symmetry

MDPI

Article

Cross-Camera Tracking Model and Method Based on Multi-Feature Fusion

Peng Zhang [1,†], Siqi Wang [1,†], Wei Zhang [1,*], Weimin Lei [1], Xinlei Zhao [2], Qingyang Jing [1] and Mingxin Liu [1]

[1] School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China; 1910609@stu.neu.edu.cn (P.Z.); 2172018@stu.neu.edu.cn (S.W.); leiweimin@mail.neu.edu.cn (W.L.); 2110649@stu.neu.edu.cn (Q.J.); 2310710@stu.neu.edu.cn (M.L.)
[2] Shenyang Er Yi San Electronic Technology Co., Ltd., Shenyang 110023, China; zhaoxinlei@china213.net
* Correspondence: zhangwei1@mail.neu.edu.cn
† These authors contributed equally to this work.

Abstract: Multi-camera video surveillance has been widely applied in crowd statistics and analysis in smart city scenarios. Most existing studies rely on appearance or motion features for cross-camera trajectory tracking, due to the changing asymmetric perspectives of multiple cameras and occlusions in crowded scenes, resulting in low accuracy and poor tracking performance. This paper proposes a tracking method that fuses appearance and motion features. An implicit social model is used to obtain motion features containing spatio-temporal information and social relations for trajectory prediction. The TransReID model is used to obtain appearance features for re-identification. Fused features are derived by integrating appearance features, spatio-temporal information and social relations. Based on the fused features, multi-round clustering is adopted to associate cross-camera objects. Exclusively employing robust pedestrian reidentification and trajectory prediction models, coupled with the real-time detector YOLOX, without any reliance on supplementary information, an IDF1 score of 70.64% is attained on typical datasets derived from AiCity2023.

Keywords: multi-camera tracking; trajectory prediction; appearance features; spectral clustering

Citation: Zhang, P.; Wang, S.; Zhang, W.; Lei, W.; Zhao, X.; Jing, Q.; Liu, M. Cross-Camera Tracking Model and Method Based on Multi-Feature Fusion. *Symmetry* **2023**, *15*, 2145. https://doi.org/10.3390/sym15122145

Academic Editors: Dmitri Donetski and Aristeidis Tsitiridis

Received: 23 October 2023
Revised: 21 November 2023
Accepted: 29 November 2023
Published: 2 December 2023

1. Introduction

Multi-camera surveillance systems have found extensive applications in various domains such as urban traffic regulation, public safety monitoring and crowd behavior analysis. Nevertheless, there are considerable challenges that remain to be addressed. Most importantly, the task of manually monitoring the voluminous target data from multiple cameras is considerably immense. Subsequently, due to factors such as different asymmetric perspectives of cameras, lighting conditions and obstructions, achieving accurate multi-target multi-camera tracking continues to confront formidable challenges.

Multi-target multi-camera tracking consists of the following steps:

1. Pedestrian detection;
2. Single-camera tracking;
3. Pedestrian Re-ID feature extraction;
4. Cross-camera association.

First, the pedestrians are located in each frame of the surveillance video, then the objects are tracked to obtain short-term trajectories. Then, reidentification (Re-ID) features are extracted using the Re-ID model for cross-camera trajectory association. In theory, accurate pedestrians Re-ID can solve the whole task. However, the actual tracking performance is affected by various environmental factors such as ambiguity due to object occlusion.

In recent years, cross-camera tracking methods based on occluded Re-ID models have become the subject of increased attention among researchers. Such methods introduce auxiliary modules such as human body analysis and pose estimation on the basis of

occluded Re-ID models [1–4]. This strategy allows the system to perform more accurate feature matching in the non-occluded core area, thereby reducing the matching error caused by occlusion and further improving the matching accuracy. Although substantial progress has been made in the development of these methods, the introduced auxiliary modules undoubtedly increase the amount of computation and model complexity, which poses some challenges for actual deployment. In addition, Re-ID models themselves still have some inherent limitations in actual surveillance systems:

In the overhead view of surveillance, different objects with similar clothing may have high similarity in appearance information, and the Re-ID model does not perform well in distinguishing objects with similar appearances.

Due to the different angles of cameras and different backgrounds, the appearance information of the same object may vary greatly in different surveillance views, making it difficult for the Re-ID model to match the same object across views.

In crowded public places, there may be situations where objects overlap with each other or are partially occluded by obstacles, resulting in incomplete or distorted appearance features, which affects the matching accuracy of the model.

To address the above issues, this paper proposes an efficient cross-camera tracking model with multi-feature fusion. In this method, the implicit social model [5] predicts the position of objects in future frames based on their motion features in previous frames and the behavior patterns between objects, which helps to compensate for trajectory interruptions and ID switching issues caused by occlusion. To further enhance feature robustness, a Transformer-based Re-ID model is integrated to capture the relative positional relationships between objects more carefully. The trajectory prediction and Re-ID model output features are weighted and summed to obtain a fused feature representation. This feature is first used for temporal association within a single camera, and object association across different views is completed by a multi-round clustering strategy based on single-camera tracking results.

The main contributions of this paper include

- An implicit social model-based trajectory prediction is introduced to achieve trajectory prediction by simulating the interactions and motion patterns between objects, instead of traditional linear prediction models, making it more suitable for object tracking in crowded public scenes;
- A single-camera tracking model based on multi-feature fusion is proposed, which achieves more accurate associations between detection and trajectory by comparing the Intersection over Union of trajectory predictions and calculating the Euclidean distance similarity of Re-ID;
- A cross-camera association strategy is proposed which associates objects across cameras using spectral clustering based on single-camera tracking results, and validates the clustering results using appearance features to reduce incorrect associations.

2. Relevant Works

2.1. Trajectory Prediction

Trajectory prediction is an important component in a variety of real-world applications such as autonomous driving, robotics and smart cities. In these applications, a generative model is utilized to predict the future trajectory of an agent. Trajectory prediction models can be classified into two categories on the basis of the output that they produce. In the first category, the approach taken is to explicitly model the future as a continuous or discrete parameter distribution, e.g., S-LSTM [6] and S-STGCNN [7] use Gaussian distributions for modelling. Since the Gaussian distribution is unimodal, it cannot capture multimodal future trajectories. In contrast, PRECOG [8] and AgentFormer [9] can support discrete or continuous multimodal information by learning the object latent behavior distribution. In these studies, Gaussian distributions for trajectory prediction were generated based on the sampling of latent information. In the second category, the approach is to implicitly model the future as a non-parametric distribution, e.g., S-GAN [10], SoPhie [11] and S-BiGAT [12]

use a GAN architecture with inputs added to randomly sampled noise and trained by an adversarial loss mechanism to output deterministic trajectory predictions.

We use an implicit social trajectory prediction model based on convolutional neural networks to obtain an efficient trajectory prediction model by integrating both the personal information of the object's individual and social context into the model. The social context helps the model focus on the global object movement, dynamic relationships between objects and group behavior information, which enables more accurate prediction of the object's future actions.

2.2. Pedestrian Re-Identification

Pedestrian re-identification reduces tracking errors caused by occlusion and person-ID switching by comparing appearance information between objects. In order to solve the problem of occlusion resistance, methods for occlusion resistant pedestrian re-recognition can be basically classified into three categories: manual segmentation-based methods, additional cue-based methods and Transformer-based methods [2]. Manual segmentation-based methods chunk pedestrians into different regions and compare the different regions for similarity. Additional cue-based methods use additional auxiliary models to localize body parts, such as segmentation, pose estimation or body resolution. Although convolutional neural networks (CNNs) currently dominate the field of pedestrian re-recognition, their receptive field regions are small and downsampling operations reduce the spatial resolution of the output feature maps, which reduces the ability to discriminate between similar looking objects. In contrast, it is more appropriate to embed the attention mechanism of transformers in a deep network, since deep networks are more suitable for dealing with larger continuous regions and difficult to extract detailed features. The Transformer-based approaches introduce multi-head attention modules, remove convolution and downsampling and perform multi-scale feature fusion at different layers, which can better capture the details and global information of the image, with powerful feature extraction and global context capture.

In this paper, the TransReID-SSL [13] pedestrian re-identification model is chosen, because it extracts robust and discriminative Re-ID features.

2.3. Single-Camera Multi-Object Tracking

Single-camera multi-object tracking methods can be classified into separate detection and embedding (SDE) and joint detection and embedding (JDE). SDE first detects all the objects in a single frame and extracts their Re-ID features and then uses an association algorithm to associate the objects between the previous frames. This approach splits object tracking into two independent steps of detection and association, e.g., SORT [14] and DeepSORT [15]. JDE obtains both detection and embedding information, and then joins the appearance features or motion prediction to the detection frame by association matching, which achieves better tracking performance through lower computational cost, e.g., FairMOT [16] and ByteTrack [17]. ByteTrack uses only motion features for association matching, which has better real-time performance, but is prone to identity switching problems due to occlusion and does not track well in crowded public scenes.

Recently, Bot-SORT [18], which utilizes appearance and motion information, surpassed ByteTrack in tracking accuracy. In this paper, we replace the Kalman filtered trajectory prediction algorithm with a deep learning-based trajectory prediction model on the basis of ByteTrack, add Re-ID features for secondary matching and contribute to the model's improved tracking speed through Intersection over Union matching, as well as achieving high accuracy in similarity matching.

2.4. Multi-Target Multi-Camera Tracking

Multi-target multi-camera tracking (MTMCT) involves object detection, decentralized consensus estimation, appearance feature extraction via re-ID models, trajectory prediction and cross-camera trajectory association. Decentralized consensus-based estimation offers

a decentralized approach to target estimation that enhances the collaborative aspects of cross-camera tracking systems [19]. But the main problem of multi-target multi-camera tracking is the trajectory association problem between different views, which focus on how to reduce the search space and time for the same object association of different views.

Data fusion techniques among various sensors are typically categorized into two research streams: centralized tracking methods [20] and single-view tracking methods [21]. Centralized tracking methods combine data from various views to facilitate detection and tracking, specifically by composing a global occupancy graph for tracking based on the detection nodes of each time frame [22]. Its benefit is that it can utilize the multi-view information of the scene and reduce the effect of occlusion and noise in crowded environments. The limitation is that it requires accurate camera synchronization and geometrical relationships between cameras, and since the position of the tracked object is transmitted continuously every few frames, it requires the camera to cover the full field of view, and the amount of data transmitted is relatively large, which affects real-time performance. The single-view tracking method first detects and tracks each view individually, and then obtains the final object trajectory through data correlation and trajectory fusion, and does not rely on limited information sharing between different views [23]. The advantages of the single-view tracking method are that it does not require camera synchronization and precise geometrical relationships between cameras and is more suitable for real-time tracking; the disadvantages are that it cannot solve the serious occlusion problem, and its performance relies on the effect of single-camera detection. In this paper, we use the single view method to represent multi-view trajectories via a graph model and obtain cross-camera trajectory matching through spectral clustering.

3. Overall Framework for Cross-Camera Tracking

The overall framework diagram of cross-camera multi-object tracking based on fused features is shown in Figure 1, and the process of specific cross-camera tracking is as follows:

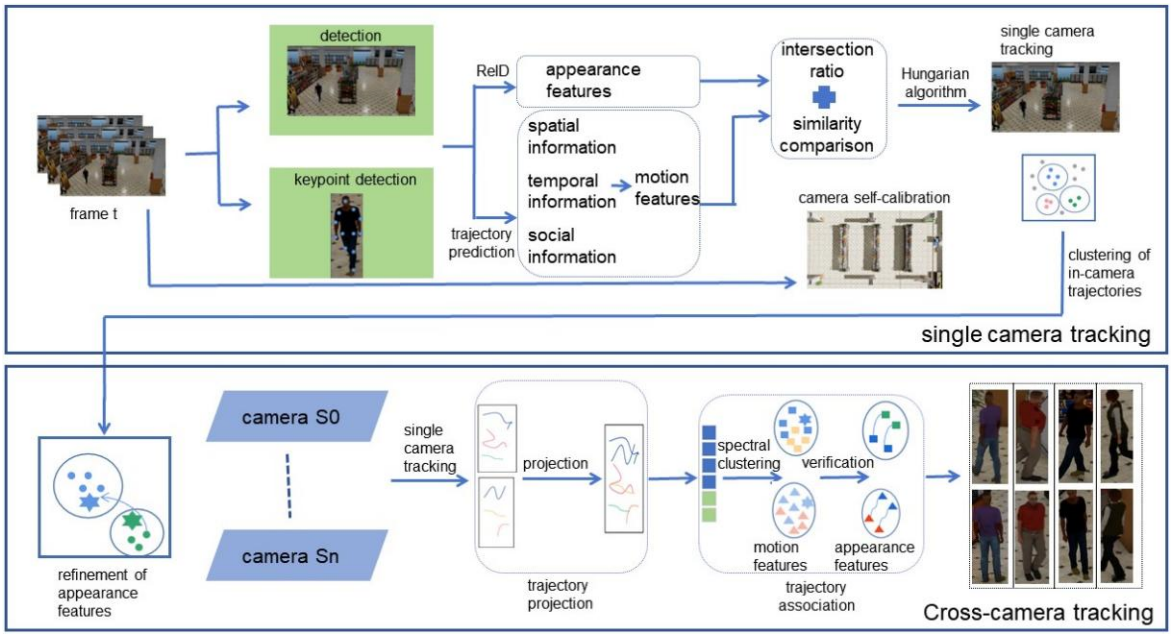

Figure 1. Overall framework diagram.

Firstly, camera self-calibration is performed to obtain the geometrical relationship between camera network viewpoints to get the homography matrix. The images from different camera viewpoints can be mapped to a common reference coordinate system through the homography matrix. In this paper, the homography matrix is used to integrate

the single-camera tracking trajectories obtained from different cameras into the same ground plane, which facilitates cross-camera multi-object tracking.

Next, single-camera tracking is performed. Tracking is done in an occlusion situation where the detection frames may fail, but the key points provide some important information which helps to maintain continuous tracking of the object. Person detection and key-point detection are first performed on consecutive frames to obtain spatial information about the object's location and temporal information between consecutive frames, while combining social information between different persons to obtain motion features. After person detection, the appearance features of the persons are obtained by the Re-ID module, followed by weighted fusion of the appearance and motion features. The implicit social model realizes the prediction of person trajectory through motion features and the temporal correlation of single-camera tracking through the Intersection over Union between the detection box and the trajectory prediction box of the next frame, while the appearance features are temporally correlated through the similarity comparison.

Eventually, the single-camera tracking trajectories obtained from different cameras are mapped onto the same plane by the projection of a homography matrix. Subsequently, all the trajectories on this plane are clustered and stacked based on the motion information to obtain the association of trajectories on different views. Finally, the cross-camera tracking results are obtained by verifying whether the matching is successful through appearance features.

The proposed model in this study reduces the frequency of appearance feature matching and eliminates reliance on additional auxiliary models, thereby reducing the overall complexity of the model. Target association through multiple rounds of clustering enhances the accuracy of the model in object tracking.

4. Cross-Camera Association Technique Based on Fused Features

The cross-camera tracking technology is based on a single view tracking method and incorporates camera calibration technology. This model uses a multi-feature extraction module for single-camera tracking and uses the Re-ID model based on Transformer to divide images into blocks, so as to extract more fine-grained appearance features with context connection. The trajectory prediction technology based on the implicit social model is also introduced to combine spatio-temporal information and social information according to the motion features to predict the position of the object in the next frame. Two feature weighting yields single-camera tracking results through the cost matrix.

In the cross-camera tracking phase, the homography matrix obtained by the camera calibration technology is used to project all the tracks onto a common plane for cluster matching. The matching results are then verified and re-matched using appearance features, and finally we get the cross-camera tracking results.

4.1. Implicit Social Modelling Based on Motion Information

ByteTrack introduces a two-step matching algorithm grounded in object detection frame thresholds, employed to track individual objects by harnessing trajectory prediction information. In this investigation, the ByteTrack framework replaces the conventional Kalman filter model with an implicit social model. The object's prospective positions in forthcoming frames are prognosticated through the implicit social model, and subsequently, the Intersection over Union between the predicted bounding box and the real detection box in successive frames are computed to ascertain correlation.

The implicit social model adopts implicit maximum likelihood estimation (IMLE) as its trajectory prediction mechanism [24]. IMLE, through a straightforward mechanism, trains the model by introducing additional noise into the model to predict multiple samples, selecting the sample closest to the real value in the distribution, and utilizing this sample for backpropagation during training. IMLE minimizes the model based on distance optimization.

The implicit social model clusters pedestrian trajectories based on the maximum speed changes observed in the trajectories. The training mechanism of the model, as depicted

in Algorithm 1, relies on the motion information of objects in the spatial domain between consecutive frames, framing it as a regression task with sequences as input and output. The motion state of an object, monitored from frame t_1 to frame t_{obs}, is represented as $m_{t_1:t_{\text{obs}}} = \{m_t | t \in [t_1, \ldots, t_{\text{obs}}]\}$, where $m_t \in R^{D*\text{obs}}$, with D signifying the dimension of the input motion state. In the context of this study, the model takes the (x, y) coordinates of an object's position as input, resulting in $D = 2$ dimensions. Neglecting to distinguish and collectively train pedestrians with disparate velocities may engender a bias towards swifter pedestrians in trajectory prediction, potentially misclassifying stationary individuals as mobile entities. To surmount this challenge, this research adopts a strategy of segregating social domains characterized by varying speeds.

Algorithm 1: Implicit social modelling algorithm

Input: Model $\theta(\cdot)$ and parameters $\alpha_1, \alpha_2, \alpha_3$
Output: Model θ with a distribution similar to the real samples

Initialize model parameters, dataset $D = \left(d_o^i, d_p^i\right)_{i=1}^{n}$
Perform clustering on D to obtain Z
Initialize the loss function $\mathcal{L}(\cdot)$
for e = 1 to Epochs **do**
Select a random batch (d_o, d_p) from D
Predicting independent identically distributed samples $\overline{d}_p^1, \ldots, \overline{d}_p^m$ from $\theta(d_o)$
Compute the loss function values $\mathcal{L}_{\text{triplet}}, \mathcal{L}_{\text{G-distance}}, \mathcal{L}_{\text{G-angle}}$
$\mathcal{L}(\cdot) = \| d_p - d_p^1 \|_1 + \alpha_1 \mathcal{L}_{\text{triplet}} + \alpha_2 \mathcal{L}_{\text{G-distance}} + \alpha_3 \mathcal{L}_{\text{G-angle}}$
$\sigma(i) \leftarrow \text{argmin}_i \mathcal{L}\left(d_p - \overline{d}_p^i\right), \forall i \in m$
$\theta \leftarrow \theta - \eta \nabla_\theta \sigma(i)$
end for
return θ

As shown in Figure 2, clustering pedestrians based on the maximum speed changes yields multiple social regions. With these social regions, a better understanding of social relationships between different pedestrians can be achieved. Each social unit deals with one social region, taking input as $\text{P} * \text{T}_0 * \text{N}$ and producing output as $\text{P} * \text{T}_\text{P} * \text{N}$, where P represents observed positions, T_0 and T_P are the lengths of the observation and prediction steps and N is the number of individuals within the social region. Training for pedestrian trajectory prediction is performed using maximum likelihood estimation.

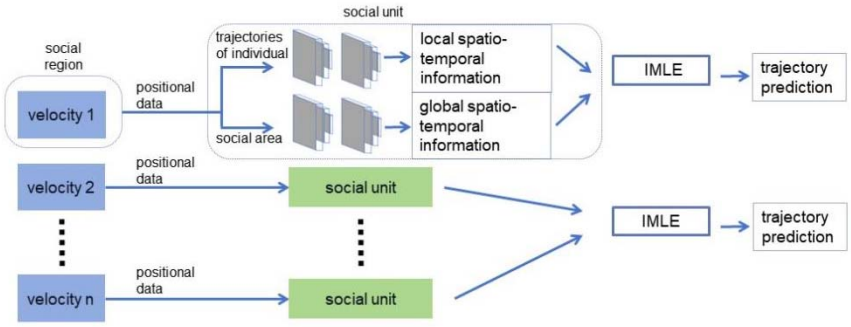

Figure 2. Implicit social model.

The trajectory prediction process is as follows: Firstly, clustering is performed based on the maximum speed change of pedestrians to obtain the position information of each pedestrian. Subsequently, this positional information is entered into the social unit, which is divided into two parts. One locally handles the trajectory of an individual and the other globally handles the trajectory relationships among all pedestrians in the social area. Both parts obtain spatio-temporal information via the employment of consecutive

residual-connected convolutional neural network (CNN) layers. Ultimately, the outcomes of trajectory prediction are realized through implicit maximum likelihood estimation:

$$V = w_{\mathrm{g}} * v_{\mathrm{g}} + w_{\mathrm{l}} * v_{\mathrm{l}} \tag{1}$$

where w_{g} and w_{l} represent the respective global and local weighting factors, while v_{g} and v_{l} embody the global and local spatio-temporal information. The resulting V signifies the anticipated future positional data of the object as predicted by the trajectory analysis.

4.2. Representation of Fusion Features

Within the context of single-camera tracking, the imperative task is to derive the corresponding trajectory information by means of associating the detection frames within the video sequence. In this research endeavor, we employ ByteTrack as our chosen tracking algorithm and adopt the implicit social model as our trajectory prediction model. To enrich the feature representation of the objects using the appearance features, we employ TransReID-SSL as the re-identification (Re-ID) model, which is dedicated to the extraction of said appearance features. The model is trained on a composite datasets encompassing Market-1501 [25], MSMT17 [26], CUHK-SYSU [27] and the AiCity2023. The model's weight initialization is founded upon the TransReID-SSL pre-training model and subsequently fine-tuned using input images sized at 256 × 128 pixels. The optimization process is orchestrated through the amalgamation of cross-entropy loss and ternary loss, with the cross-entropy loss function formulated as follows:

$$L_{\mathrm{ce}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{2}$$

y_i denotes the identification tag corresponding to the ith image, while N signifies the total number of images in the composite datasets. The formulation for the ternary loss is as follows:

$$L_{\mathrm{tri}} = \sum_{i=1}^{N}\left(\max\left(m + d(f_i^a, f_i^n) - d\left(f_i^a, f_i^p\right), 0\right)\right) \tag{3}$$

In the given equation, the variable d represents the L2 distance, while f_i^p signifies the positive samples and f_i^n denotes the negative sample. The parameter m represents the difference between ternary losses.

The fusion feature was adopted to integrate motion features with appearance features to enhance the tracking process. The motion features are used for nonlinear trajectory prediction, while the re-identification (Re-ID) model serves as a means for comparing appearance similarities. The synthesis of these two distinct feature sets yields a more precise approach to object tracking. To balance the weights of motion features and appearance features, the computation of the cost matrix for both is carried out by means of weighting. This total cost matrix, denoted as C, is expressed as follows:

$$C = \alpha A_{\mathrm{motion}} + (1 - \alpha) A_{\mathrm{appearance}} \tag{4}$$

In this equation, A_{motion} represents the motion cost matrix, $A_{\mathrm{appearance}}$ represents the appearance cost matrix and α is a weighting parameter constrained within the range of 0 to 1. α plays a pivotal role in determining the relative significance of appearance and motion costs when computing the overall cost. More specifically, as α approaches 1, the influence of motion costs becomes more pronounced, whereas as α approaches 0, the impact of appearance costs becomes more dominant.

The formula for the Intersection over Union related to motion costs can be expressed as

$$\mathrm{IoU}(A, B) = \frac{\mathrm{area}(A \cap B)}{\mathrm{area}(A \cup B)} \tag{5}$$

A and B in the above formula are the prediction and detection bounding boxes, respectively. In the total cost formula, the Euclidean formula for the appearance cost is expressed as follows:

$$\mathrm{dis}(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{6}$$

a and b represent the appearance feature vectors associated with two distinct objects and n signifies the dimensionality of the feature vector.

To complete the tracking process within a single-camera system, a Hungarian algorithm is applied to the cost matrix C. This algorithm effectively addresses the linear assignment problem, determining the optimal matching configuration, thereby facilitating single-camera trajectory tracking.

4.3. Cross-Camera Object Matching Based on Fused Features

Firstly, calibrating each camera determines the internal and external parameters of the cameras. Subsequently, the homography matrix for each camera was calculated using the calibration information. The homography matrix is used to establish the mapping between the camera coordinate and the global map. The method presented in this paper is to further realize across-camera multi-object tracking based on the single-camera tracking results. Trajectories of all persons were extracted from each camera view and then projected onto the global map using the homography matrix. The homography matrix H was calculated for the corresponding coordinate transformation by using the P_1' point in the given frame and the corresponding P_2' point in the global map:

$$P_2' = H\begin{bmatrix} c_x & b_y & 1 \end{bmatrix} \tag{7}$$

Here, c_x and b_y represent the center and bottom coordinates of the detection frame, respectively. The ensuing step involves the normalization of the global map coordinates for P_2' points to derive the trajectory positions within the global map:

$$P_2 = \frac{P_2'}{P_2'z} \tag{8}$$

The projection of trajectories onto the global map facilitates a comprehensive analysis of object motion within a standardized coordinate system. Subsequently, an affinity graph is constructed on the global map, denoted as $G = (V, E)$, where V signifies the set of trajectories, E signifies the set of feature-weighted edges connecting trajectories and the affinity matrix A can be formally defined as

$$A_{ij} = \exp\left(-\frac{\left|f_i - f_j\right|^2}{\sigma^2}\right) \tag{9}$$

Here, f_i and f_j signify the motion features associated with trajectories i and j, respectively, while σ represents the scaling factor.

The subsequent step involves the application of spectral clustering on the graph to partition trajectories into distinct clusters. This process requires the computation of the graph Laplacian matrix L, which is expressed as

$$L = D - A \tag{10}$$

D represents the degree matrix, where $D_{ii} = \sum_j A_{ij}$. Subsequently, the k smallest eigenvectors of L are extracted to form the matrix $U \in R^{n*k}$, with n denoting the number

of trajectories. To achieve relative weightings within matrix U, row normalization is performed, resulting in the formation of matrix T:

$$T_{ij} = \frac{U_{ij}}{|U_i|} \tag{11}$$

The rows of matrix T are then subjected to K-means clustering into k clusters. Each cluster corresponds to a distinct global object. To enhance the precision of tracking results, the trajectories are subjected to appearance feature verification, as depicted in Equation (3). Subsequently, objects that do not meet the appearance-matching criterion are subjected to re-clustering and re-matching procedures. This approach contributes to the attainment of accurate and efficient trajectory associations across various views, ultimately enhancing tracking performance.

5. Experimental Validation and Analysis

5.1. Experimental Data Set

The AI City Challenge 2023 Track 1 dataset is used in this paper, which contains both real and virtual synthetic data. This cross-camera multi-object tracking dataset includes a total of 130 cameras covering 1491 min of high-resolution (1920 × 1080) 30FPS video. The dataset is divided into 22 subsets, of which 10 are used for training, 5 for validation and 7 for testing. In addition, three publicly available Re-ID datasets (Market-1501, MTMC17, and CUHK-SYSU) were used to train the Re-ID model.

5.2. Evaluation Metrics

We use IDF1, IDP and IDR as evaluation metrics for cross-camera tracking. The IDF1 score is a metric employed to assess the performance of an object detection model. It quantifies the ratio of the average number of correct identifications to real objects and computed detections in a detection, which is formulated as follows:

$$\text{IDF1} = \frac{2 \times \text{IDTP}}{2 \times \text{IDTP} + \text{IDFP} + \text{IDFN}} \tag{12}$$

$$\text{IDP} = \frac{\text{IDTP}}{\text{IDTP} + \text{IDFP}} \tag{13}$$

$$\text{IDR} = \frac{\text{IDTP}}{\text{IDTP} + \text{IDFN}} \tag{14}$$

where IDTP is the number of true positive identities, IDFP is the number of false positive identities and IDFN is the total number of false negative identities. A higher IDF1 value signifies superior algorithmic performance in the context of multi-object tracking.

5.3. Experimental Details

The experimental operating system is Ubuntu 20.04.1, 64-bit OS, the graphics card is NVIDIA GTX3090Ti, the compiler setting is Python 3.8, Pytorch 1.11.0 deep learning framework is used as the experimental platform, YOLOX is used to generate the bounding box for object detection. The model is pre-trained using the COCO with a threshold value of 0.1. Image dimensions were consistently standardized to 1333 × 800 pixels, and the training process involved 300 epochs with a learning rate of 0.001 and a batch size of 8. The experiments were conducted using MMDetction toolbox to train the models and YOLOv7 pose estimation model and pre-trained model for key point estimation. An implicit social model is used for the trajectory prediction part, a TransReID model is used for the pedestrian re-identification module and a modified ByteTrack is used for single-camera tracking; based on this result, cross-camera tracking is performed.

5.4. Experimental Results and Analyses

We validate the proposed cross-camera tracking system on the 2023 AI City Challenge test dataset. As demonstrated in Table 1, a higher IDF1 score indicates better tracking performance, the system successfully achieves an IDF1 score of 0.7064. Compared to the ByteTrack baseline model, the IDF1 score is significantly improved by 11.88% by using an implicit social model for trajectory prediction. This result demonstrates that a nonlinear trajectory prediction model provides more accurate predictions for pedestrian tracking in public scenes compared to Kalman filtering in the baseline model. In addition, AI2023Team20 [28] further improved the IDF1 score by 2.31% by employing an FFT component to eliminate noise and extract key motion information.

Table 1. Performance comparison of different network models.

Network Models	IDF1 ↑	IDP ↑	IDR ↑
Baseline	0.4752	0.4989	0.4537
Baseline + TP	0.5940	0.6156	0.5738
AI2023Team20	0.6171	0.6392	0.5965
Our network	0.7064	0.6923	0.6532

To solve the occlusion problem, we introduce a fused feature-based strategy. The core of this strategy is cross-camera tracking assistance by incorporating appearance features embedded by Re-ID. As shown in Figure 3, the pedestrian reappears after complete occlusion, and the model is still able to accurately recognize the object with correct ID assignment. This approach effectively improves the accuracy of cross-camera tracking and significantly reduces the tracking error caused by identity switching. Ultimately, our model achieves 70.64% accuracy on the IDF1 evaluation metric.

Figure 3. Visualization of full occlusion tracking of objects.

A major advantage of the method proposed in this paper is its ability to address the problem of person occlusion without the need for high computational cost. The person occlusion problem is effectively dealt with by predicting the object's trajectory, and the accuracy of occluded tracking is verified through appearance similarity comparison while keeping the computational cost relatively low. Table 1 proves that the technical route of this paper is able to solve the problems in the baseline model.

6. Conclusions

In this study, we design a cross-camera tracking model based on fused features. To address common occlusion issues in cross-camera tracking, we introduced implicit social models and Re-ID techniques. Meanwhile, to reduce identity confusion errors caused by tracking mismatches in multi-pedestrian scenarios, we proposed a specific cross-camera tracking strategy. This strategy first performs preliminary clustering matching based on fused features, and further verifies using appearance features to reduce probability of clustering mismatches. Extensive experiments were conducted on complex scenes to validate the model. The experimental results demonstrate the superior performance of the

proposed model in handling occlusions, reducing identity switching errors and lowering computational costs. This indicates that our method not only enhances the accuracy and stability of multi-object tracking, but also provides an efficient and practical solution for real-world multi-object tracking applications.

Author Contributions: Conceptualization, P.Z.; methodology, S.W.; software, S.W.; validation, P.Z. and S.W.; formal analysis, P.Z.; investigation, S.W.; resources, P.Z. and W.L.; data curation, P.Z. and S.W.; writing—original draft preparation, S.W. and P.Z.; writing—review and editing, W.Z., Q.J. and M.L.; visualization, S.W. and X.Z.; supervision, W.L. and W.Z.; project administration, W.Z.; funding acquisition, W.L. and P.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the 'Jie Bang Gua Shuai' Science and Technology Major Project of Liaoning Province in 2022 (No. 2022JH1/10400025), the Fundamental Research Funds for the Central Universities of China (No. N2216010).

Data Availability Statement: The datasets and codes are available from the first and corresponding authors on reasonable requests. The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the project being currently in progress.

Acknowledgments: The authors wish to thank the editors and reviewers for their reviews and advices on this paper.

Conflicts of Interest: Author Xinlei Zhao was employed by the company Shenyang Er Yi San Electronic Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Wang, T.; Liu, H.; Song, P.; Guo, T.; Shi, W. Pose-Guided Feature Disentangling for Occluded Person Re-Identification Based on Transformer. *AAAI* **2022**, *36*, 2540–2549. [CrossRef]
2. Somers, V.; Vleeschouwer, C.D.; Alahi, A. Body Part-Based Representation Learning for Occluded Person Re-Identification. In Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2–7 January 2023; pp. 1613–1623.
3. Zhao, Y.; Zhu, S.; Wang, D.; Liang, Z. Short Range Correlation Transformer for Occluded Person Re-Identification. *Neural Comput. Appl.* **2022**, *34*, 17633–17645. [CrossRef]
4. Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; Chen, X. Feature Completion for Occluded Person Re-Identification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 4894–4912. [CrossRef] [PubMed]
5. Mohamed, A.; Zhu, D.; Vu, W.; Elhoseiny, M.; Claudel, C. Social-Implicit: Rethinking Trajectory Prediction Evaluation and the Effectiveness of Implicit Maximum Likelihood Estimation. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022.
6. Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Fei-Fei, L.; Savarese, S. Social LSTM: Human Trajectory Prediction in Crowded Spaces. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 961–971.
7. Mohamed, A.; Qian, K.; Elhoseiny, M.; Claudel, C. Social-STGCNN: A Social Spatio-Temporal Graph Convolutional Neural Network for Human Trajectory Prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
8. Rhinehart, N.; Mcallister, R.; Kitani, K.; Levine, S. PRECOG: PREdiction Conditioned on Goals in Visual Multi-Agent Settings. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2821–2830.
9. Yuan, Y.; Weng, X.; Ou, Y.; Kitani, K. AgentFormer: Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 9793–9803.
10. Gupta, A.; Johnson, J.; Fei-Fei, L.; Savarese, S.; Alahi, A. Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2255–2264.
11. Sadeghian, A.; Kosaraju, V.; Sadeghian, A.; Hirose, N.; Rezatofighi, H.; Savarese, S. SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical Constraints. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1349–1358.

12. Kosaraju, V.; Sadeghian, A.; Martín-Martín, R.; Reid, I.; Rezatofighi, H.; Savarese, S. Social-BiGAT: Multimodal Trajectory Forecasting Using Bicycle-GAN and Graph Attention Networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Vancouver, Canada, 2019; Volume 32.
13. He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; Jiang, W. TransReID: Transformer-Based Object Re-Identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2023.
14. Can, A.B.; Bhargava, B. SORT: A Self-ORganizing Trust Model for Peer-to-Peer Systems. *IEEE Trans. Depend. Secur. Comput.* **2013**, *10*, 14–27. [CrossRef]
15. Wojke, N.; Bewley, A.; Paulus, D. Simple Online and Realtime Tracking with a Deep Association Metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3645–3649.
16. Zhang, Y.; Wang, C.; Wang, X.; Zeng, W.; Liu, W. FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking. *Int. J. Comput. Vis.* **2021**, *129*, 3069–3087. [CrossRef]
17. Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; Wang, X. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. In *Computer Vision–ECCV 2022*; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Lecture Notes in Computer Science; Springer Nature Switzerland: Cham, Switzerland, 2022; Volume 13682, pp. 1–21. ISBN 978-3-031-20046-5.
18. Aharon, N.; Orfaig, R.; Bobrovsky, B.-Z. BoT-SORT: Robust Associations Multi-Pedestrian Tracking. *arXiv* **2022**, arXiv:2206.14651.
19. Milos, S.S.; Nemanja, I.; Srdan, S. *Decentralized Consensus-Based Estimation and Target Tracking*; Akademska misao: Beograd, Srbija, 2021; ISBN 978-86-7466-859-7.
20. You, Q.; Jiang, H. Real-Time 3D Deep Multi-Camera Tracking. *arXiv* **2020**, arXiv:2003.11753.
21. Quach, K.G.; Nguyen, P.; Le, H.; Truong, T.-D.; Duong, C.N.; Tran, M.-T.; Luu, K. DyGLIP: A Dynamic Graph Model with Link Prediction for Accurate Multi-Camera Multiple Object Tracking. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13779–13788.
22. Hou, Y.; Zheng, L.; Gould, S. Multiview Detection with Feature Perspective Transformation. In *Computer Vision–ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2020; Volume 12352, pp. 1–18. ISBN 978-3-030-58570-9.
23. Nguyen, D.M.H.; Henschel, R.; Rosenhahn, B.; Sonntag, D.; Swoboda, P. LMGP: Lifted Multicut Meets Geometry Projections for Multi-Camera Multi-Object Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
24. Li, K.; Malik, J. Implicit Maximum Likelihood Estimation. *arXiv* **2018**, arXiv:1809.09087.
25. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable Person Re-Identification: A Benchmark. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2023.
26. Wei, L.; Zhang, S.; Gao, W.; Tian, Q. Person Transfer GAN to Bridge Domain Gap for Person Re-Identification. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 79–88.
27. Xiao, Q.; Luo, H.; Zhang, C. Margin Sample Mining Loss: A Deep Learning Based Method for Person Re-Identification. **2017**.
28. Jeon, Y.; Tran, D.Q.; Park, M.; Park, S. Leveraging Future Trajectory Prediction for Multi-Camera People Tracking. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 18–19 June 2023; pp. 5399–5408.