MOTR: End-to-End Multiple-Object Tracking with Transformer[C]// European Conference on Computer Vision, 2022, 13687: 659-675.

[21] Zhang Y, Wang T, Zhang Y, et al. MOTRv2: Bootstrapping End-to-End Multi-Object Tracking by Pretrained Object Detectors[C]// 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.