ITC 3/55
Information Technology and Control
Vol. 55 / No. 3/ 2026
pp. 1444-1466
DOI 10.5755/j01.itc.55.3.44169

Cross-camera Tracking Technology for Intelligent Communities Based on Deep Learning and Fuzzy Clustering Algorithms

Received 2026/01/28 | Accepted after revision 2026/06/09

HOW TO CITE: Zhu, Y., Han, Z. (2026) Cross-camera Tracking Technology for Intelligent Communities Based on Deep Learning and Fuzzy Clustering Algorithms. *Information Technology and Control*, 55(3), 1443-1465. https://doi.org/10.5755/j01.itc.55.3.44169

Cross-camera Tracking Technology for Intelligent Communities Based on Deep Learning and Fuzzy Clustering Algorithms

Yufang Zhu*, Zhihua Han

School of Modern Educational Technology, Shangrao Preschool Education College, Shangrao, 334000, China; e-mails: YufangZhuzz@outlook.com, sryzjwchzh@163.com

Corresponding author: YufangZhuzz@outlook.com

In the current construction of smart communities, continuous tracking across cameras has become the key to improving security efficiency. This study aims at the existing problems such as identity consistency maintenance, noise interference and perspective changes in cross-camera target tracking in intelligent communities. Furthermore, it proposes an intelligent cross-camera tracking technology based on an improved fuzzy C-means clustering algorithm and deep learning. In the image segmentation stage, this research improves the robustness of the segmentation algorithm in complex monitoring scenarios by constructing an affinity matrix that fuses pixel and regional information, and introducing spatial regularization terms and fixed clustering center strategies. In the tracking stage, this research combines large-kernel convolutional attention and masked autoencoder self-supervised learning to achieve discriminative feature extraction and efficient cross-camera matching, and improve the accuracy of cross-camera tracking. According to experimental results, the suggested improved fuzzy C-means clustering algorithm achieved segmentation accuracy of 90.5% and 83.8% on two public datasets, and the mean intersection over union under 30% speckle noise was 94.8% and 93.7%, respectively. In the actual monitoring scenario, the proposed cross-camera tracking method achieved an identity preservation rate of 94.7%, only 38 identity switches, and an average tracking delay of only 65.6ms. The findings demonstrate that the suggested approach offers a workable technical solution for intelligent community security systems and can significantly increase the precision and usefulness of cross-camera tracking.

KEYWORDS: Deep Learning, Fuzzy C-means Clustering Algorithm, Self-supervised Learning, Intelligent Community, Cross-camera Tracking

1. Introduction

With the acceleration of urbanization and the advancement of smart community construction, video surveillance systems have become increasingly important in the fields of community security, personnel management, and behavior analysis [18]. In a multi-camera surveillance environment, achieving continuous target tracking across cameras has become the key to improving the level of intelligent community security [3]. Traditional object tracking methods rely on single-camera scenarios. It is difficult to cope with challenges such as maintaining identity consistency, perspective changes, occlusion and lighting differences when the target moves between multiple cameras [20]. Deep learning (DL) technology has advanced computer vision significantly in recent years. Specifically, novel approaches to cross-camera tracking are offered by target detection and re-identification technologies based on convolutional neural networks and attention mechanisms (AM) [14]. However, relying solely on DL models still has problems such as strong dependence on annotated data and sensitivity to scene changes [28]. In addition, the fuzzy clustering (FC) algorithm shows good robustness and adaptability in image segmentation (IS) and feature aggregation, and can effectively deal with uncertainty and noise problems in images, but its adaptability and robustness in cross-camera scenarios are poor [21]. Therefore, combining the powerful representation capabilities of DL with the advantages of FC in uncertainty modeling to build an efficient and stable cross-camera tracking framework has become the key starting point of this research.

This study proposes an intelligent community cross-camera tracking method based on DL and FC algorithms. This method first improves the segmentation accuracy (SA) and robustness of FC in complex surveillance images by introducing spatial prior information and geometric consistency constraints. Then, this study designs a self-supervised feature learning network that combines large-kernel convolutional attention and masked autoencoders to achieve robust feature extraction and matching across camera targets. Finally, this research is based on the improved siamese mask network (SiamMask) and multi-target tracker to achieve continuous tracking of multiple targets in time and space. This research aims to improve the accuracy and stability of cross-camera tracking and provide a feasible technical solution for smart community security systems.

2. Related Works

A significant area of study in computer vision and intelligent surveillance is cross-camera target tracking (CCTT). Maintaining target identity constancy in a multi-camera system is the key challenge [27]. Early research is mostly based on feature matching and motion modeling, such as optical flow method, Kalman filter, etc. However, these methods are sensitive to viewing angle changes and occlusion, and have poor robustness. Neural network-based target recognition and tracking techniques have progressively gained popularity with the advancement of DL [17]. Zhang et al. [29] proposed a video multi-camera vehicle tracking method based on appearance parsing spatiotemporal trajectory matching network. This study used a multi-appearance attribute parsing module to extract salient features, and achieves cross-camera trajectory association through an appearance joint tracking mechanism and a spatiotemporal matching module. The outcomes demonstrated that this technique might significantly increase tracking accuracy [29]. Gao et al. [4] proposed a tracking system that integrates visual and spatiotemporal information to solve the problem of poor tracking performance in non-overlapping areas across cameras. The system obtained vehicle information through online tracking, trains a Gaussian model with microscopic traffic characteristics, and combined spatiotemporal and visual features for target association. The system demonstrated good tracking performance, according to the results [4]. Luna et al. [13] proposed a global correlation method based on graph neural network to solve the problem of high computational cost and reliance on predetermined similarity in data correlation across cameras. It was found that this method significantly outperformed existing techniques and does not require scene-specific training [13]. Despite the above-mentioned research advances, challenges remain in handling long-term tracking, identity consistency, and computational efficiency.

In terms of IS, FC algorithm is widely used due to its good ability to handle noise and uncertainty [5]. However, traditional FC algorithms are sensitive to noise and lack spatial constraints, resulting in poor segmentation results in complex images [11]. To solve the above problems, researchers have proposed a variety of improvement solutions. Xu proposed a fuzzy C-means clustering (FCMC) IS algorithm based on hidden Markov models, aiming to improve the anti-interference ability of images. This algorithm used a hidden Markov model to extract texture parameters to eliminate the influence of brightness, and combined standard deviation to extract features, convert color space and perform smoothing. The results showed that the algorithm had smooth segmentation boundaries and strong anti-noise ability [26]. To address the issue of FCMC's sensitivity to noise, Song et al. presented an FC picture segmentation technique based on low-rank representation. This method used an improved superpixel generation method to enhance edges, and constructed a membership matrix constrained by fuzzy low-rank structures for subspace clustering. It was found that this method outperformed existing methods on both artificial and real images [19]. Hooda and Verma [9] proposed a FC brain IS algorithm based on gravity search, which used fuzzy inference rules to dynamically control key parameters in the gravity search algorithm to optimize the search process. Results showed that this algorithm achieved higher quality cluster segmentation results on multiple datasets [9]. Nonetheless, the application of FC in cross-camera IS remains to be further explored in the context of smart surveillance.

Recently, several approaches have been proposed to address cross-camera tracking challenges. A cross-camera tracking model based on multi-feature fusion has been developed, which integrates appearance, motion, and temporal features to improve tracking robustness across non-overlapping views [22]. For multi-target multi-camera scenarios, a tracklet-to-target assignment method has been introduced, enabling global identity association by formulating tracking as an assignment problem across camera views [23]. In the domain of object tracking, a gradient descent-based method has been proposed for tracking fixed-shape moving objects, demonstrating effective convergence properties under constrained motion patterns [16]. Furthermore, graph matching technology has been applied to single camera and cross camera multi-target tracking, where target trajectories are represented as graph structures and matched between views [15].

In summary, existing research has made some progress in cross-camera tracking, IS and AMs, but there are still the following shortcomings: (1) Existing segmentation methods have limited ability to handle complex backgrounds and noise in cross-camera scenes. (2) The DL model is highly dependent on annotated data, and the robustness of cross-camera feature matching needs to be improved. (3) The identity switching and occlusion problems in multi-target tracking have not been effectively solved. Given this, the goal of this research is to enhance the precision and resilience of target tracking in intelligent community cross-camera tracking systems using DL and FC algorithms. The innovation of this study lies in: (1) the novel affinity matrix that integrates both pixel-level and region-level information for improved segmentation robustness; (2) the original combination of large-kernel convolutional attention with masked autoencoder self-supervised learning for cross-camera feature extraction; (3) the fixed clustering center strategy within a fuzzy ensemble framework, which provides a new perspective on maintaining identity consistency across different camera views.

3. Methodology

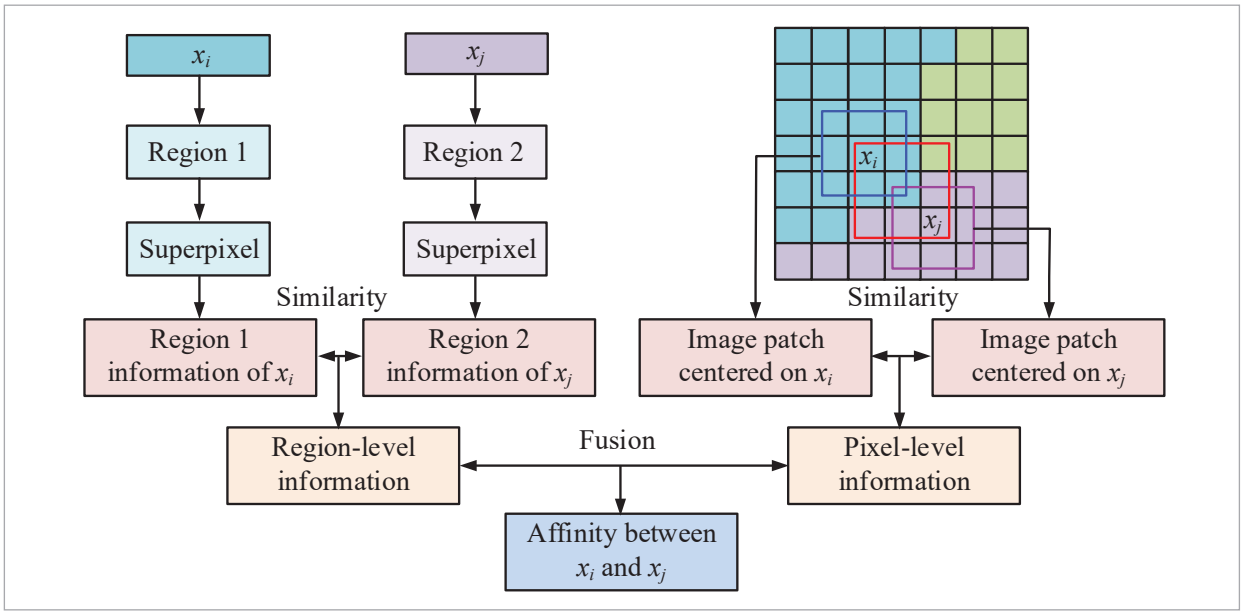
This study first proposes a cross-camera surveillance IS method based on an improved FCMC algorithm, aiming to improve the SA of cross-camera surveillance images under complex backgrounds, noise, and changes in viewing angles. Then, this study proposes a community CCTT method that combines AM and self-supervised learning to achieve efficient feature extraction, matching and continuous tracking of cross-camera targets.

3.1. Cross-camera Surveillance IS Method Based on Improved FCMC Algorithm

In cross-camera task tracking in smart communities, accurate IS is an important step to achieve target detection, feature extraction and continuous tracking. The FCMC algorithm is widely used in IS due to its advantages in handling image uncertainty [10].

Figure 1

Framework of affinity matrix construction integrating pixel-level features and superpixel region-level information.



However, in complex community environments, images will be interfered by complex backgrounds, illumination changes, occlusions, and noise. The traditional FCMC algorithm lacks spatial information constraints, is sensitive to noise, and has insufficient segmentation robustness under changes in cross-camera viewing angles [25]. Therefore, to improve the segmentation quality of cross-camera surveillance images, this study proposes an improved FCMC algorithm, which improves the SA and robustness of cross-camera surveillance images by introducing spatial prior information and geometric consistency constraints. This study first constructs a new type of affinity matrix that integrates pixel-level and regional-level information to better express the spatial and feature relationships between pixels in the image. Meanwhile, this study introduces a prescribed cluster center strategy in the clustering process to reduce the impact of noise on membership updates. The geometric consistency constraint model is then used in this work to guarantee the structural consistency between the original image and the membership matrix.

To effectively fuse the spatial structure information and visual features of the image, this study constructed an affinity matrix that fuses spatial priors. It is assumed that the input image is $I \in \boldsymbol{R}^{H \times W \times C}$. Among them, H, W, and C display the height, width, and quantity of channels, as shown in Equation (1).

$$A_{ij} = \exp\left(-\frac{\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2}{2\sigma_f^2}\right) \cdot \exp\left(-\frac{\left\|p_i - p_j\right\|^2}{2\sigma_s^2}\right) \cdot \tag{1}$$

In Equation (1), A_{ij} represents the affinity between pixels i and j. $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ represent the feature vectors (FVs) of pixels i and j respectively. p_i and p_j represent the spatial coordinates of pixels and pixels. σ_f and σ_s represent feature scale parameters and spatial scale parameters respectively. This affinity matrix takes into account both feature similarity and spatial proximity, and its calculation framework is shown in Figure 1.

In Figure 1, the computing framework of the affinity matrix first divides regions through superpixel segmentation and extracts region-level information. Then, image patches are used to extract pixel-level features. After the two are fused, the affinity between pixels is calculated. Traditional FCMC is based on the feature distance minimization objective function (OF), which ignores the spatial relationship between pixels and is susceptible to noise interference [2]. Its OF is shown in Equation (2).

$$J_{\mathrm{FCM}}=\sum_{i=1}^{N}\sum_{k=1}^{K}u_{ik}^{m}\left\|\boldsymbol{x}_i-v_k\right\|. \tag{2}$$

In Equation (2), J_{FCM} displays the OF value of the FCMC algorithm. N represents the total quantity of pixels. K represents the quantity of clusters. u_{ik} represents the membership degree (MD) of pixel i belonging to the k-th cluster. m represents the blur index. v_k represents the k-th clustering center. The core difference between the proposed pixel region fusion and traditional local region combination lies in the transition from static weighting and to spatial regularization based on dynamic affinity. The OF is shown in Equation (3).

$$J'=\sum_{i=1}^{N}\sum_{j=1}^{C}u_{ij}^{m}\left\|x_i-v_j\right\|^2+\sum_{i=1}^{N}\sum_{j=1}^{C}u_{ij}^{m}\left(\frac{1}{|N_i|}\sum_{r\in N_i}\left\|x_r-v_j\right\|^2\right). \tag{3}$$

In Equation (3), J' represents the OF and the total energy of algorithm optimization. m represents the fuzzy weighted index. x_r represents the r-th pixel in domain N_i. This method treats all pixels within the window equally, which can easily lead to edge blurring. In contrast, this article constructs spatial regularization terms through affinity matrices, whose expression is shown in Equation (4).

$$\begin{cases}J_{\mathrm{reg}}=\lambda\sum_{i=1}^{N}\sum_{k=1}^{K}Aij\sum_{k=1}^{K}(u_{ik}-u_{jk})^2\\ J=J_{\mathrm{FCM}}+J_{\mathrm{reg}}\end{cases}. \tag{4}$$

In Equation (4), J_{reg} represents the space regularization term. λ represents the regularization coefficient. $(u_{ik}-u_{jk})$ represents the difference in membership of two pixels on the k-th cluster. J represents the improved overall OF. During the iterative optimization process, dynamically changing cluster centers will drift due to noise pixels [6]. Traditional FCMC center updates are highly sensitive to outliers, and noisy pixels can cause the centroid to deviate from the true distribution. Therefore, by adopting a fixed clustering center strategy and introducing a strong spatial prior to limit the search space, the study effectively shields the member update process from random noise interference and ensures convergence to a stable, physically meaningful class representation. Its expression is shown in Equation (5).

$$u_{ik}=\frac{\left(\left\|\boldsymbol{x}_i-v_k\right\|^2+\lambda\sum_{j=1}^{N}A_{ij}(u_{ik}-u_{jk})\right)^{-\frac{1}{m-1}}}{\sum_{l=1}^{K}\left(\left\|\boldsymbol{x}_i-v_l\right\|^2+\lambda\sum_{j=1}^{N}A_{ij}(u_{il}-u_{jl})\right)^{-\frac{1}{m-1}}}. \tag{5}$$

In Equation (5), l represents the summation of clusters. $\frac{1}{m-1}$ represents the power exponent, ensuring normalization of MDs. In order to obtain more stable and geometrically consistent segmentation results, this study integrates multiple sub-models based on different feature representations. The integration goal is to find a globally consistent membership matrix, as shown in Equation (6).

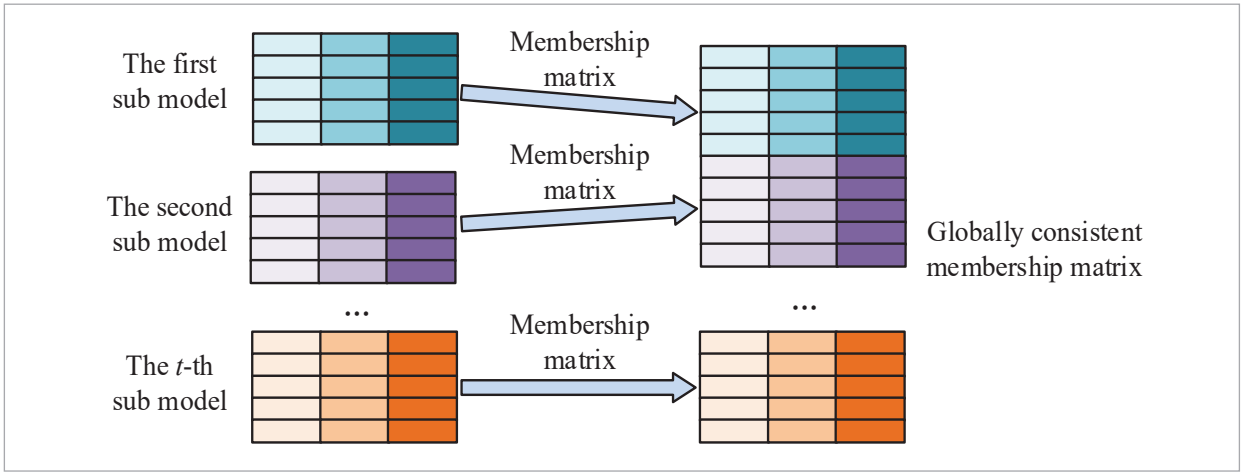
$$\begin{cases}\boldsymbol{U}^{*}=\arg\min_{\boldsymbol{U}}\sum_{t=1}^{T}\left\|\boldsymbol{U}-\boldsymbol{U}^{(t)}\right\|_F^2+\beta\cdot\mathrm{tr}(\boldsymbol{U}^T\boldsymbol{L}\boldsymbol{U})\\ \boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}\end{cases}. \tag{6}$$

In Equation (6), $\boldsymbol{U}^{*}$ represents a globally consistent membership matrix. T represents the number of sub-models. $\boldsymbol{U}^{(t)}$ represents the membership matrix generated by the t-th sub-model. β represents the balance parameter. $\boldsymbol{L}$ stands for Laplacian matrix. $\boldsymbol{D}$ represents the degree matrix. By combining the outputs of several sub-models, this matrix is intended to increase the resilience and consistency of segmentation results. Figure 2 displays its structural diagram.

In Figure 2, the membership matrix of FC integration contains multiple sub-models, and each model independently generates a membership matrix. These matrices are fused through an integration process to generate a globally consistent membership matrix. Equation (7) [7] illustrates how the effective semi-implicit gradient descent approach is applied to the optimization problem of the overall OF.

$$\boldsymbol{U}^{(k+1)}=\boldsymbol{U}^{(k)}-\eta\left(2\sum_{t=1}^{T}(\boldsymbol{U}^{(k)}-\boldsymbol{U}^{(t)})+2\beta\boldsymbol{L}\boldsymbol{U}^{(k)}\right). \tag{7}$$

Figure 2
Schematic diagram of membership matrix structure for FC ensemble.



In Equation (7), $\boldsymbol{U}^{(k+1)}$ represents the membership matrix updated after the k+1-th iteration. η represents the learning rate. $2\beta\boldsymbol{L}\boldsymbol{U}^{(k)}$ represents the smoothing term gradient. During the iterative process, its convergence criterion is displayed in Equation (8).

$$\left\|\boldsymbol{U}^{(k+1)}-\boldsymbol{U}^{(k)}\right\|_F<\varepsilon. \tag{8}$$

In Equation (8), $\left\|\boldsymbol{U}^{(k+1)}-\boldsymbol{U}^{(k)}\right\|_F$ represents the Frobenius norm between two adjacent iteration solutions. ε represents the preset threshold. After obtaining the optimal integrated fuzzy membership matrix, it needs to be converted into deterministic category labels to generate the final segmentation image [1]. This study achieves defuzzification by taking the maximum index of the membership vector of each pixel, as shown in Equation (9).

$$y_i=\arg\max_k u_{ik}^*. \tag{9}$$

In Equation (9), y_i represents the final category label of pixel i. u_{ik}^* displays the final MD of pixel i belonging to cluster k. In summary, the proposed cross-camera surveillance IS process based on the improved FCMC algorithm is shown in Figure 3.

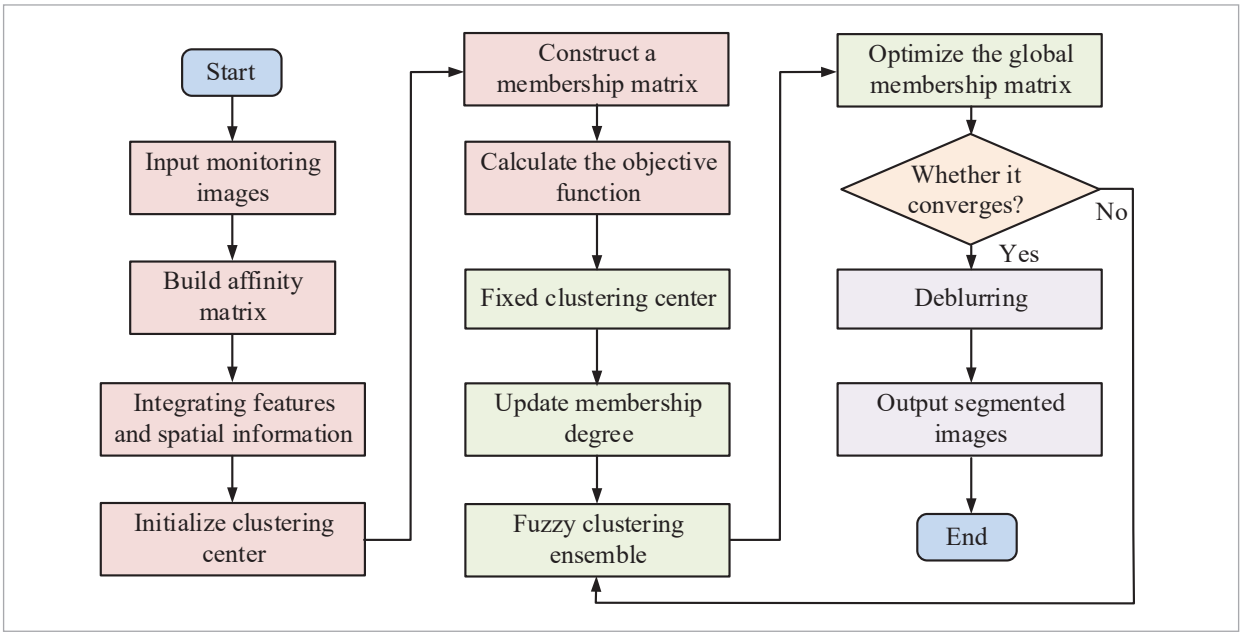
In Figure 3, the cross-camera surveillance IS process based on improved FCMC starts from inputting surveillance images. This study sequentially constructs an affinity matrix, fuses features and spatial information, and initializes the cluster center. Subsequently, the study carries out iterative optimization by constructing a membership matrix, calculating the OF, fixed clustering center and updating the membership, and then improves the robustness through FC integration and global membership matrix optimization.

Inaccurate segmentation produces fragmented or contaminated instance masks, which directly corrupt the feature extraction process. Specifically, when a target is incorrectly merged with background or partially split across multiple regions, the resulting feature vector becomes less discriminative. During cross-camera association, such degraded features increase the Euclidean distance between same-identity samples while reducing the distance to different-identity samples, leading to identity switches. Therefore, improved segmentation serves as a prerequisite for robust re-identification.

3.2. Community CCTT Method Combining AM and Self-supervised Learning

After completing the segmentation of multi-camera surveillance images, it is necessary to extract and match the deleted target features to adapt to cross-camera perspective changes, occlusions and identity consistency issues. The AM can adaptively focus on the discriminative area, effectively suppress background noise and occlusion interference, and improve the robustness of feature expression [8].

Figure 3
Cross camera monitoring IS process based on improved FCMC algorithm.



Self-supervised learning can mine potential semantic structures from massive unlabeled data, enhance the model's adaptability to changes in viewing angle and illumination, and reduce reliance on manual annotation [12]. To this end, this study proposes an intelligent community CCTT method that combines AM and self-supervised learning. This method first uses an AM based on large kernel convolution filling to perform robust feature extraction on the segmented targets. It then uses masked autoencoders for self-supervised feature learning to achieve efficient matching and re-identification of cross-camera objects. Finally, this research uses an improved SiamMask network and a multi-target tracker to achieve continuous tracking of multiple targets in space and time.

To achieve robust extraction and enhancement of cross-camera target features, this research designs a new AM. After obtaining a feature map (FM) with a broad receptive field through large-core convolution, this mechanism makes the FM evenly divided in the spatial dimension through specific filling operations, and then reorganizes and fuses the divided small feature blocks in the channel dimension to enhance the expressive ability and discriminability of the features. The FM output after large kernel convolution and padding is shown in Equation (10).

$$\begin{cases} F_{\mathrm{LKC}} = \mathrm{Conv}_{k\times k}(F) \in \boldsymbol{R}^{H\times W\times C} \\ F_{\mathrm{pad}} = \mathrm{Padding}(F_{\mathrm{LKC}}) \in \boldsymbol{R}^{N\times N\times C} \end{cases}. \tag{10}$$

In Equation (10), F_{LKC} represents the output FM after large kernel convolution. F_{pad} represents the filled FM. $\mathrm{Conv}_{k\times k}$ stands for large kernel convolution operation. F represents the input FM. Padding stands for symmetrical filling method. Then, this study spliced all small feature blocks in the channel dimension to form a reorganized feature tensor, as shown in Equation (11).

$$F_{\mathrm{blocks}} = \mathrm{Concat}_{\mathrm{channel}}(\mathrm{Split}_{K\times K}(F_{\mathrm{pad}})) \in \boldsymbol{R}^{\frac{N}{K}\times\frac{N}{K}\times C}. \tag{11}$$

In Equation (11), F_{blocks} represents the reorganized feature tensor. $\mathrm{Split}_{K\times K}(F_{\mathrm{pad}})$ means that the filled FM is evenly divided into $K\times K$ small non-overlapping feature blocks in the spatial dimension. This paper presents a lightweight fusion module to build spatial attention weights by fusing the original feature information with the rearranged block infor-

mation. The characteristics of its output are shown in Equation (12).

$$\begin{cases} A_s = \sigma(F_{\text{resized}} + F_{\text{pad}}) \in \boldsymbol{R}^{N\times N\times C} \\ F_{\text{enhanced}} = A_s \otimes F_{\text{pad}} \in \boldsymbol{R}^{N\times N\times C} \end{cases}. \quad (12)$$

In Equation (12), A_s represents spatial attention. F_{resized} represents the upsampled FM. σ stands for Sigmoid activation function. F_{enhanced} represents the enhanced feature output. By extending the receptive field and structuring feature restructuring, this AM improves the model's capacity to concentrate on target critical areas in cross-camera situations. Its schematic diagram is shown in Figure 4.

In Figure 4, the AM of large kernel convolution filling takes the input features as the starting point, and first expands the spatial size of the FM through the filling operation. Then, it adopts large kernel convolution for feature extraction to expand the receptive field and capture richer contextual information. Finally, the enhanced FM is output. To reduce dependence on annotated data, this study uses masked autoencoders for self-supervised feature learning. This method randomly masks part of the input image and trains the network to reconstruct the masked area, thereby learning a visual representation with strong generalization ability. The complete image reconstructed by its characteristics is shown in Equation (13).

$$\begin{cases} I_m = I \otimes \boldsymbol{M} \\ \boldsymbol{z} = En(I_m) \\ I_r = De(\boldsymbol{z}) \end{cases}. \quad (13)$$

In Equation (13), I_m represents the masked image. I represents the original input image. $\boldsymbol{M}$ represents a binary mask matrix. represents the encoder network. $\boldsymbol{z}$ represents the latent FV. I_r stands for reconstructed image. De represents the decoder network. In order to enhance the discriminability of features in cross-camera scenarios, this study introduces triplet loss, as shown in Equation (14) [24].

$$L_{\text{triplet}} = \max(0, d(z_a, z_p) - d(z_a - z_n) + \alpha). \quad (14)$$

In Equation (14), L_{triplet} represents the triplet loss. z_a represents the FV of the anchor point. z_p and z_n represent the FVs of positive samples and negative samples respectively. α represents the interval parameter. The proposed masked self-supervised learning network learns robust feature representations containing rich semantics by randomly masking and reconstructing input images. Its schematic diagram is shown in Figure 5.

In Figure 5, the mask self-supervised learning network starts from an input image of 128×256 pixels. After data loading and random mask processing, it enters the encoder module. The encoder consists

Figure 4

Structure of the large kernel convolution based AM with padding and feature block reorganization.

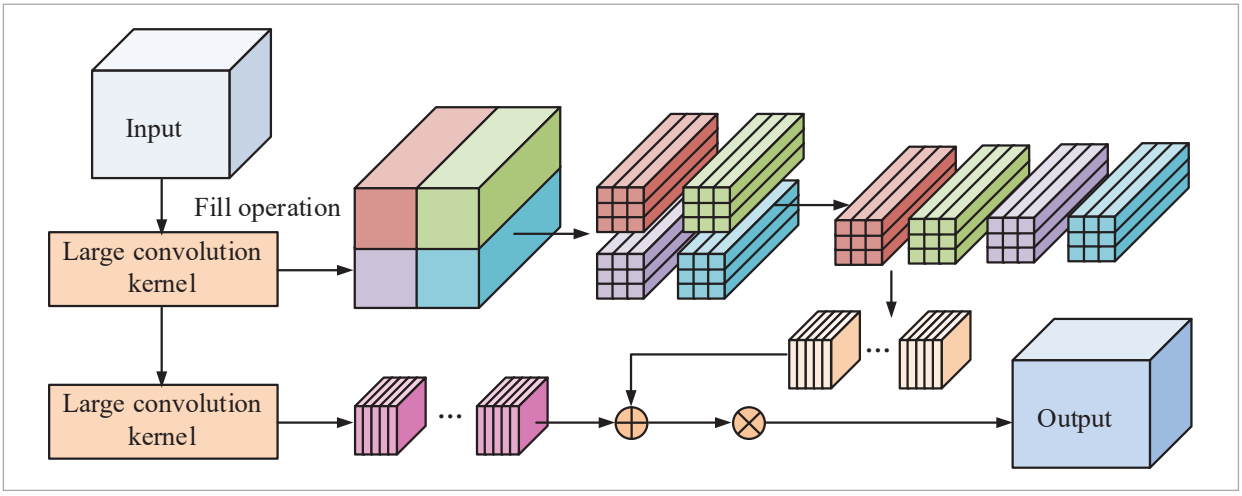


Figure 5
Architecture of the masked autoencoder for self supervised feature learning with random masking and reconstruction.

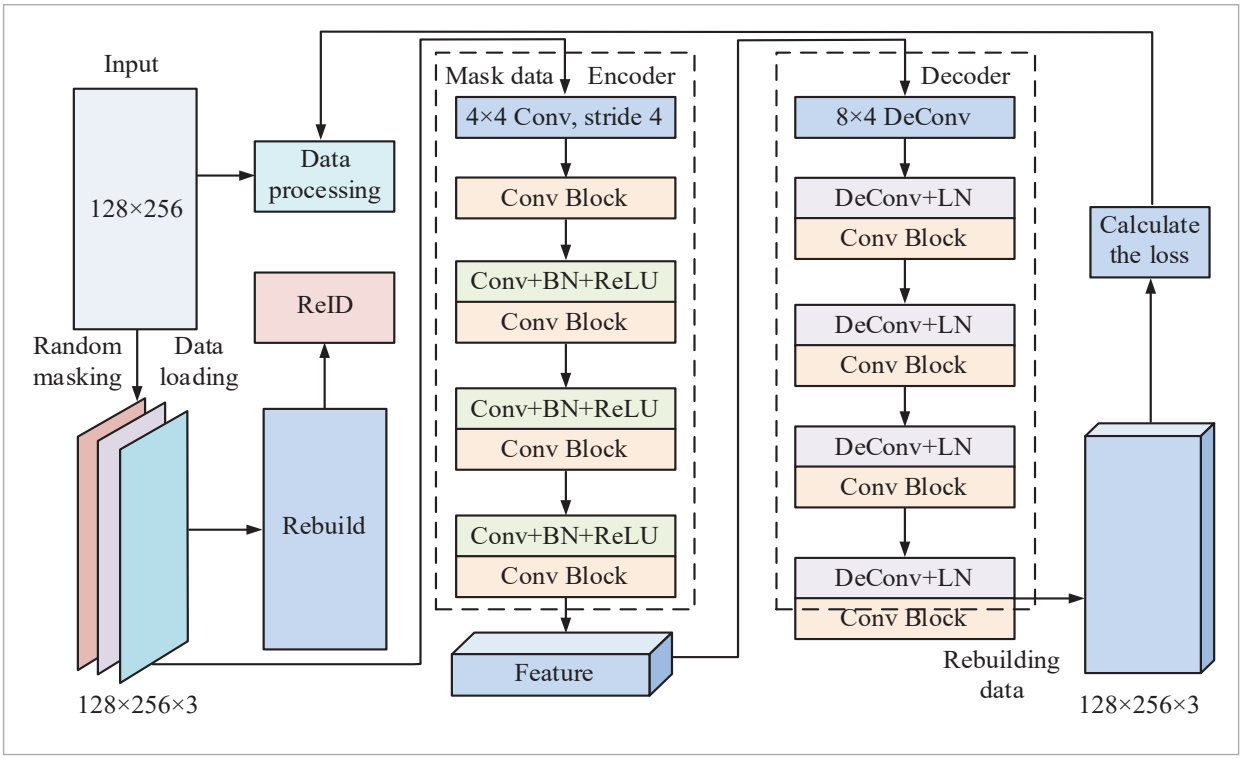
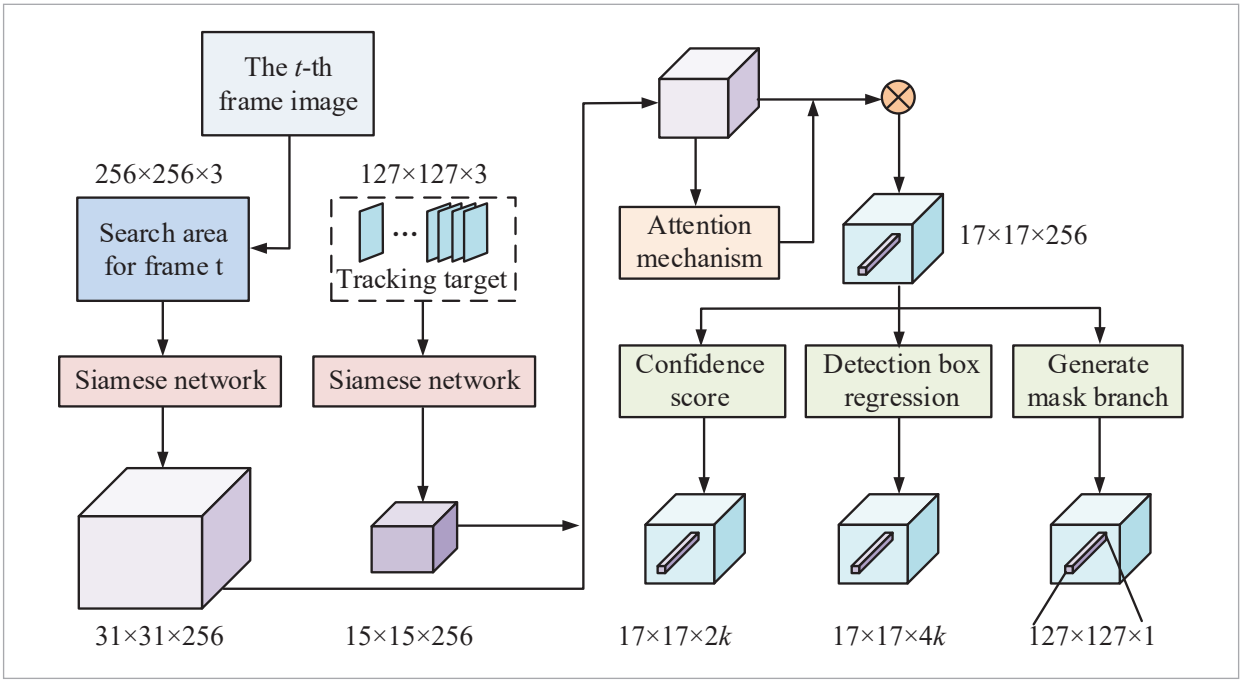


Figure 6
Improved SiamMask network with historical frame template fusion and attention enhanced feature matching.



of multiple convolutional blocks to extract deep features step by step. Subsequently, the decoder gradually reconstructs the masked area through the deconvolution layer and outputs the reconstructed image with the same size as the original image. This study further addresses the problem of insufficient template updates in long-term tracking by the traditional SiamMask network and introduces a historical frame template fusion mechanism, aiming to perform effective long-term tracking of multiple cameras, as shown in Equation (15).

$$T_{\text{fused}} = \text{Conv}_{1\times 1}(\text{Concat}(T_1, T_2, \ldots, T_g)). \quad (15)$$

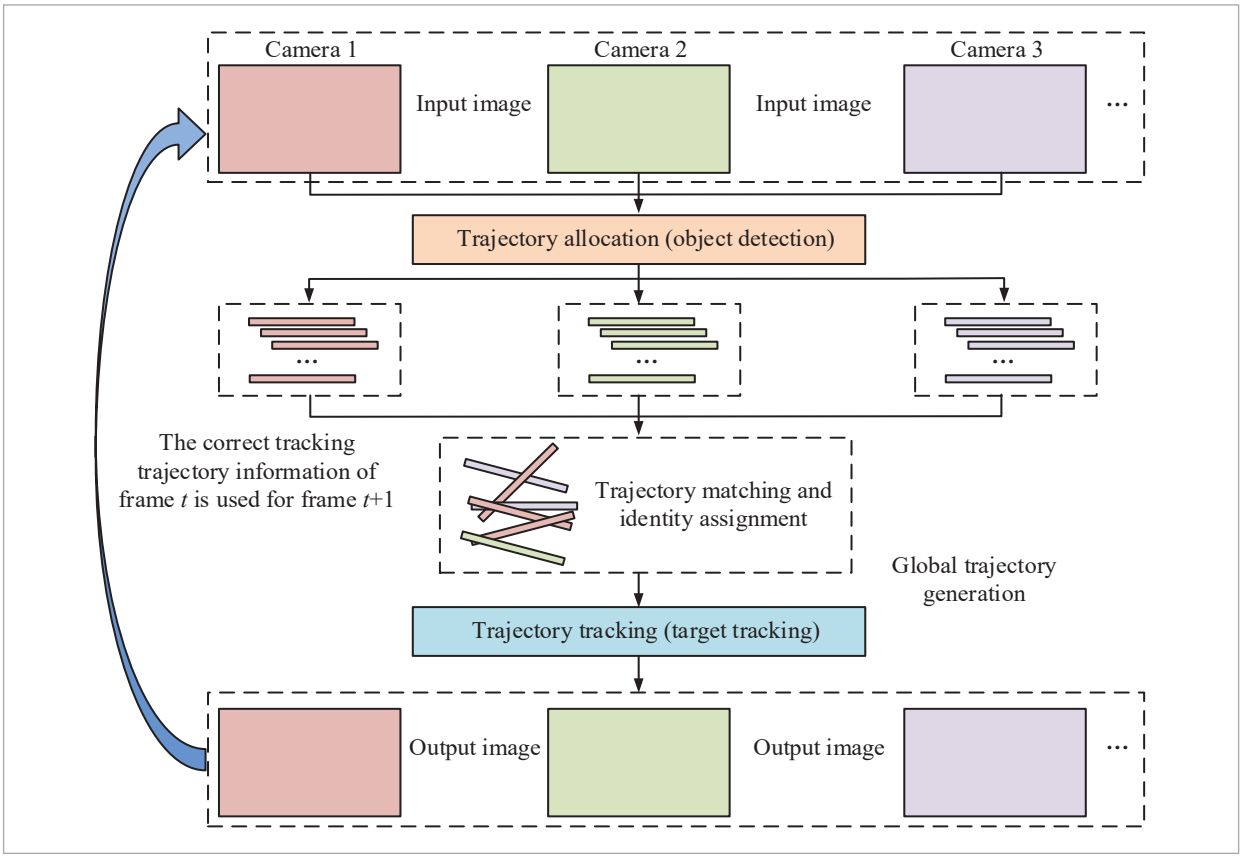
In Equation (15), T_{fused} represents the unified template feature after fusion, which is used for target matching in the subsequent tracking process. (T_1, T_2,...,T_g) represents the target template features extracted from historical frames. g is the number of historical frames. The improved SiamMask network is designed to improve the network's template adaptability and matching accuracy in long-term tracking. Figure 6 displays its structural diagram.

In Figure 6, the improved SiamMask network takes the t-th frame image as input, processes the search area and target template respectively through the siamese network, extracts features and performs similarity matching. This study introduces an AM to

Pseudocode 1: Overall pipeline of the proposed cross-camera tracking method

Input: Multi-camera surveillance image sequence
Output: Cross-camera target trajectories T_global
for each camera view **do**
  for each frame I_i **do**
    // **Step 1:** Image segmentation using improved FCMC (Section 3.1)
    Construct affinity matrix A_ij using Eq. (1)
    Initialize cluster centers v_k
    repeat
      Update membership matrix u_ik using Eq. (5) (fixed center strategy)
      Apply spatial regularization using Eq. (4)
      Enforce geometric consistency via Eq. (6)
    until convergence (Eq. (8))
    Obtain final segmentation mask y_i via Eq. (9)
    // **Step 2:** Feature extraction with attention mechanism (Section 3.2)
    Extract feature map F via large-kernel convolution (Eq. (10))
    Apply padding and feature block reorganization (Eq. (11))
    Generate spatial attention weights A_s (Eq. (12))
    Enhance features using masked autoencoder (Eq. (13))
    // **Step 3:** Target tracking and cross-camera association
    Perform multi-target tracking using improved SiamMask (Eq. (15))
    Extract feature vectors for detected targets
    Compute correlation cost matrix C_pq using Eq. (16)
    Apply Hungarian algorithm for data association
    Update trajectories with spatiotemporal constraints
end for
// **Step 4:** Global trajectory integration
Merge local trajectories from all cameras into T_global
return T_global

Figure 7
Overall framework of the proposed cross camera multi target tracking system integrating trajectory matching and identity assignment.



enhance feature expression and generate confidence scores, detection box regression and mask branch outputs. In the cross-camera scenario, this study adopts a multi-object tracker based on feature similarity for data association. For the detected target in each camera, its FV is extracted, and target association across cameras is performed based on the feature distance through the Hungarian algorithm. The correlation cost matrix is shown in Equation (16).

$$\boldsymbol{C}_{pq} = 1 - \frac{f_p \cdot f_q}{\|f_p\| \|f_q\|} \cdot \tag{16}$$

In Equation (16), $\boldsymbol{C}$ represents the correlation cost matrix. f_p and f_q represent the FVs of the p-th and q-th targets respectively. Then, this study combines spatiotemporal constraints to filter out association assumptions that are inconsistent with motion consistency, and finally forms a continuous trajectory across cameras to achieve global tracking of multi-targets across cameras in the community. The schematic diagram of its overall tracking framework is shown in Figure 7.

In Figure 7, the intelligent community CCTT obtains input images from multiple cameras, and extracts target information through the trajectory distribution module. Then, cross-camera target association and updating are performed through trajectory matching and identity assignment, combined with historical trajectory information. Finally, the trajectories of each camera are integrated to generate a globally consistent continuous motion trajectory, and the tracking result image is output.

The pseudocode of the proposed cross camera tracking method is as follows.

4. Results and Analysis

The performance of the suggested cross-camera surveillance IS approach is first confirmed by this study. Then, the study analyzes the performance of the proposed community CCTT method to verify its tracking effect in aspects such as cross-viewing angles, occlusion, and identity preservation, and evaluate its practicality and real-time performance in actual smart community scenarios.

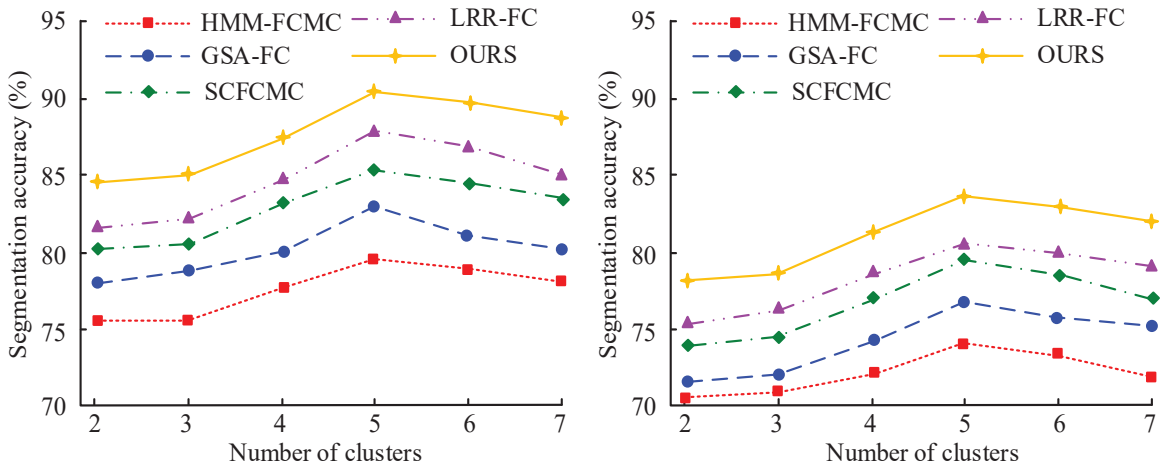
4.1 Performance Verification of Cross-camera Surveillance IS Method

This study chooses two public datasets – the Market-1501 dataset and the WildTrack dataset – for verification in order to confirm the effectiveness of the suggested enhanced FCMC algorithm in cross-camera surveillance IS. The Market-1501 dataset contains 32668 annotated bounding box images of 1501 pedestrians captured by 6 surveillance cameras. Its scenes are diverse, with dramatic changes in lighting and perspective. Market-1501 is originally a person re-identification dataset with cropped pedestrian boxes rather than a standard multi-object tracking benchmark. For tracking evaluation, the original video sequences were reconstructed by mapping cropped boxes back to frame-level coordinates, and temporal continuity was established using the provided frame IDs and tracklet labels. Pedestrian trajectories were then formed by linking detections across consecutive frames within each camera view. Based on these reconstructed trajectories, tracking accuracy, FPS, and IDS were reported to evaluate cross-camera tracking performance. The WildTrack dataset is collected in an open public space using 7 synchronized calibrated cameras, covering most of the square. There are a large number of overlapping and partially non-overlapping areas between the fields of view, forming a true wide-area surveillance network. Neither the Market-1501 nor the WildTrack dataset originally provides pixel-level segmentation masks. Therefore, pixel-level ground truth was generated by manually annotating a representative subset of images from each dataset. Specifically, 500 images were randomly selected from Market-1501 and 300 images from WildTrack. For each selected image, foreground (pedestrian) regions were labeled at the pixel level using professional annotation tools (LabelMe). The remaining unannotated images were not used for segmentation evaluation; they were only employed for tracking experiments where pixel-level masks are not required. This study compared the proposed improved FCMC algorithm with existing advanced magic drama clustering algorithms, including spatially constrained FCMC (SCFCMC) algorithm, hidden markov model-based FCMC (HMM-FCMC) algorithm, low-rank representation-based FC (LRR-FC) algorithm and gravitational search algorithm-based FC (GSA-FC) algorithm. Table 1 displays the experimental setup and parameter values. The semantic meaning of the five clusters is interpreted as follows: Cluster 1 (background, e.g., roads and walls), Cluster 2 (pedestrian body/limbs), Cluster 3 (clothing and accessories),

Table 1
Experimental environments and parameter settings.

Experimental environments		Parameter settings	
Items	Configuration	Items	Parameters
Operating system	Ubuntu 20.04 LTS	Fuzziness exponent	2.0
Processor	Intel Core i9-12900K	Regularization coefficient	0.5
Memory	64 GB DDR4	Balance parameter	0.3
Graphics card	NVIDIA GeForce RTX 4090 (24 GB)	Learning rate	0.01
DL framework	PyTorch 2.0.1	Convergence threshold	1e-5
Programming language	Python 3.9	Number of clusters	5

Figure 8
Comparison of SA of different algorithms on the Market-1501 dataset under different numbers of clusters.



(a) Segmentation accuracy of different algorithms on Market-1501 dataset

(b) Segmentation accuracy of different algorithms on WildTrack dataset

Cluster 4 (shadow and low-contrast regions), and Cluster 5 (occluding objects and boundaries). This five-cluster configuration is meaningful because it captures the intrinsic variability of surveillance scenes while avoiding over-segmentation. The downstream tracker's sensitivity was analyzed by varying the cluster number from 2 to 7.

To quantitatively evaluate the performance of the proposed cross camera tracking method, indicators such as tracking accuracy, identity retention rate, and trajectory coverage were used in the study. Among them, the tracking accuracy is the ratio of the number of correct tracking frames for a single target to the actual number of frames, as shown in Equation (17).

$$\mathrm{TA} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i + \mathrm{FP}_i}. \tag{17}$$

In Equation (17), TA represents tracking accuracy. N is the total number of targets. TP_i is the frame rate at which target i is correctly tracked. FN_i and FP_i are the number of missed detection frames and false detection frames of target i, respectively. The identity retention rate measures the ability to maintain identity tag consistency across multiple cameras, as shown in equation (18).

$$\mathrm{IPR} = \frac{1}{M}\sum_{j=1}^{M}\frac{\mathrm{ID}^{j}_{\mathrm{correct}}}{\mathrm{ID}^{j}_{\mathrm{total}}}. \tag{18}$$

In Equation (18), IPR represents the identity retention rate. M is the total number of cross camera conversions. $\mathrm{ID}^{j}_{\mathrm{correct}}$ is the number of times the identity is correctly preserved in the j-th conversion. $\mathrm{ID}^{j}_{\mathrm{total}}$ is the total number of identity assignments attempted in this conversion. The quantification of trajectory coverage estimates the completeness of the trajectory relative to the true trajectory, as shown in Equation (19).

$$\mathrm{TC} = \frac{1}{N}\sum_{i=1}^{N}\frac{L^{i}_{\mathrm{match}}}{L^{i}_{gt}}. \tag{19}$$

In Equation (19), TC represents the trajectory coverage rate. L^{i}_{match} is the length of the successfully matched part in the trajectory of target i. L^{i}_{gt} is the total length of the true trajectory of target i.

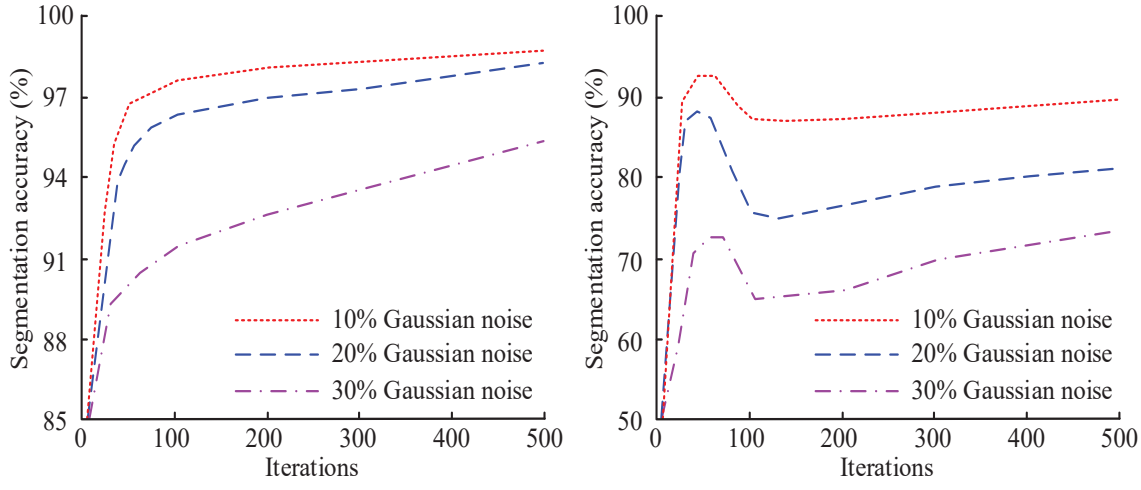
The study first analyzed the SA of different FC algorithms under different cluster numbers on two datasets. In Figure 8(a), on the Market-1501 dataset, when the quantity of clusters was 3, the SA rates of

HMM-FCMC, GSA-FC and SCFCMC were 75.6%, 78.3%, and 80.5%, respectively. The SA of LRR-FC and the proposed algorithm were 82.1% and 85.2%, respectively. When the quantity of clusters reached 5, the SA of the five algorithms were 79.8%, 82.7%, 85.4%, 87.2%, and 90.5%, respectively. In Figure 8(b), on the WildTrack dataset, when the quantity of clusters was 5, the SA of HMM-FCMC, GSA-FC, SCFCMC, and LRR-FC were 73.4%, 76.1%, 78.8%, and 80.6%, and the accuracy of the proposed algorithm was 83.8%. The findings demonstrated that the suggested approach outperformed the comparison algorithm and had the best SA on the two datasets when there were five clusters. The rationale for selecting five clusters is justified as follows. Pedestrian segmentation in surveillance scenarios typically requires distinguishing the human body from background, clothing regions, and potential occluding objects. Fewer clusters tend to merge semantically distinct regions, such as dark clothing with shadow backgrounds, leading to under-segmentation. Conversely, more cluster cause over-segmentation where a single pedestrian is fragmented into multiple small regions, complicating subsequent feature extraction. As shown in Figure 8, segmentation accuracy improved monotonically from 2 to 5 clusters on both datasets, reaching 90.5% (Market-1501) and 83.8% (WildTrack) at 5 clusters. Therefore, five clusters were empirically determined as the optimal trade-off between segmentation fidelity and computational cost.

To verify the rationality of fixed clustering, the study was conducted on the Market-1501 dataset, and the IS accuracy of the proposed algorithm in fixed clustering and non-fixed clustering was analyzed. The Gaussian noise intensities were 10%, 20%, and 30%, respectively. In Figure 9(a), when the iteration was 100, the proposed algorithm achieved 97.4%, 95.8%, and 91.2% IS accuracy in fixed clustering with Gaussian noise intensity of 10%, 20%, and 30%. When the iteration reached 500, the IS accuracy under three Gaussian noise intensities was 98.3%, 97.8%, and 94.6%. In Figure 9(b), when the iteration was 100, under three Gaussian noise intensities, the IS accuracy of the proposed algorithm in non-fixed clustering was 88.2%, 74.9%, and 67.5%. When the iteration increased to 500, the accuracy rates were 89.2%, 78.6%, and 72.3%. The effectiveness of the fixed clustering technique was confirmed by the findings, which demonstrated that the suggested algorithm had higher IS accuracy in fixed clustering.

Figure 9

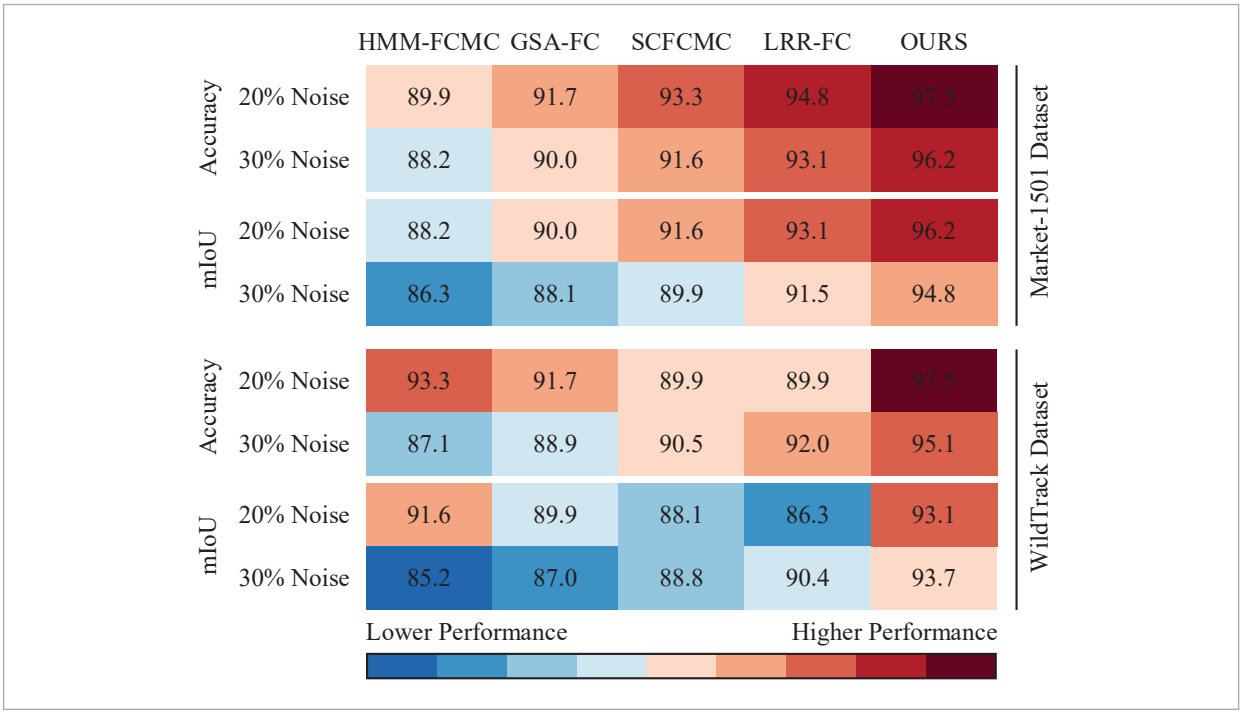
Comparison of SA of the proposed algorithm under fixed clustering and non-fixed clustering strategies on the Market-1501 dataset with different Gaussian noise intensities.



(a) Segmentation accuracy of the proposed algorithm in fixed clustering

(b) Segmentation accuracy of the proposed algorithm in non fixed clustering

Figure 10
Comparison of SA and mean intersection over union of different FC algorithms on the Market-1501 and WildTrack datasets under 20% and 30% speckle noise.



The study conducts a comparative analysis on the IS accuracy and mean intersection over union (mIoU) of different algorithms under different speckle noise on two datasets. Speckle noise was 20% and 30%, respectively. In Figure 10, on the Market-1501 dataset, the IS accuracy of HMM-FCMC, GSA-FC, SCFCMC, and LRR-FC when the speckle noise was 20% was 89.9%, 91.7%, 93.3%, and 94.8%, respectively. When the noise was 30%, the accuracy rates were 88.2%, 90.0%, 91.6%, and 93.1%, respectively. The SA of the proposed algorithm under two types of noise was 97.5% and 96.2%. On the WildTrack dataset, when the speckle noise was 30%, the IS accuracy rates of HMM-FCMC, GSA-FC, SCFCMC, LRR-FC, and the proposed algorithm were 87.1%, 88.9%, 90.5%, 92.0%, and 95.1%, respectively. In When the speckle noise was 30%, the mIoU of the five algorithms on the Market-1501 dataset were 86.3%, 88.1%, 89.9%, 91.5%, and 94.8%, respectively. On the WildTrack dataset, the mIoU of the five algorithms when the speckle noise was 30% were 85.2%, 87.0%, 88.8%, 90.4%, and 93.7%, respectively. The outcomes displayed that the IS accuracy and mIoU of the proposed algorithm were better than those of the comparison algorithm under high speckle noise, which verified the superiority of the algorithm.

The paper also compared the boundary displacement error (BDE), global consistency error (GCE), and segmentation covering (CV) of several algorithms on two datasets in order to assess the effectiveness of the suggested approach on IS boundary accuracy and region consistency. Table 2 displays the outcomes. On the Market-1501 dataset, the BDE, GCE, and CV of the proposed algorithm were 7.23 pixels, 2.15%, and 98.72% respectively, which was better than comparison algorithms such as HMM-FCMC, GSA-FC, SCFCMC, and LRR-FC. On the WildTrack dataset, the proposed algorithm also achieved the best results, with BDE, GCE, and CV being 7.67 pixels, 2.84%, and 97.95%, respectively. This once again verified its superiority in segmentation boundary accuracy and region consistency in complex scenes across cameras.

To evaluate the vulnerability of the proposed affinity matrix to superpixel segmentation quality, sensitiv-

Table 2
Comparison of boundary and region consistency metrics for different algorithms in IS.

Datasets	Algorithms	BDE (pixels)	GCE (%)	CV (%)
Market-1501	HMM-FCMC	10.23 ± 0.45 [9.8-10.6]	6.37 ± 0.31 [6.1-6.6]	94.26 ± 0.52 [93.8-94.7]
	GSA-FC	9.87 ± 0.38 [9.5-10.2]	5.02 ± 0.28 [4.8-5.3]	95.83 ± 0.44 [95.4-96.2]
	SCFCMC	8.92 ± 0.33 [8.7-9.2]	4.18 ± 0.24 [4.0-4.4]	96.58 ± 0.38 [96.2-97.0]
	LRR-FC	8.45 ± 0.29 [8.2-8.7]	3.42 ± 0.19 [3.2-3.6]	97.21 ± 0.31 [96.9-97.5]
	OURS	7.23 ± 0.21 [7.0-7.4]	2.15 ± 0.12 [2.0-2.3]	98.72 ± 0.15 [98.6-98.9]
WildTrack	HMM-FCMC	11.08 ± 0.52 [10.6-11.5]	7.45 ± 0.42 [7.1-7.8]	92.78 ± 0.65 [92.2-93.3]
	GSA-FC	10.45 ± 0.48 [10.1-10.9]	6.21 ± 0.37 [6.0-6.5]	94.12 ± 0.58 [93.6-94.6]
	SCFCMC	9.34 ± 0.41 [9.1-9.7]	5.13 ± 0.33 [4.9-5.4]	95.70 ± 0.49 [95.2-96.2]
	LRR-FC	8.89 ± 0.36 [8.6-9.1]	4.06 ± 0.25 [3.8-4.3]	96.43 ± 0.42 [96.0-96.8]
	OURS	7.67 ± 0.25 [7.4-7.9]	2.84 ± 0.18 [2.7-3.0]	97.95 ± 0.22 [97.7-98.2]

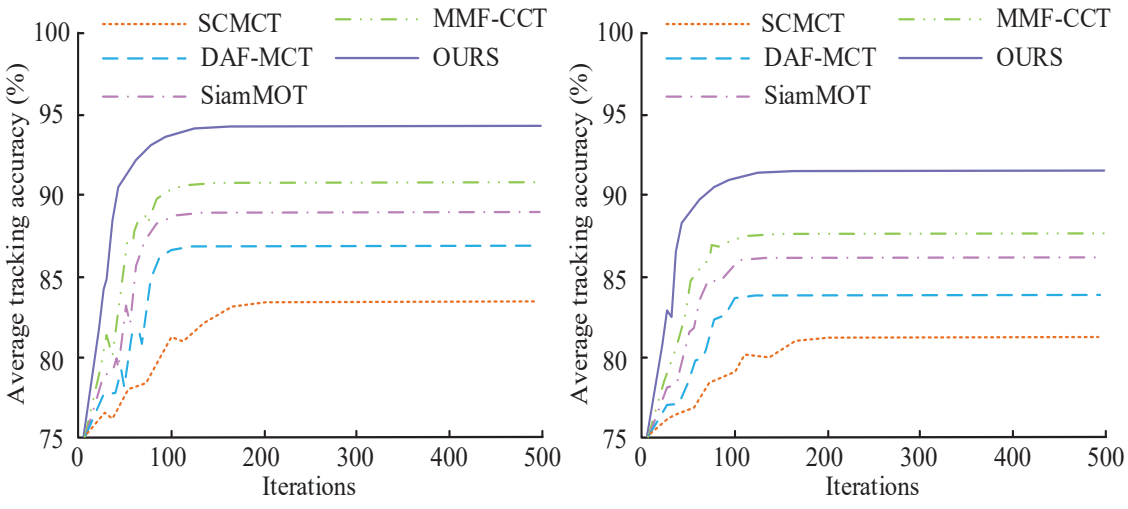
Table 3
Sensitivity analysis of superpixel segmentation algorithms and parameters on final segmentation performance.

Superpixel algorithm	Number of superpixels	SA (standard, %)	mIoU (standard, %)	SA (30% noise, %)	mIoU (30% noise, %)
SLIC	100	89.2 ± 0.62 [88.6-89.7]	93.5 ± 0.48 [93.1-93.9]	85.6 ± 0.55 [85.1–86.1]	90.2 ± 0.45 [89.8–90.6]
	300	90.5 ± 0.55 [90.0–91.0]	94.8 ± 0.39 [94.4–95.2]	86.3 ± 0.48 [85.9–86.7]	94.0 ± 0.33 [93.7–94.3]
	500	90.3 ± 0.58 [89.8–90.8]	94.6 ± 0.42 [94.2–95.0]	86.1 ± 0.51 [85.6–86.6]	90.8 ± 0.40 [90.4–91.2]
Felzenszwalb	100	88.7 ± 0.71 [88.1–89.3]	93.1 ± 0.54 [92.6–93.6]	84.9 ± 0.68 [84.3–85.5]	89.6 ± 0.52 [89.1–90.1]
	300	89.8 ± 0.64 [89.2–90.4]	94.2 ± 0.42 [93.8–94.6]	85.7 ± 0.58 [85.2–86.2]	90.4 ± 0.44 [90.0–90.8]
	500	89.5 ± 0.66 [88.9–90.1]	93.9 ± 0.47 [93.5–94.3]	85.4 ± 0.61 [84.8–86.0]	90.1 ± 0.49 [89.6–90.6]
Watershed	100	88.1 ± 0.75 [87.4–88.8]	92.6 ± 0.59 [92.1–93.1]	84.2 ± 0.72 [83.6–84.8]	89.0 ± 0.58 [88.5–89.5]
	300	89.3 ± 0.68 [88.7–89.9]	93.8 ± 0.51 [93.3–94.3]	85.0 ± 0.64 [84.4–85.6]	89.8 ± 0.51 [89.3–90.3]
	500	89.0 ± 0.70 [88.4–89.6]	93.5 ± 0.55 [93.0–94.0]	84.8 ± 0.67 [84.2–85.4]	89.5 ± 0.54 [89.0–90.0]

ity experiments were conducted on the Market-1501 dataset. For each algorithm, three different numbers of superpixels (100, 300, and 500) were tested. Recorded SA and mIoU under standard conditions and 30% speckle noise conditions. The results are shown in Table 3. As shown in Table 3, the proposed method maintained consistent stable performance under different superpixel algorithms and parameter settings. Under all configurations, the standard deviation of SA was only 0.71% under standard conditions and 0.68% under 30% noise conditions. The maximum performance degradation caused by suboptimal superpixel segmentation relative to the optimal configuration was only 2.1% on SA and 2.0% on mIoU. These results quantitatively confirmed the robustness of the proposed method to changes in superpixel segmentation quality. This robustness was theoretically attributed to the fact that the affinity matrix combines pixel level features and region level information derived from superpixels through mul-

Figure 11

Comparison of average tracking accuracy of different cross-camera tracking methods on the Market-1501 and WildTrack datasets over increasing iteration steps.



(a) Average tracking accuracy of different methods on Market-1501 dataset

(b) Average tracking accuracy of different methods on WildTrack dataset

Table 4

Ablation study on the contribution of different modules in the proposed tracking framework.

Experimental configuration	Datasets	Precision (%)	Recall (%)	F1-Score (%)	Parameters (M)
Baseline (SiamMask)	Market-1501	88.03	84.91	86.45	12.50
	WildTrack	86.14	82.87	84.48	
+ Large-kernel attention	Market-1501	91.22	88.71	89.95	14.76
	WildTrack	89.53	86.32	87.90	
+ Masked autoencoder SSL	Market-1501	92.48	90.12	91.28	16.23
	WildTrack	90.97	88.42	89.69	
Full model	Market-1501	94.31	92.53	93.41	17.58
	WildTrack	93.12	90.24	91.66	

tiplication. When the region information was inaccurate, the pixel level similarity term based on FVs served as a correction mechanism.

4.2. Performance Analysis of Community CCTT Method

To evaluate the performance of the proposed community CCTT method, this study conducted experiments on the Market-1501 and WildTrack datasets. This study compares the proposed cross-camera object tracking method with existing mainstream methods, including spatially constrained multi-camera tracking (SCMCT) method, deep appearance feature-based multi-camera tracking (DAF-MCT) method, siamese network-based online multi-object tracking (SiamMOT) method and multi-modal fusion-based cross-camera tracking (MMF-CCT) method.

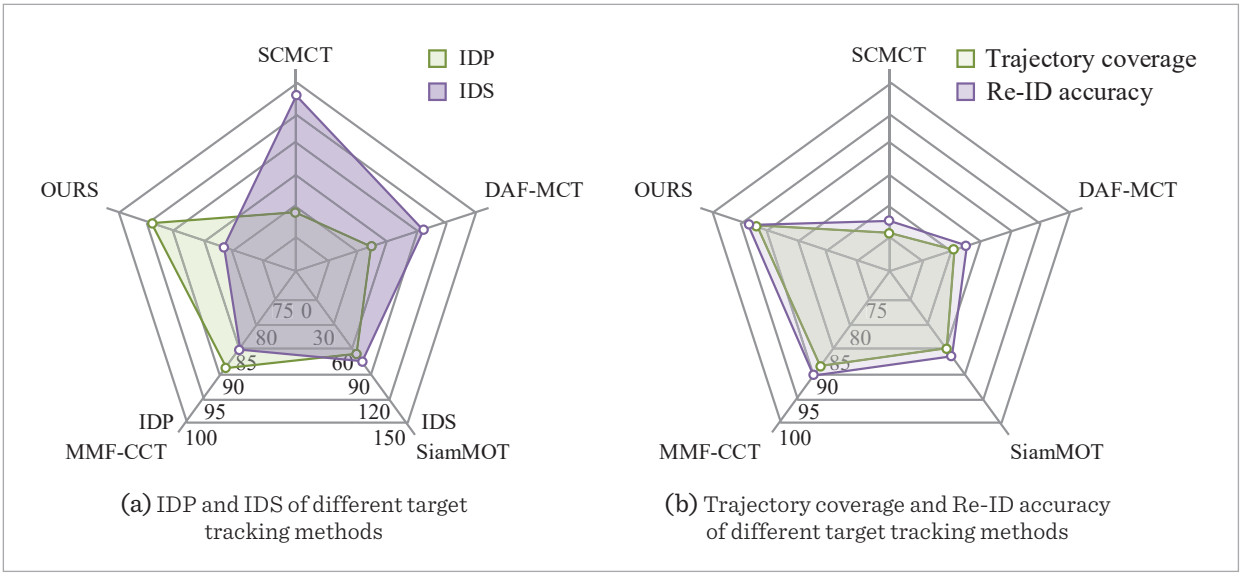
This study analyzed the average tracking accuracy of different target tracking methods on two datasets. In Figure 11(a), on the Market-1501 dataset, when the iteration was 100, the average tracking accuracy of SCMCT, DAF-MCT, SiamMOT, and MMF-CCT were 81.2%, 86.0%, 87.8, and 90.1%. The accuracy of the proposed method was 93.7%. When the iteration reached 300, the average tracking accuracy of the five methods was 82.5%, 86.3%, 88.9%, 90.2, and 94.3%. In Figure 11(a), on the WildTrack dataset, when the iteration was 300, the average tracking accuracy of SCMCT, DAF-MCT, SiamMOT, and MMF-CCT were 80.6%, 83.6%, 85.4, and 87.1%. The suggested approach had a 91.6% accuracy rate. The outcomes validated the superior cross-camera tracking performance of the suggested technique by demonstrating its ability to retain high tracking accuracy on both datasets.

To explore the specific contribution of each module to CCTT performance, this study conducts ablation experiments on two datasets. The evaluation indicators include precision rate, recall rate, F1-Score, and parameter amount. The results are shown in Table 3. On the Market-1501 dataset, the F1-Score of the baseline model was 86.45%. After adding large-core attention, it increased to 89.95%, and after further adding masked autoencoder self-supervised learning, it reached 91.28%. The complete model reached the optimal 93.41%. The F1-Score of the entire model was 91.66%, a considerable improvement over the baseline model, and the WildTrack dataset exhibited the similar trend. The outcomes displayed that each module had a positive contribution to performance, and the complete model performed best in both precision and recall. To reveal why large-kernel attention improves tracking, feature response maps were visualized for the same target across different cameras. Without this module, responses were scattered over backgrounds and occluded regions. After adding it, the network consistently focused on discriminative parts under scale changes and partial occlusions. This confirmed that the performance gain primarily stemmed from enhanced robustness to scale variation and occlusion, not merely increased parameters.

To verify the practical application effect of the proposed method, this study analyzes the identity preservation rate (IDP), identity switches (IDS), trajectory coverage and re-identification accuracy of different CCTT methods in an actual monitoring scenario of a smart community. In Figure 12(a), the IDPs of SCMCT, DAF-MCT, SiamMOT, and MMF-CCT were 79.5%, 83.2%, 86.8%, and 89.4%, respec-

Figure 12

Comparison of IDP, IDS, trajectory coverage, and re-identification accuracy of different cross-camera tracking methods in an actual smart community monitoring scenario.



(a) IDP and IDS of different target tracking methods

(b) Trajectory coverage and Re-ID accuracy of different target tracking methods

Table 5
Tracking performance under different segmentation inputs.

Input type	Condition	IDP (%)	IDS	Tracking accuracy (%)	mIoU (%)
Raw image (no segmentation)	Standard	82.3	87	86.5	/
Mask R-CNN segmentation	Standard	88.6	62	90.2	89.4
Proposed FCMC segmentation	Standard	94.7	38	94.3	94.8
Raw image (no segmentation)	Noisy (30%)	68.5	156	71.3	/
Mask R-CNN segmentation	Noisy (30%)	79.2	108	82.6	81.5
Proposed FCMC segmentation	Noisy (30%)	91.8	51	92.4	93.1

tively. The IDS are 142, 98, 73, and 61, respectively. The IDP and IDS of the proposed method are 94.7% and 38, respectively. In Figure 12(a), the trajectory coverage rates of SCMCT, DAF-MCT, SiamMOT, MMF-CCT, and the proposed method are 76.2%, 80.7%, 85.2%, 88.5%, and 92.8%, respectively. The re-identification accuracy rates were 77.5%, 82.4%, 86.9%, 90.1%, and 93.6%. This fully verified the advantage of the suggested method in maintaining target identity continuity and tracking integrity in actual complex scenarios.

To verify whether the proposed improved FCMC segmentation is an essential component or just an interchangeable preprocessing module, a controlled ablation study was conducted. Feed three different types of inputs into the proposed large kernel attention tracking network: (1) the original monitoring image (baseline), (2) the segmentation results of Mask R-CNN (a classic instance segmentation method), and (3) the segmentation results of the proposed improved FCMC algorithm. All other tracking parameters and modules remained the same. The experiment was conducted on the Market-1501 dataset under standard conditions and noise conditions (30% speckle noise). The results are shown in Table 5. In Table 5, when the original image was directly used, the tracking network only achieved 82.3% IDP and 86.5% tracking accuracy under standard conditions, while its performance was severely degraded under noisy conditions. The introduction of Mask R-CNN segmentation improved performance, but it was still significantly lower than the pipeline proposed in this paper. Under 30% speckle noise, the FCMC segmentation in this paper achieved 91.8% IDP and 92.4% tracking accuracy, which were 12.6% and 9.8% higher than Mask R-CNN, respectively. These results quantitatively confirmed that the proposed improved FCMC segmentation was not replaceable by classical segmentation methods. In contrast, Mask R-CNN trained on fixed domain data and lacked this adaptive spatial constraint, making it sensitive to domain offset and noise, resulting in higher identity switching times and lower tracking accuracy. Therefore, in the entire tracking framework proposed, the segmentation module is a necessary and critical component, rather than an independent and interchangeable module.

To investigate the actual tracking performance of the proposed method, an analysis was conducted on the visualization results of pedestrian tracking across cameras in an intelligent community system deployed in a residential area. This study tracked pedestrians with three different cameras, and the results are shown in Figure 13. In Figure 13, the proposed method successfully achieved cross camera continuous tracking of pedestrian targets in actual intelligent community scenarios. From the perspective of camera 1, the target is moving at a speed of 1.1 m/s. From the perspective of camera 2, the target is located in the central path and maintains a stable tracking state. From the perspective of camera 3, the target ultimately reaches Building B. Three cameras in different positions can accurately recognize the same target and maintain its identity consistency without identity switching. This result verifies that the proposed method has good cross camera target matching and continuous tracking capabilities under actual non overlapping perspectives and different scene changes.

Figure 13
Visualization results of pedestrian tracking using the proposed method.



Tracking visualization camera 1 Tracking visualization of camera 2



Tracking visualization of camera 3

This study further compares the average processing frame rate (frames per second, FPS) and average tracking delay of different CCTT methods in actual monitoring scenarios of smart communities to verify the real-time tracking performance of the proposed method. The FPS of SCMCT, DAF-MCT, and SiamMOT were 28, 32, and 36, respectively. The FPS of MMF-CCT and the proposed method were 39 and 42, respectively. The average tracking delays of SCMCT, DAF-MCT, SiamMOT, and MMF-CCT were 105.2ms, 93.8ms, 82.5ms, and 77.3ms, respectively. In comparison, the average tracking delay of the suggested method was 65.6ms. The findings demonstrated that the suggested approach could match the genuine requirements of high efficiency and low latency for intelligent community monitoring while achieving effective real-time processing capabilities while keeping high tracking accuracy.

To evaluate the practicality of the proposed cross camera tracking method, a computational efficiency analysis was conducted. The inference speed (frames per second, FPS), GPU memory consumption, GPU utilization, and CPU utilization of all comparison methods were measured on the Market-1501 dataset. The results are shown in Table 6. In Table 6, the proposed method achieved the highest inference speed (42 FPS) among all compared methods. The GPU memory consumption of the proposed meth-

Table 6
Computational efficiency comparison of different cross-camera tracking methods.

Methods	Inference speed (FPS)	GPU memory (GB)	GPU utilization (%)	CPU utilization (%)
SCMCT	28.0 ± 1.2	5.2 ± 0.15	62.0 ± 2.5	45.0 ± 2.0
DAF-MCT	32.0 ± 1.0	6.8 ± 0.12	58.0 ± 2.2	38.0 ± 1.8
SiamMOT	36.0 ± 0.9	7.5 ± 0.10	71.0 ± 2.0	42.0 ± 1.5
MMF-CCT	39.0 ± 0.8	8.3 ± 0.08	68.0 ± 1.8	35.0 ± 1.2
OURS	42.0 ± 0.6	7.9 ± 0.07	65.0 ± 1.5	32.0 ± 1.0

Table 7
Comparison with recent state-of-the-art cross-camera tracking methods.

Index	TransTrack	QDTrack	OURS
Tracking accuracy (%)	91.2 ± 0.5	92.5 ± 0.4	94.3 ± 0.3
IDP (%)	91.8 ± 0.6	92.9 ± 0.5	94.7 ± 0.4
IDS	52 ± 4	44 ± 3	38 ± 2
Inference speed (FPS)	38.5 ± 0.7	40.2 ± 0.6	42.0 ± 0.6
GPU memory (GB)	8.1 ± 0.09	7.8 ± 0.08	7.9 ± 0.07
GPU utilization (%)	67.5 ± 1.6	66.8 ± 1.4	65.0 ± 1.5
CPU utilization (%)	34.2 ± 1.1	33.5 ± 1.0	32.0 ± 1.0

od was 7.9 GB, which was comparable to SiamMOT and MMF-CCT5. The GPU utilization and CPU utilization of the proposed method were at a moderate level, at 65.0% and 32.0%, respectively, indicating high resource utilization efficiency. These results confirmed that the proposed method not only performed well in accuracy, but also had computational feasibility in the actual deployment of intelligent community security systems.

The study conducted additional comparisons between the proposed method and the latest and most advanced cross camera tracking methods, including Transformer-based Tracking (TransTrack) and Quasi-Dense Tracking (QDTrack), The results are shown in Table 7. In Table 7, the proposed method outperformed the two latest state-of-the-art methods in both tracking performance and computational efficiency. In terms of tracking accuracy, this method achieved 94.3%, higher than QDTrack's 92.5% and TransTrack's 91.2%. The IDP was 94.7%. In terms of inference speed, the proposed method achieved 42.0 FPS, which was higher than the comparison methods. The above results fully verified the superiority and efficiency of the proposed method in cross camera tracking tasks.

5. Discussion

This study suggested a cross-camera tracking method for intelligent communities based on DL and FC algorithms, which showed excellent performance in cross-camera monitoring scenarios. In terms of IS, the SA of the suggested algorithm on the Market-1501 and WildTrack datasets reached 90.5% and 83.8% respectively. Compared with existing FC algorithms, the proposed algorithm effectively improved its robustness in noisy environments after introducing spatial regularization and fixed clustering center strategies. The fixed strategy significantly enhanced robustness by preventing centroid drift in high-noise environments, such as community surveillance with 30% speckle noise. However, this involved a theoretical trade-off: While it provided a stable prior that ensures identity consistency, it may slightly reduce local adaptability in scenes with extreme illumination shifts or highly dynamic backgrounds. This constraint is mitigated in the proposed framework by the ensemble of multiple sub-models, which restores global flexibility without sacrificing the noise-rejection benefits of the fixed centers. Compared with existing research, the SA of the HMM-FCMC proposed in [26] was lower than 80% in a noisy environment. The method proposed in the study still maintained an accuracy of more than 90% under the same conditions, showing stronger anti-interference ability. The LRR-FC in [19] performed well in edge preservation. However, the method proposed in the study further reduced the BDE index by about 1.2 pixels, indicating that the boundary segmentation was more accurate.

In terms of tracking performance, the F1-Score of the proposed method reached 93.41% and 91.66% on the Market-1501 and WildTrack datasets respectively, which was better than existing methods based

on twin networks or multi-modal fusion. In the actual scenario test, the proposed method achieved a processing speed of 42 FPS and an average delay of 65.6ms while maintaining high tracking accuracy, meeting the needs of real-time monitoring. Compared with the multi-camera perception system proposed in literature [27], the proposed method showed higher IDP and IDS in CCTT, which were 94.7% and 38, respectively. Compared with the spatiotemporal trajectory matching network based on appearance analysis in the literature [29], the self-supervised learning mechanism introduced in the study reduced the dependence on annotated data, and the tracking accuracy on the WildTrack dataset was increased by about 4%. The tracking latency of 65.6ms suggested potential applicability to high-speed targets. However, motion blur remains challenging.

In summary, the method proposed in the study achieves high SA, tracking accuracy and identity consistency in both public datasets and actual smart community scenarios. At the same time, this approach offers an effective and dependable technical path for intelligent security systems in complicated open environments and has good real-time processing capabilities.

6. Conclusion

This research proposed an intelligent community cross-camera tracking method based on DL and FC algorithm. It achieved robust IS through an improved FCMC algorithm, and combined the AM and self-supervised learning to achieve efficient target feature extraction and continuous cross-camera tracking. According to experimental results, the suggested approach outperformed current mainstream approaches in both real-world settings and public datasets. It also did well in SA, tracking accuracy, identity retention, and real-time performance. It provided effective technical support for smart community security systems. The limitation of this research is that the proposed algorithm still has room for improvement in its adaptability to extreme lighting and dense occlusion scenes, and the deployment efficiency of the model on embedded devices needs to be optimized.

Future research will focus on weak target recognition in cross camera scenes. Based on the theory of ultra high speed weak target enhancement, the tracking system will be optimized to improve the continuous correlation of small targets at long distances in non-overlapping camera views.

References

1. Ali, N. A., Abbassi, A. E., Cherradi, B. The Performances of Iterative Type-2 Fuzzy C-Mean on GPU for Image Segmentation. The Journal of Supercomputing, 2022, 78(2), 1583-1601. https://doi.org/10.1007/s11227-021-03928-9
2. Alomoush, W., Alrosan, A., Alomari, Y. M., Alomoush, A. A., Almomani, A., Alamri, H. S. Fully Automatic Grayscale Image Segmentation Based on Fuzzy C-Means with Firefly Mate Algorithm. Journal of Ambient Intelligence and Humanized Computing, 2022, 13(9), 4519-4541. https://doi.org/10.1007/s12652-021-03430-3
3. Azmi, N., Kamarudin, L. M., Zakaria, A., Zakaria, S. M. M. S. AI-Based Analytics for Hawkers Identification in Video Surveillance for Smart Community. Journal of Techno-Social, 2023, 15(2), 77-87. https://doi.org/10.30880/jts.2023.15.02.008
4. Gao, B., Li, Z., Zhang, D., Liu, Y., Chen, J., Lv, Z. Roadside Cross-Camera Vehicle Tracking Combining Visual and Spatial-Temporal Information for a Cloud Control System. Journal of Intelligent and Connected Vehicles, 2024, 7(2), 129-137. https://doi.org/10.26599/JICV.2023.9210034
5. Gao, X., Zhang, Y., Wang, H., Sun, Y., Zhao, F., Zhang, X. A Modified Fuzzy Clustering Algorithm Based on Dynamic Relatedness Model for Image Segmentation. The Visual Computer, 2023, 39(4), 1583-1596. https://doi.org/10.1007/s00371-022-02430-4
6. Gao, Y., Li, H., Li, J., Cao, C., Pan, J. Patch-Based Fuzzy Local Weighted C-Means Clustering Algorithm with Correntropy Induced Metric for Noise Image Segmentation. International Journal of Fuzzy Systems, 2023, 25(5), 1991-2006. https://doi.org/10.1007/s40815-023-01485-2
7. Gong, M., Zhao, Y., Li, H., Qin, A. K., Xing, L., Li, J., Liu, Y. Deep Fuzzy Variable C-Means Clustering Incorporated with Curriculum Learning. IEEE Transactions on Fuzzy Systems, 2023, 31(12), 4321-4335. https://doi.org/10.1109/TFUZZ.2023.3283046
8. He, M., Wang, R., Wang, Y., Zhou, F., Guo, N. DMPH-Net: A Deep Multi-Scale Pyramid Hybrid Network for Low-Light Image Enhancement with Attention Mechanism and Noise Reduction. Signal, Image and Video Processing, 2023, 17(8), 4533-4542. https://doi.org/10.1007/s11760-023-02687-9
9. Hooda, H., Verma, O. P. Fuzzy Clustering Using Gravitational Search Algorithm for Brain Image Segmentation. Multimedia Tools and Applications, 2022, 81(20), 29633-

29652. https://doi.org/10.1007/s11042-022-12336-x

10. Jiao, J., Wang, X., Wei, T., Zhang, J. An Adaptive Fuzzy C-Means Noise Image Segmentation Algorithm Combining Local and Regional Information. IEEE Transactions on Fuzzy Systems, 2023, 31(8), 2645-2657. https://doi.org/10.1109/TFUZZ.2023.3235392
11. Lei, L., Wu, C., Tian, X. Robust Deep Kernel-Based Fuzzy Clustering with Spatial Information for Image Segmentation. Applied Intelligence, 2023, 53(1), 23-48. https://doi.org/10.1007/s10489-022-03255-3
12. Li, X., Pei, W., Wang, Y., He, Z., Lu, H., Yang, M. H. Self-Supervised Tracking via Target-Aware Data Synthesis. IEEE Transactions on Neural Networks and Learning Systems, 2023, 35(7), 9186-9197. https://doi.org/10.1109/TNNLS.2022.3231537
13. Luna, E., SanMiguel, J. C., Martínez, J. M., Carballeira, P. Graph Neural Networks for Cross-Camera Data Association. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 33(2), 589-601. https://doi.org/10.1109/TCSVT.2022.3207223
14. Mokayed, H., Quan, T. Z., Alkhaled, L., Sivakumar, V. Real-Time Human Detection and Counting System Using Deep Learning Computer Vision Techniques. Artificial Intelligence Applications, 2023, 1(4), 205-213. https://doi.org/10.47852/bonviewAIA2202391
15. Nguyen, T. T., Nguyen, H. H., Sartipi, M., Fisichella, M. Real-Time Multi-Vehicle Multi-Camera Tracking with Graph-Based Tracklet Features. Transportation Research Record, 2024, 2678(1), 296-308. https://doi.org/10.1109/TCSVT.2022.3207223
16. Peng, C., Zeng, Z., Gao, J., Zhou, J., Tomizuka, M., Wang, X., et al. PNAS-MOT: Multi-Modal Object Tracking with Pareto Neural Architecture Search. IEEE Robotics and Automation Letters, 2024, 9(5), 4377-4384. https://doi.org/10.1109/LRA.2024.3379865
17. Remmelzwaal, L. Object Detection and Tracking for Crate and Bottle Identification in a Bottling Plant Using Deep Learning. Artificial Intelligence Applications, 2023, 1(3), 175-179. https://doi.org/10.47852/bonviewAIA3202798
18. Sharma, H., Kanwal, N. Video Surveillance in Smart Cities: Current Status, Challenges & Future Directions. Multimedia Tools and Applications, 2025, 84(16), 15787-15832. https://doi.org/10.1007/s11042-024-19696-6
19. Song, S., Jia, Z., Yang, J., Kasabov, N. Image Segmentation Based on Fuzzy Low-Rank Structural Clustering. IEEE Transactions on Fuzzy Systems, 2023, 31(7), 2153-2166. https://doi.org/10.1109/TFUZZ.2022.3220925
20. Urse, M.-A., Spînu, S. End-to-End Multi-Camera Live Streaming and Tracking Framework for Multiple Persons in Unconstrained Scenes. Journal of Military Technology, 2025, 8(2), 21-28. https://doi.org/10.32754/JMT.2025.2.03
21. Wang, C., Zhou, M., Pedrycz, W., Li, Z. Comparative Study on Noise-Estimation-Based Fuzzy C-Means Clustering for Image Segmentation. IEEE Transactions on Cybernetics, 2022, 54(1), 241-253. https://doi.org/10.1109/TCYB.2022.3217897
22. Wang, G., Peng, C., Gu, Y., Zhang, J., Wang, H. Interactive Multi-Scale Fusion of 2D and 3D Features for Multi-Object Vehicle Tracking. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(10), 10618-10627. https://doi.org/10.1109/TITS.2023.3275954
23. Wang, S., Sheng, H., Zhang, Y., Yang, D., Shen, J., Chen, R. Blockchain-Empowered Distributed Multicamera Multitarget Tracking in Edge Computing. IEEE Transactions on Industrial Informatics, 2024, 20(1), 369-379. https://doi.org/10.1109/TII.2023.3261890
24. Wu, H., Shen, F., Zhu, J., Zeng, H., Zhu, X., Lei, Z. A Sample-Proxy Dual Triplet Loss Function for Object Re-Identification. IET Image Processing, 2022, 16(14), 3781-3789. https://doi.org/10.1049/ipr2.12593
25. Xu, H., Li, Y., Zhang, M., Tong, P. Sonar Image Segmentation Using a Multi-Spatial Information Constraint Fuzzy C-Means Clustering Algorithm Based on KL Divergence. International Journal of Machine Learning and Cybernetics, 2025, 16(4), 2761-2778. https://doi.org/10.1007/s13042-024-02419-6
26. Xu, R. Fuzzy C-Means Clustering Image Segmentation Algorithm Based on Hidden Markov Model. Mobile Networks and Applications, 2022, 27(3), 946-954. https://doi.org/10.1007/s11036-022-01917-7
27. Yang, H., Cai, J., Zhu, M., Liu, C., Wang, Y. Traffic-Informed Multi-Camera Sensing (TIMS) System Based on Vehicle Re-Identification. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(10), 17189-17200. https://doi.org/10.1109/TITS.2022.3154368
28. Zaman, M. I., Bajwa, U. I., Saleem, G., Raza, R. H. A Robust Deep Networks-Based Multi-Object Multi-Camera Tracking System for City-Scale Traffic. Multimedia Tools and Applications, 2024, 83(6), 17163-17181. https://doi.org/10.1007/s11042-023-16243-7
29. Zhang, X., Yu, H., Qin, Y., Zhou, X., Chan, S. Video-Based Multi-Camera Vehicle Tracking via Appearance-Parsing Spatio-Temporal Trajectory Matching Network. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(10), 10077-10091. https://doi.org/10.1109/TCSVT.2024.3398658

This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 (CC BY 4.0) License (http://creativecommons.org/licenses/by/4.0/).